Methods for assessment and prediction of QoE, preference and visual discomfort in multimedia application with focus on S-3DTV

Jing Li

l'université nantes angers le mans
PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Jing LI

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le label de l'Université de Nantes Angers Le Mans

Discipline : Informatique et applications
Spécialité : Informatique
Laboratoire : Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN)

Soutenue le 6 décembre 2013

École doctorale : 503 (STIM)
Thèse n° : ED 503-211

Methods for assessment and prediction of QoE, preference and visual discomfort in multimedia application with focus on S-3DTV

Méthodes pour l'évaluation et la prédiction de la Qualité d'expérience, la préférence et l'inconfort visuel dans les applications multimédia. Focus sur la TV 3D stéréoscopique

JURY

Rapporteurs :	**M. Sanghoon LEE**, Professeur, Yonsei University **M. Gilles COPPIN**, Professeur, Telecom Bretagne
Examinateurs :	**M^me^ Luce MORIN**, Professeur, Institut National des Sciences Appliquées **M^me^ Anne GUÉRIN DUGUÉ**, Professeur, Université Joseph Fourier
Directeur de thèse :	**M. Patrick LE CALLET**, Professeur, Université de Nantes
Co-encadrant de thèse :	**M. Marcus BARKOWSKY**, Maître de conférences, Université de Nantes

Acknowledgment

At the end of my PhD journey, I would like to take this opportunity to thank all of the people who have helped me and supported me during the three years of my PhD in France.

Firstly, I would like to express my deepest gratitude to my supervisors Prof. Patrick Le Callet and Marcus Barkowsky for offering me the opportunity to be a PhD student under their supervision. During these years, they have taught me how to do a research scientifically, systematically and rigorously with their professional expertise. I also want to thank them for their helps when I encounter any problems during my daily life in France. They are nice, patient, intelligent, professional, and with charming personality. It is very lucky for me to be a PhD student of them.

I would also like to express my great appreciation to my colleagues Romain Cousseau and Romuald Pépion, who helped me a lot in all of the subjective experiments of my thesis. I would like to say without their efforts, my research work would not be so productive. Many thanks also go to Matthieu Urvoy for our enjoyable collaborations on the psychophysical studies and an output of a book chapter on visual discomfort. I also want to thank the team members of the first Qualinet Summer school in Ilmenau, they are Francesca De Simone from Telecom ParisTech, Ondřej Kaller from Brno University of Technology, Jussi Hakala from Aalto University School of Science and Dawid Juszka from AGH-University of Science and Technology. I enjoyed the collaboration work in the summer school very much. I learned a lot from them.

Many thanks also goes to Jesús Gutiérrez from Universidad Politécnica de Madrid. We helped and supported each other to get through the most difficult period of our collaboration research work. I would also like to thank all the co-authors of my

research work, Dr. Fernando Jaureguizar, Prof. Julián Cabrera, and Prof. Narciso García from Universidad Politécnica de Madrid, Prof. Kjell Brunnström and Kun Wang from Department of Netlab in Acreo AB and Mid Sweden University, Taehwan Han, Sungwook Youn, Jiheon Ok, Chulhee Lee from Yonsei University, and Christer Hedberg, Indirajith Vijai Ananth from Department of Netlab in Acreo AB.

Furthermore, I feel very honored to have such highly renowned and competent experts in my Ph.D committee. Many thanks to them to review my thesis and provide me with very valuable comments. My great appreciation goes to Prof. Sanghoon Lee from Yonsei University, Prof. Gilles Coppin from Telecom Bretagne, Prof. Luce Morin from INSA and Prof. Anne Guérin Dugué from Université Joseph Fourier.

Finally, I would express my appreciation to all of my friends and my family for their helps and supports during the three years of my stay in France. My dear father and mother, thanks for your understanding. Many thanks to my dear friends Emilie, Sofiane(XiaoPang), Bangge, Xiaoyi, Kunshu, Wenzi, Zhujie, Fengjie, BaoBao, Yuwei, Jiazi, Haiyang, Zhaolei, Zhangyu, Dingbo, Chuanlin, Zeeshan, Pierre and so on. Last but not the least, I would like to express my great appreciation to my dear husband Junle Wang for his support and understanding to a PhD wife.

Contents

1

Introduction

1.1 Context

Stereoscopic-3D (S3D) technology is changing human's viewing experience nowadays. It provides the viewers with a more immersive and natural video scene. S3D releases become increasingly popular over the last few years, peaking with the epic visual effects in James Cameron's Avatar, Martin Scorsese's Hugo and most recently Ang Lee's Life of Pi.

This new wave of S3D movies leads to an increasing expectation for the industry to explore the possibility of the S3D technology for the home entertainment. A preliminary step is S3D movies through the packaged media (e.g. DVD and Blu-ray). In the long run, the wide application of S-3DTV broadcasting system is also necessary. A pioneer for this was a pay-television operator BSkyB who introduced a stereoscopic S-3DTV channel in the United Kingdom in October 2010 (For convenience, S3D and S-3DTV are replaced by 3D and 3DTV in the following sections of this thesis).

To achieve this goal, despite the high requirements on hardware, extensive efforts on the study of 3D viewing experience have been made by researchers and international standardization organizations. The research topics covered a wide range of disciplines, including neurology, psychology, optics, multimedia and broadcasting.

1.2 Motivations

Stereoscopic 3D has recently received much attention as a result of a strong push from the cinema industry. However, in recent years, an increasing amount of people started to reconsider the question: “is the 3D viewing really worth the hope and the money they pay for it?” .

Many factors may lead to this question, for examples, the inconvenience of wearing the 3D glasses, the limited sources of 3D video, and the experienced visual discomfort induced by 3D videos. For industry, the balance between the immersive viewing experience and the perceived visual discomfort is one of the most important concerns. A goal of the 3DTV society would be that the immersive viewing experience is enhanced as much as possible while the visual discomfort or visual fatigue is perceived as little as possible. To achieve it, three important questions are raised: 1) how to measure the 3D viewing experience subjectively; 2) which factors would influence the 3D viewing experience; and 3) how to model the 3D viewing experience objectively for the purpose of optimizing or controlling the multimedia processing or broadcasting system.

Recently, many international standardization organizations have been working on these issues. For example, the International Telecommunication Union (ITU) has published the standard subjective methodologies for the assessment of stereoscopic 3DTV systems including general test methods, grading scales and viewing conditions in ITU-R BT.2021[56]. The Society of Motion Picture & Television Engineers (SMPTE) focuses on the standardization related to stereoscopic 3DTV in production environments, e.g., the exchange of 3D content amongst mastering facilities, and between a mastering facility and the ingest facility of a distribution system. The Digital Video Broadcasting Project (DVB) and the European Broadcasting Union (EBU) provide interim recommendations regarding producing, exchanging, archiving, distributing, coding, and transmitting 3D programs using 2D compatible or newly developed 3D infrastructure and transmission technologies for home viewing. The 3DTV group of the Video Quality Experts Group (VQEG) [139] is currently working on three distinct projects, which concern the subjective assessment methodology, the influence factors evaluation, and the objective models. IEEE P3333 [1], which is an individual-based project approved by the IEEE Standards Association, recently started their work on the quality assessment of 3D displays, 3D contents and 3D devices based on human factors, such as photosensitive seizures, motion sickness, visual fatigue, and identification and quantification of the causes of those factors.

All these works have made big progress for the popularization of 3DTV in the home entertainment. However, we are still facing with numerous challenges:

(1) Reliable subjective assessment methodologies for 3D viewing experience

are still missing. 3D viewing experience, which is also defined as "Quality of Experience (QoE)" in 3DTV, is a combined viewing experience of image quality, depth quality and visual comfort. Thus, it is a multi-dimensional concept. Though ITU-R BT.2021[56] has published a recommendation for subjective assessment methodology for stereoscopic 3DTV system, these methods are based on the traditional 2D quality assessment methods, which only provide the subjective judgement on each particular aspect in 3D viewing, for example, image quality or visual discomfort. A multidimensional scale for the combined experience on 3DTV may not be provided by these methods. Therefore, a reliable method to measure this combined experience is highly required.

(2) What are the influence factors of the 3D viewing experience and how these factors affect it need investigations. There are numerous possible factors that would influence the 3D viewing experience, for example, the display technology, the test environment, or the characteristics of the viewers. Due to its complexity, a large number of cross-lab study is needed. In addition, this work should be based on a reliable ground truth obtained by a reliable subjective assessment methodology. However, nowadays, most of the studies on this issue are based on the subjective experiment by using traditional 2D assessment methods.

(3) As one of the most important dimensions of QoE in 3DTV, visual discomfort is an issue often complained by the viewers after watching the 3D videos. Thus, an objective visual discomfort model is needed to monitor and optimize the broadcasting systems automatically and then achieve the best viewing experience. However, due to 1) the lack of the 3D video databases, 2) the uncertainty of the reliability of the subjective assessment methodology, and 3) the incomplete study on influence factors, no objective visual discomfort models have been validated by a large enough number of databases.

(4) Besides the subjective assessment methods and objective models, the objective psychophysical prediction on viewing experience is required. In particular, the degree of visual discomfort or visual fatigue induced by 3DTV should be predicted as they are safety issues for consumers.

In order to solve the problems listed above, some researches have been conducted and will be presented in the remainder of this thesis. Basically, there are two main topics in this thesis: subjective assessment methodology for QoE in 3DTV, and the study on visual discomfort in 3DTV. The main content and the structure of the thesis are shown in the next section.

1.3 Thesis overview

The diagram in Figure 1.1 shows the outline of this thesis. Chapter 2 introduces some important concepts regarding the 3D display technology and 3D viewing experience, including the way humans perceive stereopsis, the way 3D display systems work, the definition of 3D QoE, the factors that influence the 3D viewing experience, the subjective and objective way of measuring the 3D QoE, and the challenges and difficulties the current 3D industry is facing.

Following Chapter 2, the remainder of the thesis is structured into two parts.

The first part of the thesis focuses on the subjective assessment methodology for 3DTV. In particular, an assessment methodology which might generate reliable results for 3DTV called "Paired Comparison" is investigated. An overview of the state-of-the-art researches on the paired comparison method is firstly introduced in this part. Three of my research works are then introduced: 1) the proposal of a set of efficient designs for paired comparison; 2) the evaluation of the proposed designs by a series of subjective assessment experiments in 3DTV; 3) an application of the proposed paired comparison designs in evaluating the influence factors of QoE in 3DTV. More details of these studies are presented as follows:

- Chapter 3 introduces the state-of-the-art research work on Paired Comparison. This chapter illustrates why and how the paired comparison method is feasible in assessing the multi-dimensional viewing experience when watching 3DTV. The standardized paired comparison method and some existing efficient designs for paired comparison are introduced. In addition, the mathematical tools for analyzing the paired comparison data are introduced, including the Thurstone-Mosteller (TM) model, the Bradley-Terry (BT) model and the Elimination By Aspects (EBA) model.

- Chapter 4 introduces the proposed efficient designs for paired comparison. Generally, the existing efficient designs for paired comparison are based on theory analysis under perfect conditions (e.g., no observation errors in subjective tests). However, in a real subjective assessment experiment, there are often various errors which would influence the test results, for example, the observer's unintentional mistakes on voting. Based on statistical analysis on the possible errors, a set of efficient designs which are also robust to these errors is proposed. The proposed designs are evaluated by Monte-Carlo simulation experiments.

- In Chapter 5, the proposed designs on paired comparison are evaluated by subjective visual discomfort experiments. Five paired comparison experi-

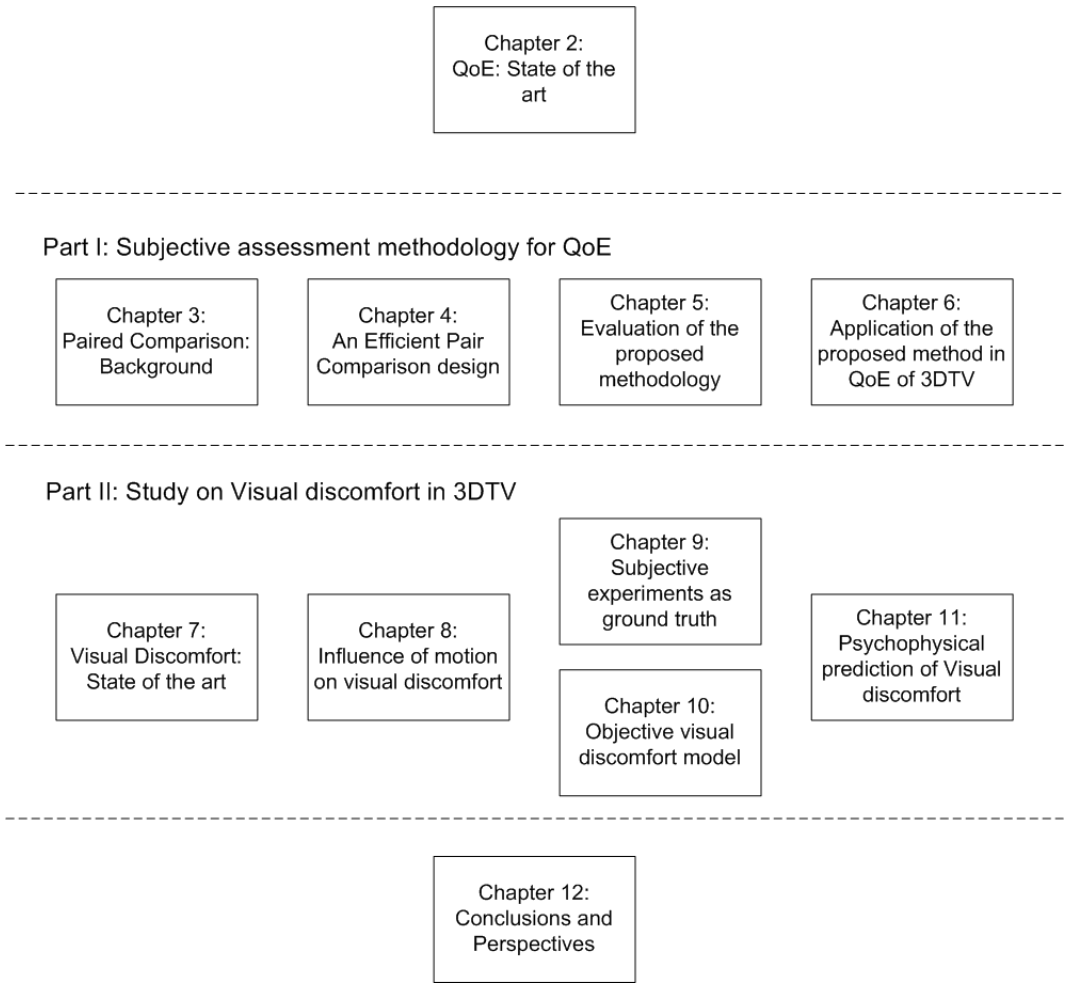


Figure 1.1: Overview of the thesis chapters.

ments aiming at different objectives were conducted. The performance of the proposed efficient designs are evaluated under different test scenarios, for example, under the influence of the observation errors, and under the influence of irrelevant stimuli.

– Chapter 6 provides an example about how to utilize the proposed pair comparison design on the study of QoE in 3DTV. A set of joint experiments were conducted in two labs considering various influence factors, for example, different 3D displays, test environments, viewers, etc. The significant factors of QoE in 3DTV are studied in this chapter, which is an important hint for the standardization of QoE in 3DTV.

The second part of the thesis focuses on the study of visual discomfort in 3DTV. Firstly, an overview of the state-of-the-art researches on visual discomfort is introduced, then, four of my research works are described: 1) the study on the influence of 3D motion on visual discomfort; 2) the comparison analysis on two different subjective assessment methodologies; 3) evaluation of the proposed objective visual discomfort model on natural stereoscopic video sequences and 4) the study on the relationship between the psychophysical signals and the experienced visual discomfort. The details are shown as follows:

– Chapter 7 introduces the state-of-the-art research work on visual discomfort, including the definitions of visual discomfort and visual fatigue in 3DTV, the possible factors that would induce visual discomfort, the widely used subjective assessment methodologies, and the psychophysical prediction of visual discomfort by devices.

– Chapter 8 focuses on the influence of 3D motion on visual discomfort. In the subjective experiments, the paired comparison method was used. Synthetic video sequences were used for precise control of the motion type, velocity and disparity values in the stimuli. This study clarifies a series of questions, including (1) which types of motion have more significant influence on visual discomfort, (2) how disparity affects visual discomfort, and (3) what is the inter-observer difference in the perception of visual discomfort. In addition, a psychophysical visual discomfort model is proposed according to the test results.

– Chapter 9 introduces a comparative study on the influence of different subjective test methodologies on visual discomfort. The subjective experimental results obtained by the ACR methodology and the Paired Comparison method-

ology are compared. The correlation between the two test results, the discriminability of the two test methodologies and the viewers' behavior in two tests are analyzed.

– In Chapter 10, the proposed psychophysical visual discomfort model is evaluated by natural stereoscopic video sequences. Two frameworks are proposed for the model. One is based on the tracked moving objects over sequences, the other is based on the moving objects in each frame. The subjective experimental results presented in Chapter 9 are used as the ground truth. The results indicate that the performances of the two frameworks are comparable and both showed higher correlation with the subjective data than an existing objective visual discomfort model in [65].

– Chapter 11 introduces a study on the psychophysical prediction of visual discomfort in 3DTV. The synthetic video sequences in Chapter 8 with different types of motion are used in the subjective test. An electro-physiological device is utilized to record the various eye movements signals of the viewers, particularly the eye blinking signals. The results showed that eye blinking rate is capable of predicting visual discomfort in 3DTV. Nevertheless, the relationship between eye blinking and visual discomfort is highly dependent on the type of motion in the videos.

At the end of the thesis, a summary of the contributions and some perspective for the future work are presented in Chapter 12.

2

Quality of Experience (QoE) in 3DTV

Television is one of the most important devices for home entertainment. Looking back at the history of television, from the black-and-white television to color television; from the analog television to the digital television; from SDTV to HDTV or even 3D HDTV, there is no doubt that improving viewers' viewing experience is a main driving force for the development of television technology as well as the broadcasting system.

3D technology is still relatively new to consumers. The "pros and cons" of this technology are somehow obvious. On one hand, the 3D technology provides the viewers with the experience of "being part of it"; on the other hand, viewers also often complain about the visual discomfort and visual fatigue after watching the 3D movies.

To improve the viewing experience of 3DTV, three basic questions are raised: 1) what is the mechanism of 3D displays, 2) which factors affect the viewing experience of 3D content, and 3) how to measure the viewing experience. In this chapter, we firstly introduce the mechanisms of the stereoscopic perception of human beings in Section 2.1, and the 3D display technologies in Section 2.2. Then, we introduce a particular terminology for 3D viewing experience, i.e., "Quality of Experience" in Section 2.3, where its multi-dimensionality and the possible influence factors are illustrated. Finally, in Section 2.4, the challenges of measuring 3D QoE and the problems of the existing subjective assessment methods are discussed. A candidate solution for these challenges and problems is then proposed.

Table 2.1: List of depth cues

Monocular cues	Binocular cues
Interposition	Vergence
Linear perspective	Binocular disparity
Light and shade	
Relative size	
Height in the visual field	
Texture gradient	
Aerial perspective	
motion parallax	
Oculomotor (Acommodation)	
Defocus blur	

2.1 Stereoscopic perception

Human beings have the ability to visually perceive the world in three dimensions. This ability is called depth perception. Depth perception relies on a variety of depth cues. As proposed in [91], these depth cues can be classified into two categories: monocular cues and binocular cues. As the name implies, monocular cues require the input from only one eye while binocular cues require inputs from both eyes. Each of these depth cues and the mechanisms by which they influence the depth perception are introduced in Table 2.1.

2.1.1 Monocular depth cues

– Interposition: Objects occluding each other suggest their depth ordering, in particular, a more distant object is partial blocked by a nearer object.

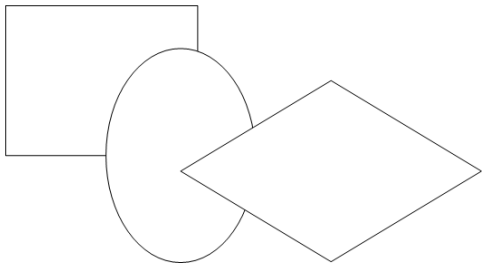


Figure 2.1: An example of interposition.

– Light and shade: It is usually agreed for all natural lights, and for most artificial lights, that the light comes from above to some degree. Thus, shadow plays a broader role in defining depth between objects since objects in shadow must be farther from the light than objects that are not in shadow.

– Relative size: An object with smaller retinal image is judged further away than the same object with a larger retinal image.

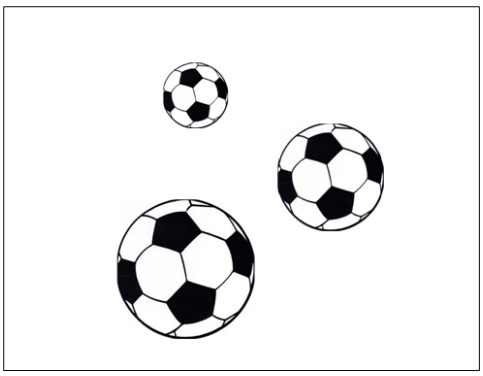


Figure 2.2: An example of relative size.

– Height in the visual field: This is a depth cue based on the vertical position of a point in the visual field. Objects further away are generally higher in the visual field.

– Texture gradient: Most surfaces, such as walls and roads and a field of flowers in bloom, have a texture. As the surface gets farther away from us this texture gets finer and appears smoother. Furthermore, the shape, size and density of the texture also affect the depth perception.

Figure 2.3: An example of texture gradient, which was taken in Bordeaux by the author.

– Linear perspective: It is the phenomenon that parallel lines that recede into the distance appear to converge.

– Aerial perspective: the atmosphere affects light traveling through it, for example due to fog, dust or rain. As light travels long distances it is scattered, colors loose saturation, sharp edges are diffused and color hue is shifted towards blue.

Figure 2.4: An example of linear perspective, which was taken in Saint-Emilion by the author.

Figure 2.5: An example of aerial perspective [39].

– Motion parallax: This is a depth cue existing in a dynamic scene. When we move or an object in the scene moves, objects that are closer to us move faster across our field of view than the objects in distance.

– Oculomotor (Accommodation): This is a nonvisual depth cue which concerns the change of the lens of the eyes. It is controlled by the ciliary muscles to maintain a sharply focused image of the fixated point. Fixation on a relatively near point corresponds to a relatively relaxed state of the muscles. Therefore, information on the state of the ciliary muscles provides the information of absolute fixation distance.

2.1.2 Binocular depth cues

– Vergence: This is a nonvisual depth cue. A vergence is the simultaneous movement of both eyes in opposite directions to obtain or maintain single binocular vision [15]. To look at an object closer by, the eyes rotate towards each other. For an object further away, the eyes rotate away from each other. When looking at an object in infinite distance, the eyes diverge until parallel.

– Binocular disparity: As human eyes are horizontally separated, the retinal images received by two eyes are slightly different. The brain fuses the left and right retinal images and then extracts the relative depth information from retinal disparity, i.e., the difference between corresponding points in these images.

2.2 3D displays

The history of 3D technology dates back to around 300 BC. The human binocular vision was firstly discovered by a Greek scientist Euklides. In 1838, Charles Wheatstone invented one of the first recorded devices for displaying three-dimensional images [142]. In 1922, a 3D mainstream film "The Power of Love" was created. It was recorded using two dissimilar colors and viewers wear anaglyph eyewear to perceive the 3D effect. In 1928, stereoscopic 3D television was shown for the first time by John Logie Baird [128].

3D technology has been rapidly developing during recent years. Basically, the 3D displays take advantage of mechanisms of the Human Visual System (HVS) on stereopsis perception. The human brain can fuse the images input from the left eye and the right eye. And then, from retinal disparity, the HVS extracts the relative depth information, i.e. the distance between corresponding points in these images [26]. Thus, the basic technique of 3D displays is to present the left and right views separately to the left and right eye. Then the two retinal images can be combined in the brain to generate the perception of 3D depth.

2.2.1 3D display classification

According to the demand of glasses or not, 3D displays can be classified into two types: stereoscopic and autostereoscopic displays. When using a stereoscopic display, the viewers have to wear an optical device to direct the left and right images to the appropriate eye, which is called aided viewing. For autostereoscopic displays, the technology of separating both views is integrated in the display, which is called free viewing.

Stereoscopic displays with aided viewing are widely used. They are classified into time-parallel or time-sequential displays. In time-parallel displays, the left and right views are displayed simultaneously on one or two screens. In such systems, the methods used to direct the distinct views to the appropriate eyes include: 1) location multiplexing; 2) anaglyph or color-multiplexing; and 3) polarization multiplexing.

In time-sequential displays, the left and right views of a stereo image pair are presented in rapid alternation. The stereo pairs are viewed using synchronized active shuttering glasses which open alternately for the appropriate eye while closing

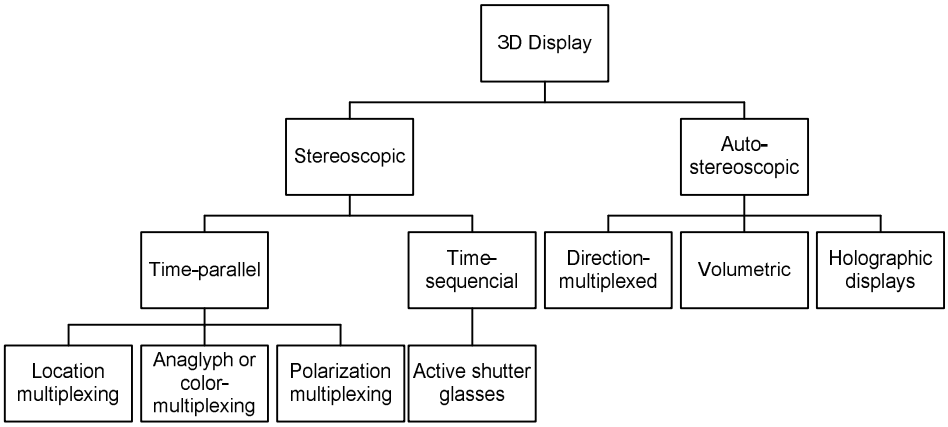


Figure 2.6: Classification of 3D displays

the other eye's view. This system exploits the characteristic of human visual systems that integrates a stereo pair across a time-lag of up to 50 ms [104].

Unlike stereoscopic displays, autostereoscopic displays do not need any glasses to present the two views. This type of displays sends the left and right images directly to the corresponding eyes. Currently, autostereoscopic displays can be classified into: 1) direction-multiplexed; 2) volumetric; and 3) holographic displays. For more details readers are referred to [135]. It is worth to note that, due to the free viewing, autostereoscopic displays are probably best suited for the application of 3DTV for home entertainment.

2.2.2 Definition of binocular disparity for 3D display

In the same viewing conditions, a larger disparity between the left and right view generally corresponds to a larger perceived depth. Disparity can be measured by various ways. A direct way is to use length units, e.g. pixels or centimeters. However, when using length units to measure disparity, the viewing condition (e.g. corresponding pixel size and viewing distance) should be provided as well.

Another way to measure disparity is to use the degree of visual angle [51] (see Figure 2.7). The binocular angular disparities ϕ_A and ϕ_B for point A and B can be calculated by Equation (2.1) and (2.2), respectively. For a point which is on the screen plane, the binocular angular disparity is 0 degree. A positive value represents crossed disparity, such as the point A; a negative value represents uncrossed disparity, such as the point B. In this way, the disparity is comparable at different conditions as the viewing distances and the distance between the two views have been taken into consideration in the expression of visual angle.

$$\phi_A = \beta - \alpha \tag{2.1}$$

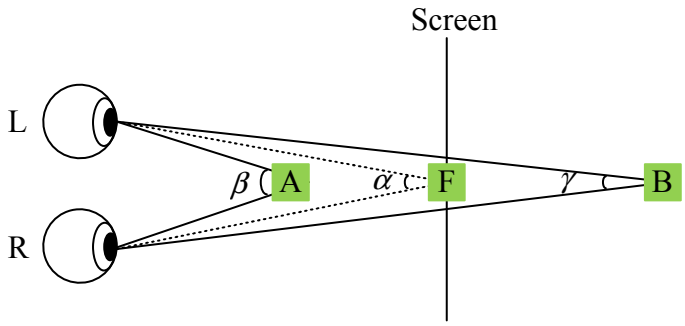


Figure 2.7: The definition of the binocular angular disparity, where F is the fixation point.

$$\phi_B = \gamma - \alpha \tag{2.2}$$

2.3 3D Quality of Experience

In 2D image/video quality assessment, "quality" is related to perceived image degradation, typically, compared to an ideal or perfect image. For 3D content, due to the added "depth" information provided by 3D displays, "quality" is not enough to describe the viewer's experience, particularly the immersive feeling, the accompanying emotion status and visual discomfort when watching 3D images/videos. Therefore, Quality of Experience (QoE) is proposed to replace "quality" in 3DTV visual quality assessment field.

The term QoE unites a multitude of meanings. Some of them were attributed to QoE and similar terms such as "Quality of Service" (QoS) in an ambiguous manner. According to ITU-T Rec. P.10 [108], QoE is defined as "the overall acceptability of an application or service, as perceived subjectively by the end user". Recently, representatives of more than 20 internationally recognized research institutions discussed this issue within the European Network of Excellence "Qualinet" (COST IC2003). They decided for the following working definition: "Quality of Experience (QoE) is 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 user's personality and current state" [80].

This definition has been triggered partially by the recent development of 3D video quality assessment methodologies. While it is evident that multimodal services, such as audiovisual services, require multidimensional quality analysis, 3D video quality assessment is a particularly interesting example of a monomodal ser-

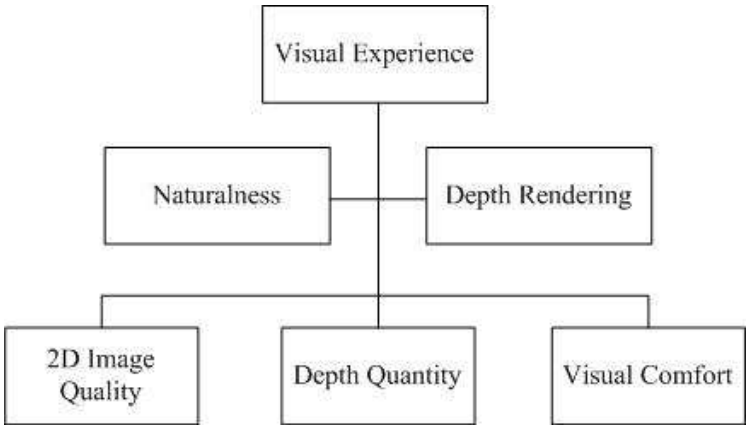


Figure 2.8: Quality of Experience model starting from primary factors on the bottom to more complex factors on higher levels.

vice stimulating the human's quality perception in a complex manner that may be modeled in a multidimensional approach.

2.3.1 Multidimensional perceptual scales for 3D QoE

The multidimensionality in 3D QoE is explained by the enhanced depth perception due to the stereopsis effect implemented, in most cases, by projecting two different images to each of the two human eyes and thus mimicking the real world situation in a fixed head position. The technical implementation in 3D capture systems and 3D display devices has deficiencies leading to visual annoyances such as visual discomfort sensations or visual fatigue symptoms.

Several models have been proposed to explain the human's integration of the different aspects, an excerpt will be provided here. Seuntiens et al. proposed a combination of perceived depth, binocular image distortion, and visual strain to model viewing experience in the presence of crosstalk in 2005 [112]. Kaptein et al. proposed to enhance the well-known 2D image quality measurement by adding a depth evaluation and the combination would then lead to a notion of naturalness [67]. This model has further been refined by Lambooij et al. towards a two level perceptual process which measures image quality and amount of depth as primary indicators and naturalness and viewing experience as derived, higher level indicators [78]. Chen et al. added visual comfort as primary indicator to the model and noted that two levels of derived perceptual criteria may be appropriate. He positioned naturalness and depth rendering on the second level and visual experience on the third level [19] leading to the pyramidal representation shown in Figure2.8.

Added value of depth

The depth perceived in stereoscopic 3D reconstruction maintains all previously perceived 2D depth cues, such as occlusion, relative size and relative density, height

in the visual field, aerial perspective, texture gradients, light and shading, and linear perspective. Most of the current 3D displays are limited to two views, such as polarized passive or active shutter glasses displays. These displays would then add binocular disparity and eventually the convergence state of the eyes as depth cue. Autostereoscopic displays provide more than two views and may therefore also reconstruct motion parallax to a certain extent.

Cutting and Vishton have analyzed the Just Noticeable Difference (JND) of object's depth position [27]. They observed that binocular disparities may offer an important depth position cue at short distances which decreases linearly with log-distance. At a viewing distance of about 1.5 m, a typical viewing distance for a 42 inches screen, the depth resolution of the human eye would correspond to about 1.5 cm. Using a visual depth acuity threshold of 20 arcsec, the minimum perceivable depth difference would correspond to about 9.4 cm in the same situation. On an autostereoscopic display, a psychophysical test has shown that the perceived JND may be in between these values [4].

The disparity distribution as shown to the observers mostly influences the perceived depth quantity effect. The qualitative effect of depth also relates to the reconstruction of the depth volume, in particular the relationship between horizontal and vertical compared to depth extents. Extreme depth compression may lead to cardboard effects or even puppet theater effect. To improve perceived depth quality, a stereoscopic shooting rule was developed to allow for improved reconstruction of S-3D content using two camera models [20].

Visual discomfort and visual fatigue

Visual discomfort and visual fatigue contain a wide range of visual symptoms, for example, headaches, tiredness, eye strains. Usually, in our daily life, visual discomfort or visual fatigue are due to some work demanding focusing or converging the eyes on an object for a long time. For example, watching TV for a long time, reading in a dark room, or reading texts with tiny font.

The added binocular depth introduced by 3D technology may provide viewers not only a totally different and enhanced viewing experience, but also visual discomfort and visual fatigue issues. Recently, it is often complained by the viewers that watching stereoscopic 3D content would induce more visual discomfort or visual fatigue when compared with the 2D video content. Thus, visual discomfort and visual fatigue are gaining increasing attention as besides the decreased viewing experience, it related to the viewers' health and safety issues.

Generally, it is believed that the imperfect simulation of depth cues on 3D display is the main cause of visual discomfort or visual fatigue, for example, the geometrical distortions between the left view and right view, the bright difference

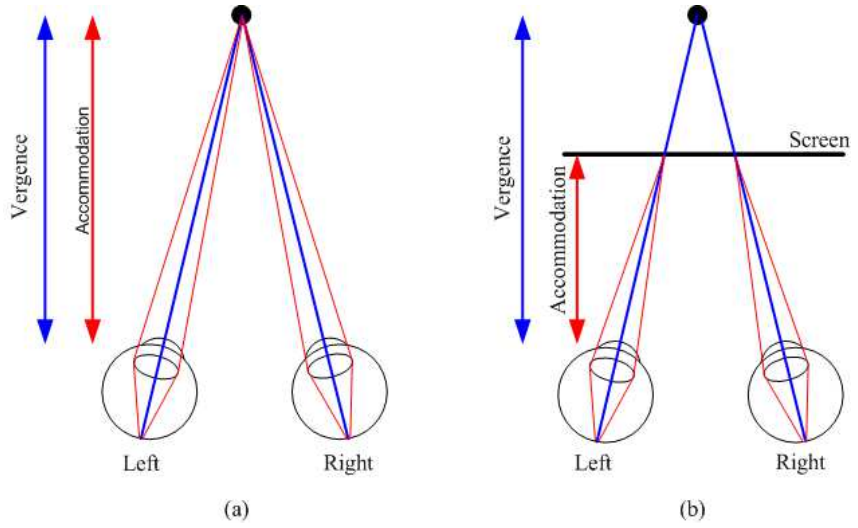


Figure 2.9: Comparison of Vergence-Accommodation conditions in (a) daily life and (b) watching 3D displays.

between the left and right views, and the conflict between vergence eye movement and accommodation. Here, the vergence and accommodation conflict is briefly introduced.

As illustrated in Section 2.1, both accommodation and vergence are nonvisual depth cues. Accommodation is controlled by the ciliary muscles which adjust the focal distance of the eyes by changing the shape of the lens. To keep the retinal image sharp, the focal distance of accommodation should be kept within a certain range around the object. On the other hand, to keep the stimuli fused, the eyes should converge to a distance close to the object distance. In our daily life, the accommodation and vergence are normally coupled. Accommodative changes can evoke vergence changes, and vergence changes can also evoke accommodative changes [36]. Focal and vergence distances are always close.

However, when watching 3D content by 3D displays, accommodation and vergence may be decoupled. As shown in Figure 2.9(b), when viewing an object by means of a 3D screen, the eyes will converge to the virtual object. However, the accommodation has to be performed at the screen depth level to keep the image sharply focused. This discrepancy is unnatural and will not happen in our daily life. The larger this discrepancy between vergence and accommodation gets, the higher the probability that observers would perceive visual discomfort.

This section just provides the readers a brief understanding about visual discomfort in 3DTV. For more details, the readers are referred to Chapter 7.

2.3.2 Influence factors of 3D QoE

QoE might be influenced by many factors. Generally, the influence factors (IF) can be grouped in three categories, namely System IF, Context IF and Hu-

man IF [80]. System IFs refer to "properties and characteristics that determine the technically produced quality of an application or service." [63]. They are related to media capture, coding, transmission, storage, rendering and display. Context IFs are factors that "embrace any situational property to describe the user's environment in terms of physical, temporal, social, economic, task and technical characteristics" [64]. A human IF is "any variant or invariant property or characteristic of a human user", e.g., the user's visual and auditory acuity, gender, age, previous experiences, education background [119].

System factors

In 3DTV broadcasting chain, one of the most important factors of QoE is the System IFs. For example, due to the limitation of the broadcasting bandwidth, the video sequences have to be compressed by encoders before transmission. The performance of different coding schemes on QoE is different [140][81]. Another influence factor for 3D QoE is the image/video format. There are several formats for 3D videos. For example, frame sequential (e.g. frame packing) format and frame compatible (e.g. side-by-side) format. The frame sequential format allows each view to have Full High Definition (HD) resolution while in frame compatible format, the left and right images are grouped into a single 2D HDTV frame with the resolution halved. Current 3DTV broadcasting systems are mostly limited to Side-by-Side (SBS) contents, as it can be processed by traditional 2D Full HD broadcasting chains. In [147], the influences of video formats on QoE were studied. The results showed that for uncompressed video, the quality and depth perception of the frame-sequential video have higher QoE than SBS video. But for the encoded video, the quality and depth perception depend on the amount of spatial and temporal information of the video sequences.

As introduced in Section 2.2, there are various types of 3D display. Different 3D display technologies influence the viewing experience differently and none of them can be considered as "transparent" as 2D displays. In [66], the authors studied the angular characteristics of polarization multiplexed and time multiplexed 3D displays. Through the evaluation of the viewing angle-related imperfections, i.e., crosstalk (one eye's view leaking to another eye's view), brightness and relative color saturation, the time multiplex 3D display performed better than the polarization multiplexed display in terms of the image quality and perceived depth. In [114], an LCD display with polarized glasses, a plasma display with shutter glasses and a projection system with shutter glasses were compared in terms of achievable QoE in different situations. The results showed that the performance of the studied display technologies were comparable in terms of the intensity of 3D effect, depth perception of the scene and user involvement. However, the plasma display with ac-

(a) Non-standard lab environment (b) Public test environment

Figure 2.10: Test environments used in [106].

tive shutter glasses were significantly better in the sharpness of the scene and visual comfort than the other two systems. Depth rendering ability of different displays were studied in [18], and the results showed that it mainly depends on the viewing distance and the properties of display. For example, for line or column interleaved displays, if the viewing distance is 3 times of the screen height, there might be a problem of visible dark stripes because the line or column will exceed the extent of 1 min of arc which is the visual acuity threshold.

Context factors

How dependent are the results of a 3D subjective quality assessment test on the particular viewing conditions used? This is a typical question about the influence of context. In 3D QoE, context influence factors may include the test environment (standard or non-standard lab environment) and terminal equipment (display or mobile phone), etc. Many studies have been conducted on the influence from test environment when considering quality assessment of 2D multimedia content. The main differences from test environments are lighting, background noise, wall color, objects on the wall, etc., which have been standardized in ITU-R BT.500 [58] for controlled lab environment. For example, in [106], the results obtained for 2D audio-visual quality assessment in a non-standard laboratory versus public environment (coffee room) have been compared, as shown in Figure 2.10. The results indicated that the impact of the environment is not significant when a wide range of quality is considered.

Human factors

A recent cross-lab study already confirmed the conclusion that human factor is a predominant influence factor on QoE [106]. Generally, human factors include observer's gender, experience, age, etc. In traditional 2D quality assessment, observer's experience is an important factor which has been defined in ITU-R

BT.500[58] to classify the observer as expert observer and naive observer. Generally, it is considered that experts may generate more consistent results [49] but experience oriented, which may influence the results when the expert observer intend (or not intend) to guess the task or objectives of the test. Naive observers may produce a general results from the perspective of customers. However, when the task of the test is hard to understand or operate, the results would be unreliable. Studies showed that expert observers are more critical of lower quality images/videos than the naive observers [46][117].

Recently, the influence from observer's gender on QoE are gaining more and more attention. For example, in a recent study on 3D QoE of coding videos in IPTV scenarios, the results showed that female observers voted slightly more positive than the male observers though there is no statistical significant difference [140]. The studies in [53] also pointed out that the male and female observers performed differently on the perception of QoE in the context of virtual acoustic environments.

2.4 Subjective assessment for QoE in 3DTV

The complexity of perceiving 3D content as opposed to real-world perception explains the difficulties that naive observers experience when asked to provide an opinion on the QoE of a particular video sample. On one hand they have limited experience with the new technology, notably as opposed to 2D television and, eventually, multimedia content. On the other hand, they may need to counterbalance positive and negative effects such as added depth value and visual discomfort.

2.4.1 Observer context dependency

An observer participating in a subjective assessment experiment cannot be considered isolated from his previous experience and current status. He bases his internal vote on many influence factors which he then expresses towards the outside world, mostly in the form of a vote on a limited scale. Figure 2.11 lists his external experience on the left, notably situations which he has encountered himself, termed "reality", experience with currently available, often wide-spread reproduction technology such as 2D television, and new reproduction technologies such as 3D. He uses his perception towards the goal of analyzing the scene information itself and the perceived artifacts which is the main task that he is asked to perform. However, he also consumes and interprets the perceived visual and eventually auditive information leading to a match or mismatch with his experience in reality. Last but not least, he also takes into consideration his overall feeling, notably his health conditions which may be divided into perception intrinsic factors, i.e. those related to eyesight, and other health factors which may or may not be related to the task at

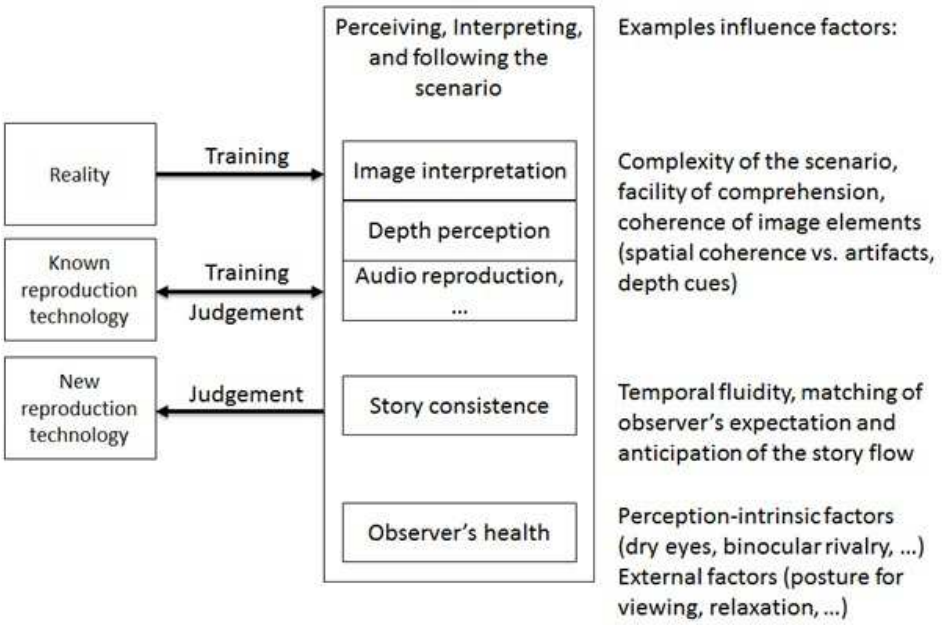


Figure 2.11: Model excerpt for a human observer in a subjective assessment task

hand.

An example of this context dependency is related to one of the major decision factors when introducing 3D services: their advantage over 2D content. From a subjective assessment point of view, the observer's habit to watch 2D content on known reproduction technologies is often misleading their judgment for 3D content shown on the new reproduction technology. Their prejudice may impact in two opposite directions. Often, they judge the 3D content mostly on their trained 2D quality aspects, i.e. perceived coding artifacts, blurring degradations, or reduced resolution, for example when judging 3D content on a vertically view-interlaced polarized display. On the opposite side, some observers overestimate the sensation of depth as a new and exciting experience as part of the so-called hype effect. Comparing 2D to 3D videos will therefore always be context dependent. Even when introducing both media types into a single subjective experiment, observers will likely change the context from presentation to presentation, therefore for example either neglecting or overestimating the added depth value.

The relationship between "observer context dependency" and human factors on QoE can be considered as but not limited to "one belongs to another". Considering the influence of observer context dependency on the assessment of QoE, the observers context dependency can be considered as part of the human factors in QoE of 3DTV, in particular, it works similarly to the observer's experience on QoE of 3DTV. However, considering the objectives of the two terms, "observer context dependency" focuses on its influence on subjective assessment task, which is in particular dependent on the subjective methodology. For example, in a subjective

test that a reference video is available for the viewers, as the viewers can compare the tested videos with the reference video, the "observer context dependency" might affect the test results less than a subjective test without reference. In contrast to "observer context dependency", the human factors "observer's experience" in 3DTV is not constrained to a subjective test task, but represents the influence of observer's previous experience on the current viewing experience. For example, if we have experienced an excellent 3D movie with the most advanced 3D technology, when we watching a new 3D movie in the theater, our viewing experience might be totally different from those who had never watched 3D movie before.

2.4.2 Multi-scale assessment methodologies

A possible solution to express the observer's opinion in complex and eventually conflicting situations concerning his internal representations of quality may be to use multiple scales. The observer may judge one aspect such as the perceived image quality independently from other aspects such as the depth quantity or visual discomfort symptoms. These scales have been proposed in Figure 2.8 as basic 3D quality factors. Several assessment methodologies have been developed to allow for assessing multiple dimensions at once or in separate experiments. Assessing all dimensions in a single experiment facilitates the de-correlation between the scales for the observer, i.e. he decides immediately which effect he assigns to which scale. The advantages of individual experiments with a single scale are the reduced experiment duration and the focus of the observer on a single quality perception aspect, i.e. he does not need to change his voting context. In most cases, one of the three following standardized methods was used:

- Absolute Category Rating with Hidden Reference (ACR-HR): A single stimulus presentation methodology where the observer votes using a fixed number of attributes per scale, such as the five attributes "excellent", "good", "fair", "poor", and "bad" [59]. High quality reference sequences are usually included in the experimental setup to allow for calibration of the observer's voting. Each video sequence is presented only once in random order.

- Double Stimulus Continuous Quality Scale (DSCQS): A double stimulus presentation methodology in which the observer watches two different video sequences with one repetition. One of the two video sequences shall be the reference, the other one a degraded version of this reference. He votes for each of the sequences on a semi-continuous integer scale from 0-100 which may be annotated with attributes for easier comprehension [58].

- Subjective Assessment Methodology for VIdeo Quality (SAMVIQ): The ex-

periment is ordered by video content. For each of the evaluated video contents, a group of degradations, usually 8-12, are presented in such an interactive interface that the observer may watch each one repeatedly. The reference video sequence is available explicitly and shall be evaluated in a hidden manner amongst the degraded versions. When the observer has provided his opinion for each scale and each video, he validates his choices and continues with the next content [8].

The International Telecommunication Union - Radiocommunications (ITU-R) has started a new 3D recommendation in 2012 [56]. Besides the three primary perceptual dimensions "Picture quality", "Depth Quality", and "Visual (Dis)Comfort", it names two additional perceptual dimensions, "Naturalness" and "Sense of Presence". Besides the above mentioned methods ACR and DSCQS, it proposes Pair Comparison and Single Stimulus Continuous Quality Evaluation which is reserved for usage when a single vote for a video sequence is not sufficient but a continuous evaluation is preferred.

All single value voting methods have the drawback that the 3D content display is interrupted after the playback of a single video sequence and a gray frame shall be shown. This distracts the 3D vision on 3D displays such that the observer requires time at the start of the next sequence before perceiving the 3D effect to its full extent [123]. A solution to this has been proposed by using a continuous playback such as a 3D movie film. Intervals that shall be voted for are marked with overlayed numbers and the observer shall provide a vote for the complete interval [42].

2.4.3 Attribute selection

Besides choosing the scales for a subjective assessment, the attributes used for voting need consideration. When using categories in different languages, important differences may occur, leading to the requirement of aligning the scales from one country to another. It was shown that in many languages the currently employed attributes are not equidistant either and that service acceptance thresholds may vary largely [109]. Assuming that the groups of observers in four different languages would vote for a common average value when judging the same video sequences, a numerical fitting of attributes has been calculated based on the attribute positions for the French scale as used by Rumsey et al.[109]. This led to Figure 2.12 which shows the experimental results for the 3D experiments with long bars [5] and the results from [109] with shorter bars. While the usual terms "Excellent", "Good", "Fair", "Poor" and "Bad" are used for both "image quality" and "depth quality", the ITU-R has introduced the scale items "Very comfortable", "Comfortable", "Mildly uncomfortable", "Uncomfortable", and "Extremely uncomfortable" for visual dis-

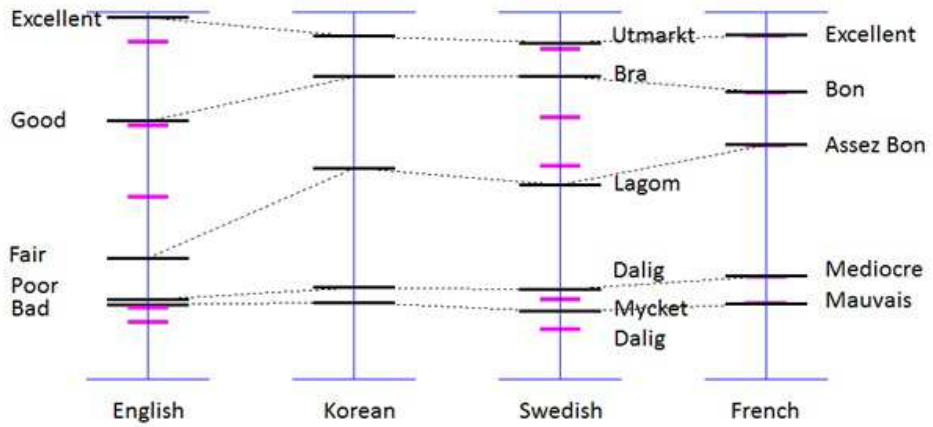


Figure 2.12: Usage of attributes in four different languages under the assumption that the same MOS value would have been obtained. The long bars indicate experimental finding in a 3D QoE experiment[5], the shorter bars represent the positions published in [109].

comfort [56]. The drawback of this scale is that the attributes are hard to associate and to distinguish for untrained observers. A typical observer question would be: "How comfortable is 2D viewing on this scale?".

2.4.4 A possible solution: Paired Comparison

As mentioned above, multi-scale experiments only evaluate a particular quality aspect. In addition, the selection of category descriptions for the scales may alter the meaning of the scales in different situations such as viewing contexts or languages and therefore determining an overall quality remains a challenge. However, Paired Comparison methodology may provide a solution.

The Paired Comparison methodology is already a standardized subjective video quality assessment method for multimedia applications [59]. The observers compare two video sequences to each other and note their preference, e.g., if there are N video sequences, the observers would compare $N(N-1)$ pairs with different presentation order in one pair (Video A first then B, or Video B first then A).

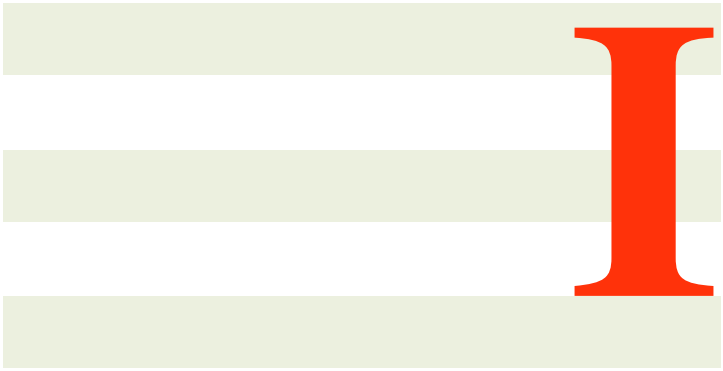
Pair comparison has its advantages to solve the "observer context dependency" and "attribute selection" issues. For example, for a naive viewer who is not used to 3D television, it might be difficult for him to vote on an absolute psychophysical scale for the "viewing experience" of a video sequence. However, when using paired comparison, the question is quite simple for the viewer: "which one do you prefer?". This question contains a combined information from the viewers about the "depth perception", "image quality", "visual discomfort", etc., which is also the definition of QoE in 3DTV. Another important thing is that viewers do not need to consider the scale problems but only judge on a binary value, "this" or "that". Thus, paired comparison might avoid the problems of language dependency and cross-lab

score alignment that occurred for example with ACR method [5].

Due to the possible reliability of paired comparison method on assessment of QoE in 3DTV, one of the objectives of this thesis is focusing on the application of this method on the assessment of QoE, in particular, why this method is reliable, what is the disadvantage of this method compared with other methods, how to improve it, how to apply it in real subjective test, and how to analyze the data, etc. For more details, the readers are referred to Part I of this thesis.

2.5 Conclusions

This chapter introduces the state-of-the-art work on Quality of Experience in S-3DTV. For better understanding, we firstly introduce some basic knowledge on stereoscopic perception and 3D display technology. Then, the definition of 3D QoE is introduced, where a distinction with the traditional concept on 2D "quality" is explained. In particular, the "added depth" perception and "visual discomfort" of S-3DTV are introduced. Due to the "multi-dimensionality" of the QoE in 3DTV, there are plenty of possible factors that might affect it. Thus, the state-of-the-art work on the study of influence factors of 3D QoE is introduced. Finally, one of the most challenging issue in QoE, i.e., subjective assessment methodology is illustrated, including the problems and possible solutions.



Paired comparison methodology: Optimization, evaluation and application

3

Subjective assessment methodology in 3DTV: Paired comparison

Part I of this thesis is focusing on the study of subjective assessment methodology on QoE of 3DTV. As introduced in Chapter 2, the subjective assessment is a challenging work due to the "multidimensionality" of the QoE in 3DTV, the context dependency from observers, and languages, etc. A possible solution called "Paired Comparison" has been mentioned in Section 2.4.4. In this chapter, more details about Paired Comparison are introduced.

3.1 Introduction

In Section 2.4.4, a possible solution for the subjective assessment on QoE of 3DTV called Paired Comparison was introduced. Recently, some studies on subjective assessment methodology in 3DTV have shown the advantages and possibilities of the Paired Comparison on assessment of QoE. Some examples of these studies are shown in the following part.

The first example is about the viewer's behavior in 3D subjective assessment test [35]. In this study, a subjective experiment on 3D video quality and comfort was conducted. The test methodology is the 5-point ACR method. Each observer was asked to provide two scores for visual quality and visual comfort after watching one video sequence. To better understand the behavior of the viewers, four typical histograms of the viewer's result on quality and comfort ratings are selected as shown in Figure 3.1. Observer 10's result is very typical in a subjective test where both the quality score and visual comfort scores are reasonably distributed. Observer 17

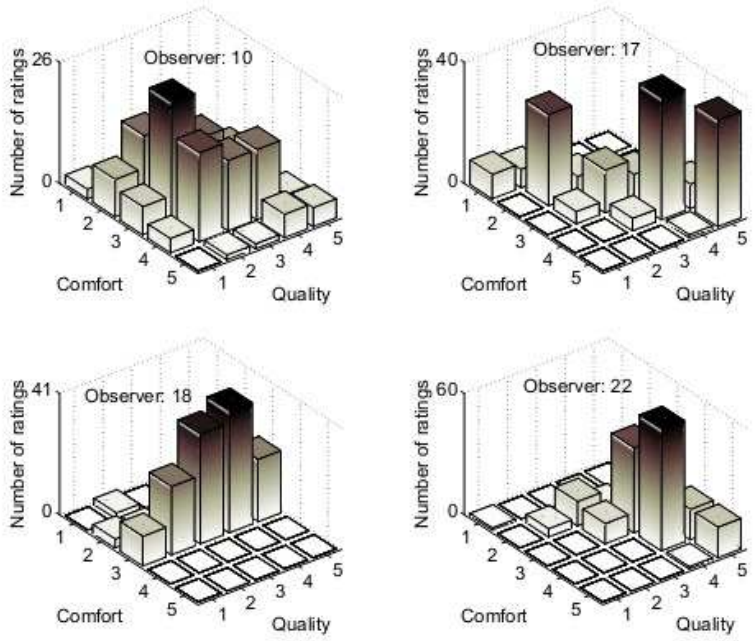


Figure 3.1: An example of 4 observers' results on co-joint quality and visual comfort from [35].

might not be able to distinguish visual comfort and overall quality very well because the comfort and quality scores are highly correlated. Observer 18 seems to utilize the quality scale very well but for visual comfort, only the middle part of the scales was utilized. Observer 22 only used a very small range of ratings to evaluate both 3D video quality and comfort. As the two scores were provided by the observers simultaneously in one individual test, observers may feel confused at some time and the possibility of voting dependency may be increased, which might affect the results. This study shows the possible problems of single-scale rating based assessment methodologies on QoE of 3DTV.

Another example is about the comparison on discriminability of the two test methodologies on 3D image quality assessment [83]. The subjective assessment of 3D image quality was conducted using SS (*Single Stimulus*) and Paired Comparison methods in two separate experiments with different viewers. To evaluate the discriminability of the two methodologies, the SS results were converted to paired comparison data. Then, the averaged preference difference of all pairs were calculated for both methodologies as the discriminability measures. The results are shown in Figure 3.2. It is indicated that besides the video "sof."(a video content), paired comparison always outperforms the SS methodology in terms of the discriminability. This conclusions is also supported by another subjective experiment about 3D synthesized view assessment in [9].

The discriminability difference between the two test methodologies in these studies might be influenced by two factors: the test methodology itself and ob-

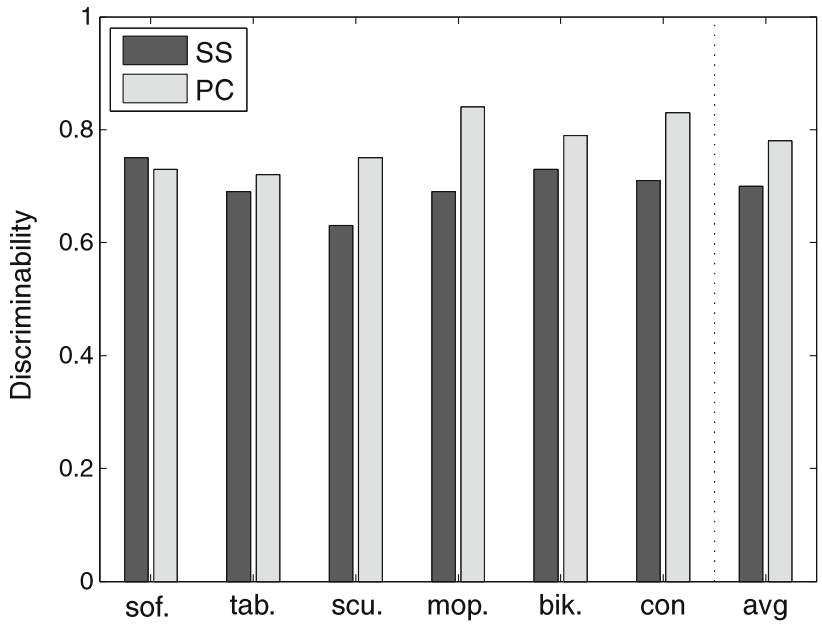


Figure 3.2: Discriminability measures of the SS and paired comparison (PC) methodologies in [83]. The x-axis represents different video contents. The last column is the averaged value for all video contents.

server's context dependency. In paired comparison, as viewers are asked to make a selection in a pair, viewers would find the most significant factors that induce the difference in a video pair. However, in a SS test, viewers only watch one video at one time, if without large number of training in 3D watching, viewers may feel difficult to point out the ratings of the test videos.

The examples listed above illustrate the possibility of the Paired Comparison methodology on the assessment of QoE in 3DTV. In the following sections, the state-of-the-art research work on Paired Comparison are introduced, including the standardized methodology, the disadvantages of Paired Comparison, some existing improved designs on it, as well as the data analysis methods.

3.2 Standardized Pair Comparison method

3.2.1 Definitions

Paired comparison method is already a standardized subjective video quality assessment method for multimedia applications [59], and has been adopted as one of the standardized subjective assessment methodologies for stereoscopic 3DTV systems in ITU-R BT.2021 [56].

In Paired Comparison, if there are m test stimuli (i.e., images or videos in quality assessment tests), 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 consider-

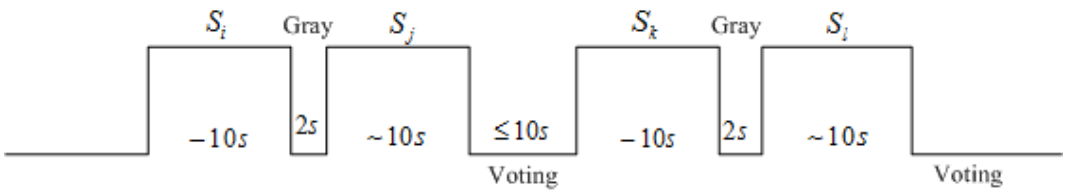


Figure 3.3: Stimulus presentation in a pair comparison experiment [59].

ing the displaying order, all the pairs should be displayed in both presentation orders (e.g. $\{S_1S_2\}$, $\{S_2S_1\}$). The number of combinations will raise to $N = m(m-1)$ for one observer. The presentation order for a pair may be either time parallel, i.e. on two screens, or time sequential, i.e. on one screen. 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.

To avoid confusing with some other paired comparison designs which will be introduced later, in this thesis, we name the Paired Comparison method as "Full Paired Comparison (FPC)" because all possible pairs are compared in this method.

According to ITU-T P.910 [59], the time pattern for the stimulus presentation can be illustrated by Figure 3.3. The voting time should be less than or equal to 10 seconds, depending upon the voting mechanism used. The presentation time should be about 10 seconds and it may be reduced or increased according to the content of the test material.

3.2.2 Comparison with other test methodologies

In a typical subjective quality assessment experiment, there are usually different source video contents (SRC) with different types of degradations under test (HRC, *Hypothetical Reference Circuit*). For example, the HRCs in a test might be different coding schemes which would induce different levels of image/video quality.

It should be noted that in a FPC test for quality assessment, usually, only the stimuli with same SRC are compared. Thus, the test procedure described above is for different HRCs under one SRC. Of course, if video content is the objective of the test, there is no such limitations any more.

To have a better understanding of different subjective assessment methodologies, a comparison of test duration between different test methodologies is conducted. Assuming there are n SRCs and m HRCs in a test, in HRC, there is one condition without inducing quality degradation, i.e., reference sequence. Each video sequence lasts 10 seconds, between each two video sequences there is a gray image lasts 2 seconds. The voting time for viewers is 5 seconds. For a single stimulus test, the duration of the test is $T_{ss} = n \times m \times (10s + 5s) + (n \times m - 1) \times 2s$. For a test using DSCQS, as the reference sequence and the degraded sequence

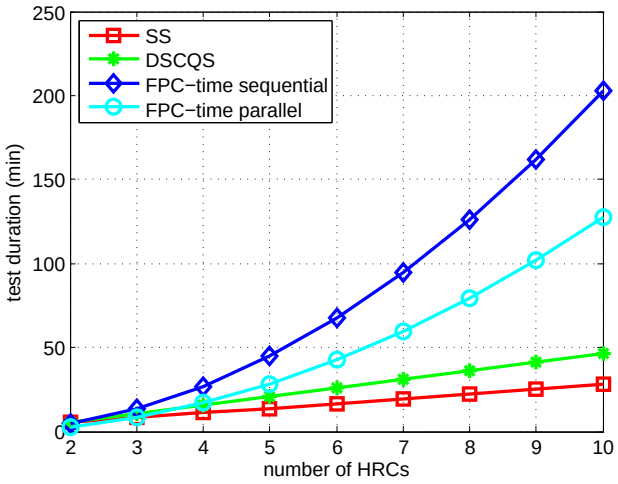


Figure 3.4: Comparison of the test duration between different test methodologies.

will be presented in one group and one repetition is required, the duration of the DSCQS test for one SRC is $T_{dscqs\_src} = m \times (10s \times 2 + 2s \times 3 + 5s)$. For n SRCs, the duration is $T_{dscqs} = n \times T_{dscqs\_src}$. For FPC method, there are in total $n \times m \times (m-1)/2$ pairs, the test duration for time-sequential method is $T_{fpc\_seq} = (n \times m \times (m-1)/2) \times (10s + 2s + 10s + 5s)$, and for time-parallel method, $T_{fpc\_parallel} = (n \times m \times (m-1)/2) \times (10s + 2s + 5s)$. For better visualization, a comparison is shown in Figure 3.4 with n = 10 and various m conditions.

3.2.3 Disadvantages of FPC method

Figure 3.4 clearly shows that compared to the SS and DSCQS methodologies, the FPC method is much more time consuming. With the increase of the number of stimuli, the number of comparisons increases exponentially and for the cases of large number of HRCs, the test becomes infeasible. For example, if there are 10 HRCs, the test duration for one observer is about 200 minutes. Thus, though the Paired Comparison method has its advantages in resolving the possible problems in subjective assessment of QoE in 3DTV, in most cases, this method is not applicable.

To solve this problem, some studies are conducted theoretically or experimentally. In the following sections, the state-of-the-art research work is introduced.

3.3 Paired comparison designs: the state of the art

The severe drawback of the FPC method has impeded its application significantly. Designs are therefore required which could reduce the number of comparisons without serious imbalance [62]. Generally, the designs are classified into non-adaptive and adaptive methods. For non-adaptive methods, each subject com-

pares a subset of the whole set of the pairs, but for all subjects the comparisons are balanced [144] [32]. In the adaptive methods, the pairs are selected adaptively according to the results of all previous observations, and these pairs are considered to be more efficient than other pairs to generate accurate results [41] [113]. In the following part, the designs for pair selection are introduced.

3.3.1 Randomised pair comparison design

Randomised Pair Comparison (R/PC) [33] is an economic extension to traditional pair comparison designs without considering balance of the stimulus or optimization of the selection of the pairs. The basic idea of the R/PC method is to divide all combinations of the FPC test pairs equally and randomly to each participant. Strictly speaking, this method is not an efficient designs, which means the total number of comparisons is not reduced. The reduction is only for each observer. However, it is quite applicable for the study which is based on crowdsourcing.

In a subjective experiment (e.g., video quality assessment), generally, the test stimuli would be designed with different factors (e.g., blur and blockness) and with different levels (little, medium, much). In R/PC design, some new definitions for pairs are proposed:

- *contrast pairs*: It is defined as any two stimuli in a pair which may differ in one or multiple factors. Usually, they are the combination of all factors and levels.

- *Reference conditions:* they are used to find unreliable assessors and outliers, thus, they are corresponding to the contrast pairs, including
  (1) *equal reference pairs*: they contain the same video at the same quality level twice. Thus, for every factor/level combination, there should be an equal reference pair.
  (2) *matched contrast pairs*: they just differ in the presentation order of the contained contrast pairs.

In a R/PC test, the subset size s, i.e., the number of pairs for each assessor should be pre-defined. The duration for each assessor should be equal. Assuming there are in total p contrast pairs with $\{S_1S_2\}$ order, and p matched contrast pairs with $\{S_2S_1\}$ order, and corresponding e equal reference pairs, then, for each assessor, $s(p/(2p+e))$ matched contrast pairs and $s(e/(2p+e))$ equal reference pairs are selected. The selection of the pairs for each assessor should be randomly. In this way, each assessor has a unique random subset of pairs, these pairs are randomly presented and ensures that:

1. The ratio of contrast to reference pairs is equal in each subset and equal to the ratio in a full design;
2. Each selected *contrast pair* is contained in both possible presentation orders (both matched contrast pairs $\{S_1S_2\}$ and $\{S_2S_1\}$ are presented);
3. *Equal reference pairs* correspond to selected *contrast pairs* (there is no reference video which does not occur in a contrast pair as well);
4. *Equal reference pairs* are contained only once, i.e., no repetitions for equal reference pairs.

As mentioned before, the design of R/PC would lead to unbalanced data. For example, the number of occurrence of each stimulus may be not balanced. Furthermore, as this method may be used for web-based studies, if assessors quit the test in the middle of the test procedure, the whole data would not be balanced as well. In this case, the traditional statistical tools, e.g. ANOVA (*ANalysis Of VAriance*) or GLMs (*Generalized linear model*) methods may not be suitable for analysis. To solve this problem, a general framework called HRRG (*HodgeRank on Random Graphs*) is proposed to analyze the imbalanced and incomplete data in randomized paired comparison experiments. For more details the readers are referred to [146].

3.3.2 Sorting algorithm based design

In [113], the authors analyzed the characteristics of paired comparison and found that comparisons between very distant samples do not provide the estimation of the distance as accurate as the nearby samples. For example, if there are three stimuli with quality ascending ordering: S_1, S_2, and S_3, to measure the distance between them, it's better to ask the viewers to select their preference on $\{S_1S_2\}$, and $\{S_2S_3\}$ rather than $\{S_1S_2\}$ and $\{S_1S_3\}$. Because in the condition of $\{S_1S_3\}$ as their difference on quality levels are more obviously for viewers, in an extreme condition, viewers may always chose one to another, which lead to a binary result, i.e., 0% viewers selected S_1 than S_3 in terms of higher quality. This binary data (0 or 1) will induce infinite estimation errors on the distance between S_1 and S_3. Thus, it is concluded that comparison should be concentrated on closer pairs. For more details about the calculation from the proportion data to distance values between each pair, the readers are referred to Section 3.4.

Based on the analysis in [113], the authors proposed to apply the sorting methods on paired comparison because firstly, the efficient sorting algorithms which are based on comparing two elements at a time (e.g., Quicksort, Heapsort) require $m\log_2 m$ rather than m^2 comparisons between samples. Secondly, a sorting algorithm must include comparisons between nearest samples. Thus, it assures the comparison between each nearest set of samples, and fewer comparisons between more distant samples.

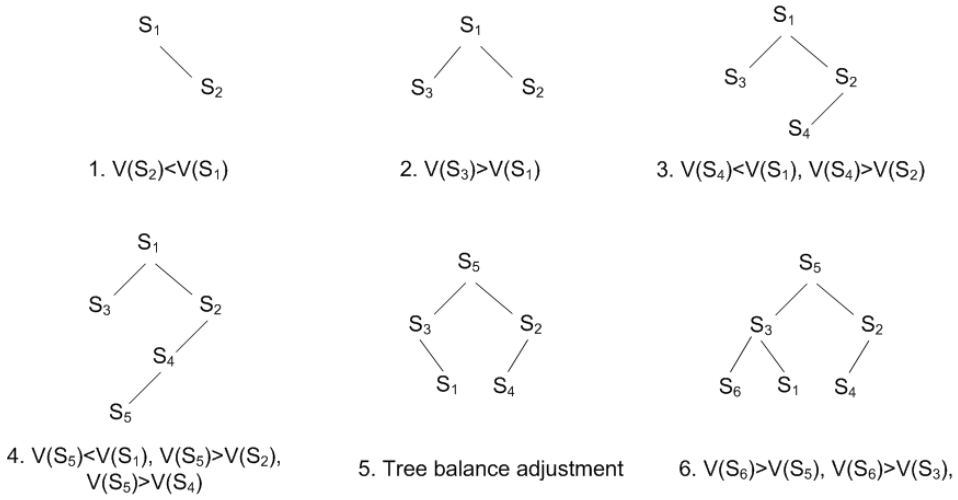


Figure 3.5: An example of a binary tree sorting for pair comparison. S_1 - S_6 are stimuli. Step 1 to 4 are binary sorting. Step 5 is reconstruction for the balance of the tree. Step 6 is for another added stimulus. V represents quality of the stimulus.

An example of using a binary tree sorting method is introduced here. A binary tree can be constructed with the stimuli as the nodes. Each node of the tree is a partitioning element for a left sub-tree and a right sub-tree. The left sub-tree consists of nodes which were judged to be higher in quality, and the right sub-tree consists of nodes which were judged to be lower in quality. During the comparison process, the stimulus is always compared from the root node. If there is no root node, this stimulus is considered as root. If this stimulus is judged as lower quality to the current node, it is then moved to the right sub-tree and compared with the root of the sub-tree, otherwise, it is moved to the left sub-tree recursively. To improve the efficiency of the sorting, a balance process is added after each comparison, which means reconstructing the tree to make it as short as possible and having as few nodes at the bottom as possible. An example of this process is shown in Figure 3.5, where V represents the quality of that stimulus.

3.3.3 Balanced sub-set design

Since it is unwieldy to run all pairs in paired comparison method, one possible way is to omit some pairs completely. Dykstra [32] proposed a "balanced sub-set" method, which means that for certain pairs $\{S_iS_j\}$, the number of comparison n_{ij} is 0 while for all other pairs it is a constant, i.e., $n_{ij} = n$. Each stimulus 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".

Group divisible designs (GDD)

Supposing there are m stimuli S_1, S_2, ..., S_m, $m = t_1 t_2$, where t_1 and t_2 are integers. The GDD method is constructed by divide the m stimuli into t_1 groups, and each group has t_2 elements. The pairs within a group are not compared. For example, $m = 6$, $t_1 = 2$ and $t_2 = 3$. The arrangement of the indices are as follows:

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

S_1, S_2, and S_3 are in one group and S_4, S_5, and S_6 are in another group. Since the pairs within one group are not compared, the pairs that need to compare are $\{S_1S_4\}$, $\{S_1S_5\}$, $\{S_1S_6\}$, $\{S_2S_4\}$, $\{S_2S_5\}$, $\{S_2S_6\}$, $\{S_3S_4\}$, $\{S_3S_5\}$ and $\{S_3S_6\}$. Each stimulus appears 3 times in the whole procedure.

Following this rule, the total number of comparisons for GDD method is $(t_1 \times (t_1 - 1)/2) \times {t_2}^2$.

Triangular designs (TD)

If the number of the stimuli m can be expressed as $t(t-1)/2$. The association scheme is given below for $m = 10$.

$$\mathbf{R} = \begin{bmatrix} X & 1 & 2 & 3 & 4 \\ 1 & X & 5 & 6 & 7 \\ 2 & 5 & X & 8 & 9 \\ 3 & 6 & 8 & X & 10 \\ 4 & 7 & 9 & 10 & X \end{bmatrix}$$

The indices of the stimuli are arranged in a square matrix of size t with the diagonal empty. The indices are symmetrically around the diagonal. There are two implementations for the TD method:

- Case 1: Only the pairs whose indices are in the same column are compared; i.e., $\{S_1S_2\}$, $\{S_1S_3\}$, $\{S_1S_4\}$, $\{S_2S_3\}$, ..., $\{S_9S_{10}\}$. Each stimulus appears 6 times in the whole procedure.
- Case 2: For the condition of $t > 4$, only the pairs whose indices are not in the same column are compared; i.e., $\{S_1S_8\}$, $\{S_1S_9\}$, $\{S_1S_{10}\}$, $\{S_2S_6\}$, $\{S_2S_7\}$, $\{S_2S_{10}\}$, $\{S_3S_5\}$,..., $\{S_8S_{10}\}$. Each stimulus appears 3 times in the whole procedure.

Following this rule, the total number of comparisons for TD method is $\frac{(t-1)(t-2)}{2}t$ for case 1, and $\frac{m(m-1)}{2} - \frac{(t-1)(t-2)}{2}t$ for case 2.

Square Designs (SD)

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 **R**, a example is shown as follows:

$$\mathbf{R} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix}$$

Only the stimuli whose indices are in the same column or row are compared. Thus, in this example, the pairs $\{S_1S_2\}$, $\{S_1S_3\}$, $\{S_1S_4\}$, $\{S_2S_3\}$, $\{S_2S_4\}$, $\{S_3S_4\}$, $\{S_1S_5\}$,..., $\{S_{12}S_{16}\}$ are compared. Each stimulus appears 6 times in the whole procedure.

Following this rule, there are in total $t^2(t-1)$ pairs for the SD method.

Comparison between the balanced sub-set designs

The balanced sub-set designs can be employed based on the number of stimuli and the efficiency of different designs. For example, in the condition of m=16, the GDD method and the SD method can be used. Which one is better is determined by its efficiency. For better visualization, the comparison between the number of trials (number of pairs) for all designs are shown in Figure 3.6.

The figure shows that the number of comparison does not necessarily increase with the number of the stimuli, which highly depends on the selection of the designs. For example, for the condition of $m = 16$, the GDD method (2×8) will generate more comparisons than the SD method (4×4).

3.4 Pair comparison models

The outcome of a paired comparison test 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 stimulus pair $\{S_iS_j\}$ is $n_{ij} = a_{ij} + a_{ji}$. Pair comparison models are mathematical tools to convert the pair comparison data to scale values for all stimuli. Meanwhile, the corresponding confidence intervals, goodness of model fit and some statistical hypothesis tests are also provided. In the following sections, the widely used Thurstone model and Bradley-Terry (BT) model, and a seldom used Elimination By Aspects (EBA) model are introduced, where the Bradley-Terry model is a special case of this model.

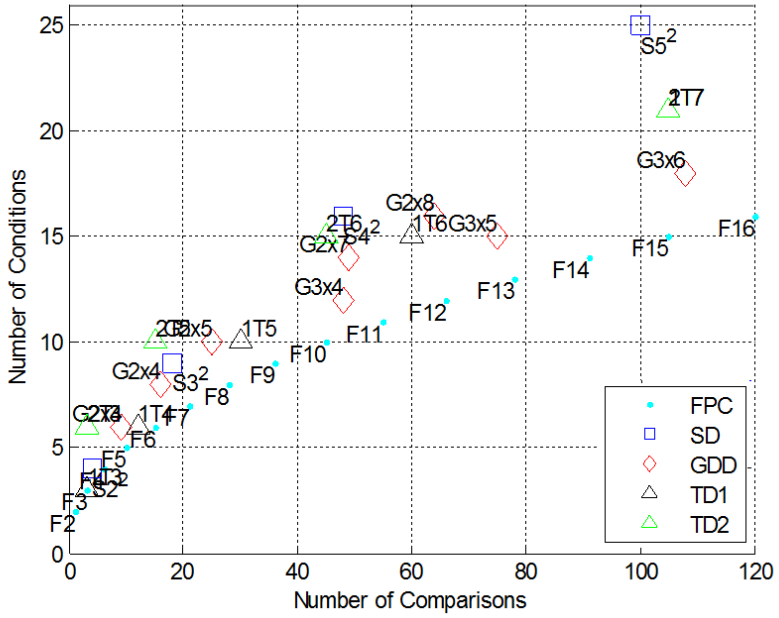


Figure 3.6: Comparison of the trial numbers for different methods. Different markers represent different methods. The number close to the markers represents the form of the matrix R. In particular, the TD has two implementations, "1T" corresponds the case 1 and "2T" represents case 2. "1T5" represents Triangular design of Case 1 with $t = 5$.

3.4.1 Thurstone model

The Thurstone model is a widely used tool where the relationship between the paired comparison procedure and the converted continuous scales origins from psychophysics. This model is based on the idea that "a given physical stimulus does not always produce the same psychological experience". The experienced scale in Thurstone model follows a Gaussian distribution[127][126].

An example will be provided in order to illustrate the conversion process. A paired comparison experiment with two stimuli S_i and S_j is considered. For each stimulus, the observed Quality of Experience X_i and X_j follow Gaussian distributions on psychophysical scales. Their distributions are shown in Figure 3.7.

For illustration it has been chosen that the mean and standard deviations of the QoE of stimuli S_i and S_j are different. The effect that observers are more undecided on the stimulus S_i than on the stimulus S_j was modeled by choosing standard deviations of σ_i is larger than σ_j. An observer samples each of the distributions and obtains a QoE value for stimulus S_i, referred to X_i and for stimulus S_j, referred to X_j. If $X_i < X_j$, the observer would prefer stimulus S_j (condition 1 in Figure 3.7), if $X_i > X_j$, then the observer prefers condition S_i (condition 2). From the diagram in Figure 3.7, it may be seen by comparing the mean value of the two distributions, that the probability of obtaining $X_i < X_j$ is larger, therefore more observers will vote

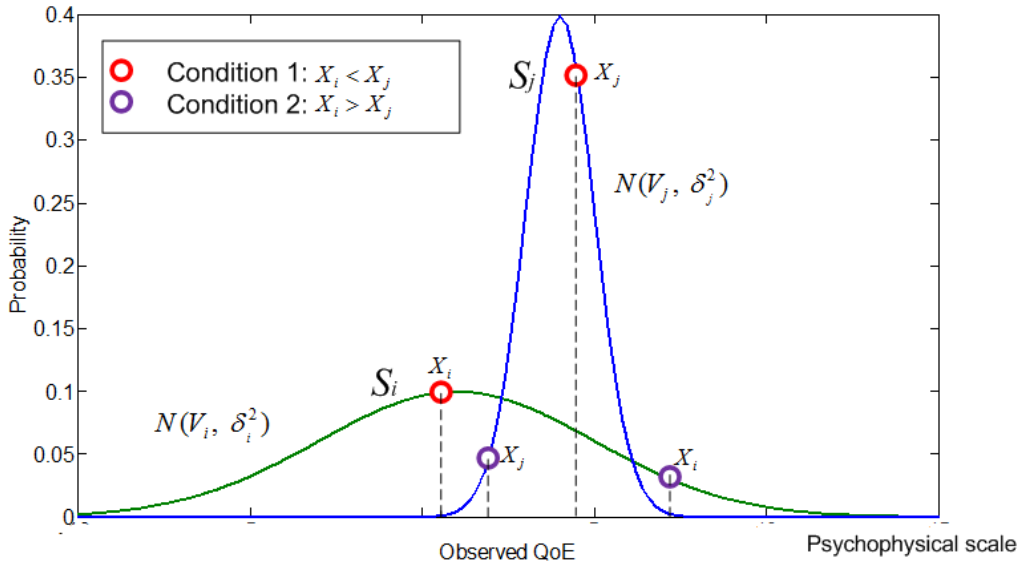


Figure 3.7: An example of the distributions of the experienced QoE for two stimuli.

for stimulus S_j as a preference. Generally, it is assumed that an observer chooses deliberately on each of the conditions. However, in some cases, this may not hold true. Let us assume that 3D is largely preferred to 2D, then a comparison between a pair containing two 2D stimuli or a pair containing two 3D stimuli will be judged in a different manner than a pair containing one 2D and one 3D sequence. This can be modeled by introducing a correlation coefficient γ_{ij}. According to the law of normal distribution, the differences of the mean value of the two distributions can be calculated by Equation 3.1, where $\Phi^{-1}(P_{ij})$ is the inverse function of normal cumulative distribution. P_{ij} is the probability that stimulus S_i is preferred to S_j.

$$V_i - V_j = \Phi^{-1}(P_{ij}) \sqrt{\sigma_i^2 + \sigma_j^2 - 2\gamma_{ij}\sigma_i\sigma_j} \tag{3.1}$$

After having obtained the difference between a sample taken from stimulus S_i and a sample from stimulus S_j, the probability of obtaining a particular difference value on QoE is dependent on the distributions of S_i and S_j. This can be seen graphically in Figure 3.8. A paired comparison experiment is usually performed as a forced choice test, therefore a threshold decision is made. In other words, if people experience the slightest positive difference, they are voting for S_i, if they experience the slightest negative difference, they are voting for condition S_j. In Figure 3.8 the area under the probability curve has been marked that corresponds to the cumulative probability of having a preference for condition S_i, that is X_i - X_j > 0.

According to different test scenarios and based on the simplicity and assumptions, Equation 3.1 is classified into five cases [126]. More assumptions adopted

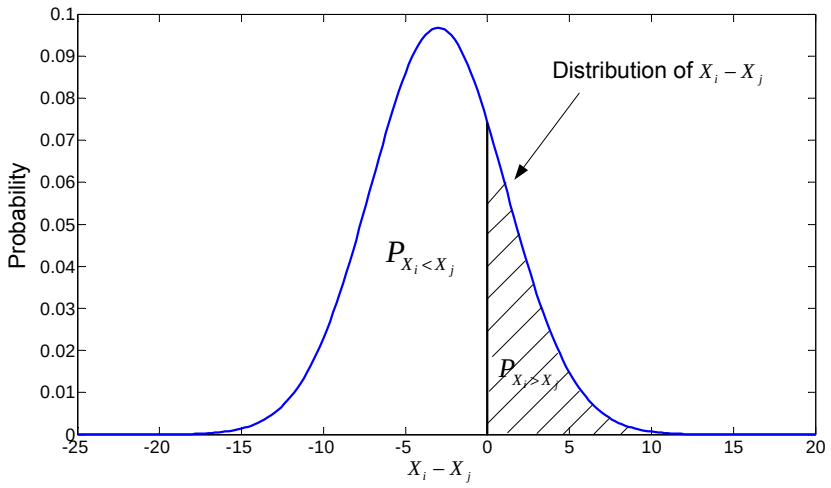


Figure 3.8: Probability of obtaining a particular difference value in terms of PoE scale value when judging condition X_i vs. X_j.

would induce simpler form of this equation. The known parameters in the pair comparison experiment is the proportion that one stimulus is preferred to another. The unknown parameters are the mean values, standard deviations and the correlation coefficients of the two test stimuli. Generally, the mean value is considered as the QoE scale values that we need.

- Case I: The equation is the same as Equation 3.1, i.e., the complete form. In this case, we assumed that the correlation coefficient γ_{ij} is a constant for each pair, i.e., $\gamma_{ij} = \gamma$. Case I is for the case that only single observer conducted the pair comparison test, but for several times.
- Case II: The equation form is exactly the same as Case I. However, this case is used to explaine the condition that a group of observers conducted the pair comparison test and each observer only conducted once.
- Case III: This case assumes that the correlation coefficient γ is zero, i.e., in this case, the test stimuli should be very homogeneous with no distracting attributes. The perceived quality of one stimulus will not have influence on the perceived quality of another stimulus. The equation then has the following form:

$$V_i - V_j = \Phi^{-1}(P_{ij})\sqrt{\sigma_i^2 + \sigma_j^2} \tag{3.2}$$

- Case IV: If assuming that the standard deviations are not subject to gross variation, and they have linear relationship such as follows:

$$\sigma_j = \sigma_i + d \tag{3.3}$$

where d is a very small value, probably a function of σ_i such as 0.1 to 0.5.

Then, the Thurstone model case IV is:

$$V_i - V_j = 0.707\Phi^{-1}(P_{ij})(\sigma_i + \sigma_j) \tag{3.4}$$

– Case V: According to Thurstone, if σ_i = σ_j is a constant, then, the equation has the following form:

$$V_i - V_j = \sqrt{2}\Phi^{-1}(P_{ij}) \tag{3.5}$$

However, Mosteller pointed out that without the strict restriction that γ = 0 and σ_i is a constant, we can also get this simple form [96]. In addition, Mosteller made a great contribution on the mathematical solution and statistical analysis on this model, e.g., transforming the observed ranking patterns to patterns of binary paired comparisons, obtaining the normal deviate corresponding to the mean of each binary variable, and estimation of the parameters of the models by using the least square estimation. Thus, the Thurstone model case V is also called "Thurstone-Mosteller" model.

3.4.2 Bradley-Terry model

The Bradley-Terry model is based on a different idea to the Thurstone model. Supposing there are m stimuli. V_i and V_j represent the "perceptual score" of the stimuli S_i and S_j, respectively. In a psychophysics setting, for example, in the visual discomfort subjective experiment, V_i represents the degree of visual discomfort on a hypothetical psychological scale. The observed score of object S_i is represented by the random variable X_i owing to observation-to-observation variation [45]. Then, the probability that X_i is larger than X_j can be defined as Equation 3.6.

$$P(X_i > X_j) = \pi_{ij} = \pi_i/(\pi_i + \pi_j) \tag{3.6}$$

where

$$\pi_i \geq 0 \qquad \sum_{i=1}^{t} \pi_i = 1 \tag{3.7}$$

The value V_i can be estimated by v_i, which can be calculated as follows:

$$v_i = \log(\pi_i) \tag{3.8}$$

The model defined above is the Bradley-Terry model. 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.

Based on the Bradley-Terry model, it would be found that the scale value v_i is not an absolute value which is dependent on the number of stimuli. However, the

distance between each two stimuli $v_i - v_j$ is an absolute value because:

$$v_i - v_j = \log \frac{\pi_{ij}}{1 - \pi_{ij}} \tag{3.9}$$

that means, $v_i - v_j$ is related to the probability that stimulus S_i is preferred to S_j which is an independent value.

Thus, for the Bradley-Terry model, one of the v_i is set as the reference, i.e., the Bradley-Terry score is set to 0, then, the distance of other stimuli to the reference can be calculated. Some literatures also use π_i as BT scores. In this case, it should be noted that this value is a ratio scale value, i.e., one of the π_i should be set as a reference value, then other stimulus's BT score is meaningful and independent only by calculating the ratio π_i / π_{ref}, which indicates the preference probability between these two stimuli.

Besides the scale values for all stimuli, the Bradley-Terry model can also provide confidence intervals, goodness of model fit and a series of hypothesis test. For more details, the readers are referred to [10][11].

3.4.3 EBA model

The EBA model is generally used to cope with subgroups consisting of similar stimuli. According to the EBA model, a subject prefers one stimulus over another because of a certain attribute this stimulus has that the other one does not have. 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.

To explain it in a formal way, let S ={S_1, S_2, S_3, ... } be the test stimuli. For each stimulus, it has a set of attributes, $S_i' = \{\alpha, \beta, \gamma, ...\}$. Different stimuli may have different attributes. According to the EBA, the probability of choosing S_i from the pair {S_i,S_j} is

$$P_{ij} = \frac{\sum\limits_{a \in S_i' \setminus S_j'} u(a)}{\sum\limits_{a \in S_i' \setminus S_j'} u(a) + \sum\limits_{b \in S_j' \setminus S_i'} u(b)} \tag{3.10}$$

where $u(a)$ is the ratio scale value of attribute a. $S_i' \setminus S_j'$ is the set of attributes that stimulus S_i has but S_j does not have. The scale value of stimulus S_i is the sum of its attributes, i.e.,

$$V_i = \sum_{a \in S_i'} u(a)$$

In this case, the Bradley-Terry model can be considered as a special case of EBA

model which only has one attribute for each stimulus.

3.4.4 Goodness of model fit

After applying the paired comparison model on the paired comparison data, the scale values for all stimuli are obtained. Usually, it is recommended to know how well the paired comparison model actually reflects the raw paired comparison data. One statistical test that addresses this issue is the chi-square test for goodness of model fit. The details about the chi-square test are not introduced here, readers are referred to [43].

It should be noted that if the chi-square test shows that the paired comparison model fails in explaining the data, i.e., the p value is less than 10% or 5% (depending on the significance level), the scale values converted by paired comparison model could not be used for further analysis. Some other statistical tools on analyzing the raw paired comparison data are necessary in this case. In Chapter 4, these methods are introduced and some novel methods are proposed.

3.5 Conclusions

This chapter is an overview of the state-of-the-art research work on paired comparison methodology. Due to the multi-dimensionality of the QoE in 3DTV, a more reliable method is necessary for the subjective assessment. In this chapter, firstly, the motivations and possibilities that why the Paired Comparison method is applicable in subjective assessment of QoE in 3DTV are introduced. However, due to its severe drawback, i.e., with the increase of the number of test stimuli, the number of comparison increases exponentially, it is usually unfeasible to conduct the paired comparison test with a reasonable number of stimuli. To resolve this problem, there are some studies focusing on the design of efficient paired comparison methods. In this chapter, the state-of-the-art efficient designs are introduced. They are applicable in different test scenarios. For example, if in a crowdsourcing study, the R/PC method might be suitable. For a condition that the number of the stimuli is a squared number, then, the SD method, or the sorting based design might be suitable.

The outcome of the paired comparison test is a paired comparison matrix. There are some different mathematical tools to convert the paired comparison data to scale values for all the stimuli. Thus, in this chapter, the widely used pair comparison model, i.e., Thurstone model, Bradley-Terry model and Elimination By Aspects model are introduced.

Based on the advantages of different efficient designs on paired comparison, and the mathematic tools for paired comparison introduced in this chapter, a set of new designs are proposed, which will be introduced in next chapter (Chapter 4).

4

Boosting paired comparison methodology: optimization on the balanced sub-set designs

In previous chapter, we introduced several different pair comparison designs and the mathematical tools to analyze the paired comparison results. In this chapter, a set of new efficient pair comparison designs is proposed. They are applicable in different test scenarios. The proposed designs are evaluated and compared with the FPC method and some other efficient designs introduced in Chapter 3. Furthermore, this chapter provides some guidelines for the readers about how to use the proposed method to conduct the subject experiments, including the selection on the number of stimuli, the number of observers to achieve reliable results and the presentation order for test stimuli, and finally, some novel statistic tools to analyze the results.

4.1 Introduction

In Chapter 3, some designs for reducing the number of pairs were introduced. These designs have their own advantages and disadvantages. For example, in R/PC method, the pairs are not optimized selected but randomly divided into several sub-sets and all participates fulfil one whole observation. However, this method is suitable for the crowdsourcing based study. The sorting based algorithm has optimized the selection of the pairs, i.e., the pairs which would generate the most accurate results are selected rather than the non-efficient pairs. However, the frequency of the occurrence of each stimulus is not balanced. Some stimuli might appear more

times than others, which would induce observer's bias on the selection. Dykstra's balanced sub-set paired comparison designs have resolved this issue. However, the sub-set pairs are not selected optimally.

Thus, it is quite necessary to develop a paired comparison design in which the pairs are selected optimally and the occurrence of each stimulus is balanced. Furthermore, most of the existing designs were developed based on perfect theoretic analysis, some issues in real subjective experiments were not taken into account. For example, what if the observers made mistakes on voting during the test? How the system errors affect the test results? A new efficient design should also be able to resolve these problems.

In this chapter, based on the balanced sub-set paired comparison designs, a set of optimized designs is proposed. Firstly, the Bradley-Terry model is used to investigate the redundancy of the pairs and the possibility to optimize the selection of the pairs in the test. Then, based on the analysis, some implementations for optimizing the balanced sub-set pair comparison designs are proposed. Furthermore, in order to generate a more reliable result, the test procedure and some constraints are defined. Finally, some statistical analysis methods for analyzing the paired comparison data are proposed. All the mentioned above can be considered as a complete system of the proposed paired comparison methodology. To evaluate the proposed designs, some Monte Carlo simulation experiments are conducted to mimic the real subjective experiments where there are observation errors. The performance of the FPC, sorting based design, original balanced sub-set designs and the proposed optimized designs are compared and evaluated.

4.2 Analysis on the selection of test pairs

Most of the existing efficient designs are based on "perfect" test condition, i.e., no error occurs during the test. However, this does not hold true in real subjective experiments. To propose an efficient design, it is highly demanded to know the errors that might happen in subjective experiments:

– Systematic errors: They are biases in measurement which lead to the situations where the mean of many separate measurements differs significantly from the actual value of the measured attribute. All measurements are prone to systematic errors. Systematic errors shift the results always in one direction and much harder to estimate. In paired comparison test, systematic errors may come from the display or test environment, etc., which might have influence on the perception of the observers and thus affect the selection results.

– Observation errors: They are related to errors induced by the observation process. Based on the source of the error, there are two categories:
1) Observer's selection error: They are related to observers' response errors. For example, in a test, the observer might press the wrong button unintentionally. This "mistake" would affect the results.
2) Sample size induced errors: They are related to the number of observations in the test. Large sized sample leads to increased precision in estimates of the "true" value.

To analyze the influences of these errors on the estimation results, the relationship between the paired comparison data and the estimated Bradley-Terry scale values is briefly repeated here. For two stimuli S_i and S_j, P_{ij} is defined as the probability that stimulus S_i is preferred to stimulus S_j, then, the distance between the quality of the two stimuli D_{ij} could be calculated according to the Bradley-Terry model [11][10]:

$$D_{ij} = \log P_{ij} - \log(1 - P_{ij}) \tag{4.1}$$

The sources of the systematic errors, observation errors and sample size induced errors are different, however, they affect the observed paired comparison results similarly, i.e., errors are added on P_{ij}. Here, we take the observation error as an example.

4.2.1 Observer selection errors

Supposing that there are N observers in a paired comparison test, for the stimulus pair S_i and S_j, m observers prefer S_i to S_j, then the ratio $p_{ij} = m/N$ is taken as the likelihood estimation of the preference P_{ij}. However, if one of the observers provided a wrong vote, i.e., he planed to select S_i but he pressed the wrong button, the influence of this error on the estimation of D_{ij} would be dependent on the "true" preference probability value P_{ij}. Here we give an example.

In the condition of two stimuli with distinct quality levels, for example, m=1, N=10, p_{ij} = 0.1. One of the observers made a mistake in the selection, i.e., the p_{ij} in fact should be 0.2 (we do not consider other errors in this example). Thus, according to Equation (4.1), the estimated distance between stimuli (S_i, S_j) should be 1.4 but the observation error makes it 2.2. The amount of estimation error is 0.8. In the condition of two stimuli with very close quality levels, supposing m=4, N=10, p_{ij} = 0.4. Similarly with the previous example, one observer made a mistake and the true p_{ij} value is 0.5. The estimated distance between stimuli (S_i, S_j) should be 0 but the observation error changes it to 0.4. In this case, the amount of estimation error is 0.4. From these two examples it could be found that the same observation error would have different influence on the estimation of the distance which is dependent

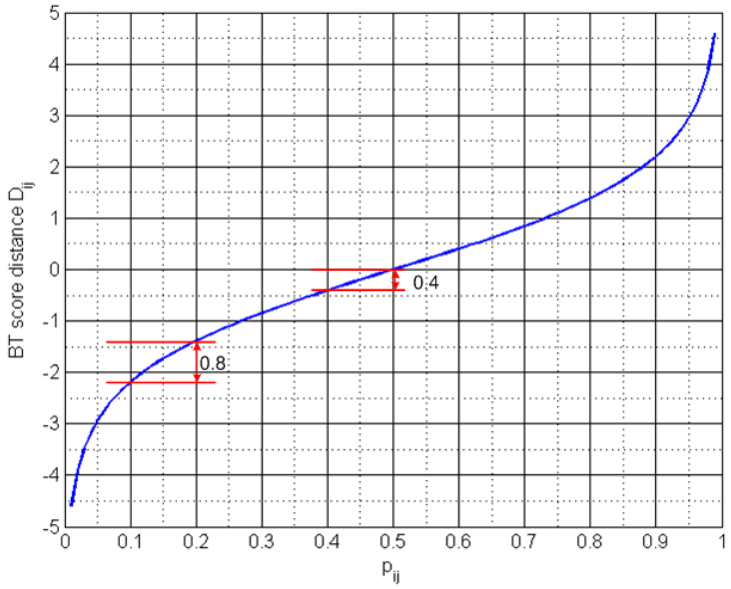


Figure 4.1: The relationship between the P_{ij} and the difference of BT scores.

on the original "true" distance of the test stimuli. Nearby pairs will be influenced less than the distant pairs. This conclusion is better visualized in Figure 4.1.

4.2.2 Sample size induced errors

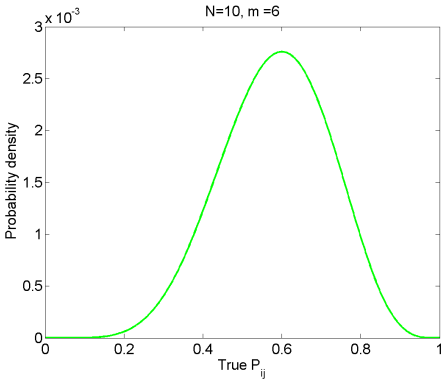
It is usually infeasible to ask a large number of observers participate in a subjective experiment, considering the expenses both on time and money. However, the limited number of observations would affect the accuracy of the estimates. For example, in a subjective test, if there are only N observations on the pair $\{S_i, S_j\}$, m observers chose S_i, thus, the observed $p_{ij} = m/N$. However, the true probability that S_i is preferred than S_j is P_{ij}, which can be approximated by infinite number of observations. A question is thus proposed: "With the observed p_{ij}, to what extent we can obtain the true P_{ij}, or what is the probability distribution of the real P_{ij}?". According to the probability distribution of the real P_{ij}, the probability distribution of the estimated distance D_{ij} can be obtained.

Based on the analysis above, we can calculate the posterior probability of P_{ij} by Bayes theory:

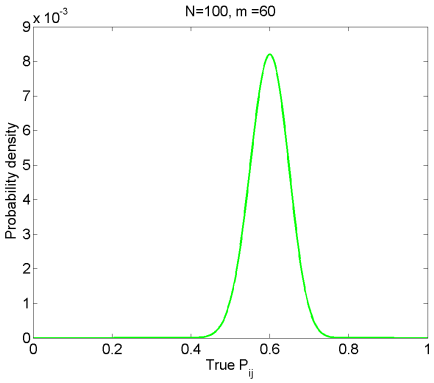
$$p(P_{ij} = P/p_{ij} = \frac{m}{N}) = \frac{p(p_{ij} = \frac{m}{N}/P_{ij} = P)p(P_{ij} = P)}{p(p_{ij} = \frac{m}{N})} \tag{4.2}$$

where $p(p_{ij} = \frac{m}{N})$ can be obtained by the Law of total probability:

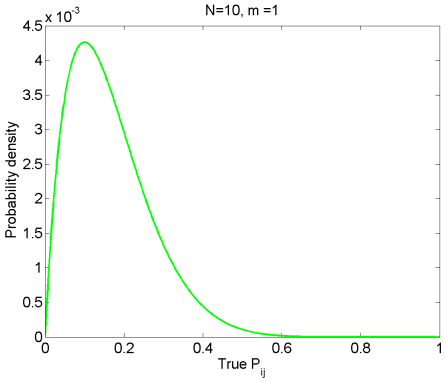
$$p(p_{ij} = \frac{m}{N}) = \sum_X p(p_{ij} = \frac{m}{N}/P_{ij} = X)p(P_{ij} = X) \tag{4.3}$$



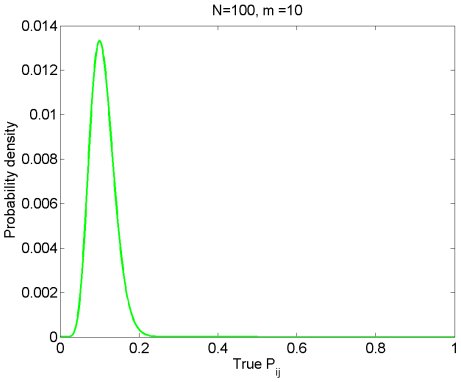
(a) closed pair condition, small number of observations



(b) closed pair condition, large number of observations



(c) distant pair condition, small number of observations



(d) distant pair condition, large number of observations

Figure 4.2: Distribution of the true P_{ij} value under different cases.

In natural world, the probability distribution of different events might be different. In this study, we chose the uniform distribution. Thus, $p(P_{ij} = X) = const = t$. So, we have:

$$p(P_{ij} = P/p_{ij} = \frac{m}{N}) = \frac{C_N^m P^m (1-P)^{N-m}}{\sum_X C_N^m X^m (1-X)^{N-m}} \tag{4.4}$$

To better visualize the distribution of P_{ij} under different number of observations, and how the number of observations affect the distribution of the true P_{ij}, some examples are shown in Figure 4.2. It is observed that the more the observations, the less standard deviations of the distribution on P_{ij}, which means the higher possibility that the observed value approximate the true value. For the condition of closer pairs, the standard deviation of P_{ij} is larger than the distant pairs. This explains the uncertainty of the viewers on voting of the closer pairs.

Considering the 95% upper and lower confidence intervals of the distribution of the P_{ij} value, the corresponding D_{ij} can be calculated. For better visualization, the confidence intervals of the P_{ij} value and the corresponding D_{ij} under different

conditions are shown in Figure 4.3.

According to Figure 4.3 it could be clearly found that:

1. The confidence intervals of P_{ij} decrease with the increasing number of observations.
2. The confidence intervals of P_{ij} for closer pairs are larger than distant pairs.
3. The confidence intervals of D_{ij} decrease with the increasing number of observations.
4. The confidence intervals of D_{ij} for closer pairs are smaller than distant pairs.

This provides the conclusions that though the uncertainty of the closer pairs would lead to a large range of possible estimations on the true P_{ij}, after the conversion from P_{ij} to D_{ij}, the uncertainty of the distance are concentrating on distant pairs rather than closer pairs. Thus, based on the objectives of the paired comparison test, i.e., assessing the values of the quality of the test stimuli, comparisons should be concentrated on closer pairs rather than distant pairs to generate accurate estimates on distances between stimulus pairs.

4.3 Optimization on the Balanced Sub-set Designs

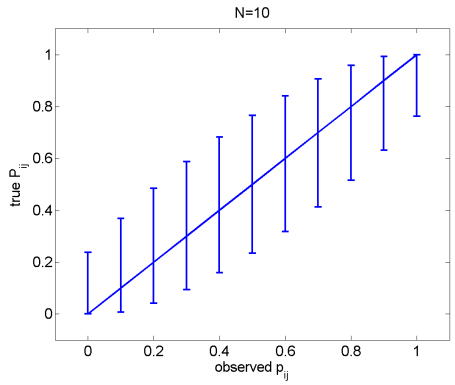
According to the analysis above, to conduct a paired comparison test more efficiently and robustly to various errors, the comparison should be concentrated on closer pairs. In this section, we propose a set of designs to optimize Dykstra's balanced sub-set designs [32]. These optimized designs are aiming at different test scenarios, i.e., the availability of the pre-tests or prior knowledge on the test stimuli.

4.3.1 Optimized Rectangular Design (ORD)

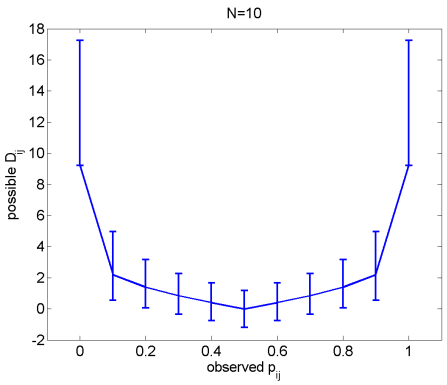
In our study, we extend the original "Square Design" to "Rectangular Design" which means the matrix **R** in SD method can be not only a square matrix but also any rectangular matrix. The rule of comparisons is kept the same, i.e., only the stimuli in the same column or row are compared.

The "**Optimized** rectangular design" is proposed for the conditions that the ranking of the stimuli in the test is known based on pre-test results or prior knowledge. The number of the stimuli m is a divisible number, i.e., $m = t_1 t_2$, where t_1 and t_2 are integers. $t_1 = t_2 = t$ is a special case for the Square Design with abbreviation **OSD** in this study.

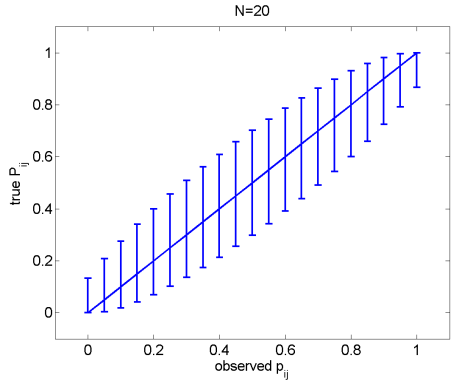
Supposing the ordering indices of the stimuli (descending or ascending) is $\boldsymbol{d} = (d_1, d_2, ..., d_m)$. The square matrix is arranged in such a way that the elements of the vector $\boldsymbol{d}$ are placed along a spiral as shown in Figure 4.4, which is defined as matrix $\mathbf{R}_{ORD}$.



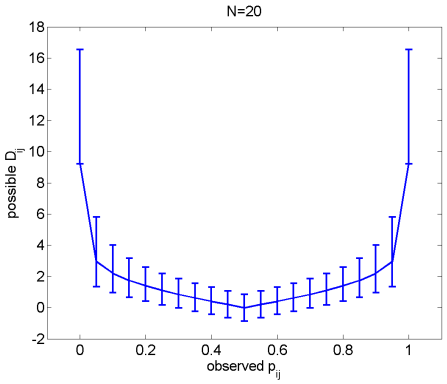
(a) $N = 10$, distribution of P_{ij}



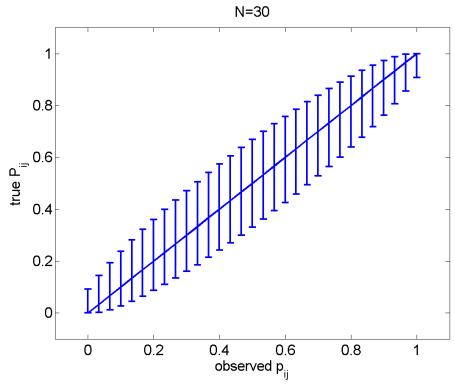
(b) $N = 10$, distribution of D_{ij}



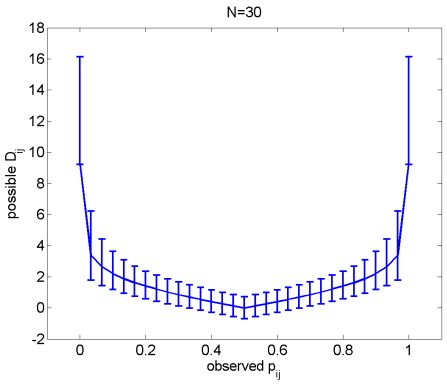
(c) $N = 20$, distribution of P_{ij}



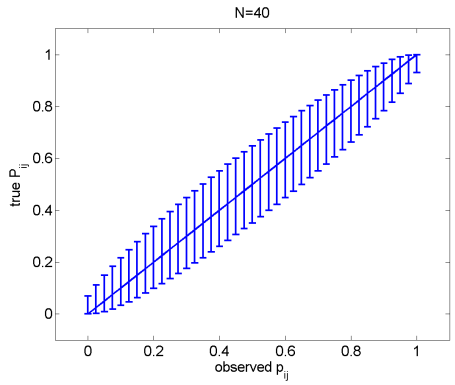
(d) $N = 20$, distribution of D_{ij}



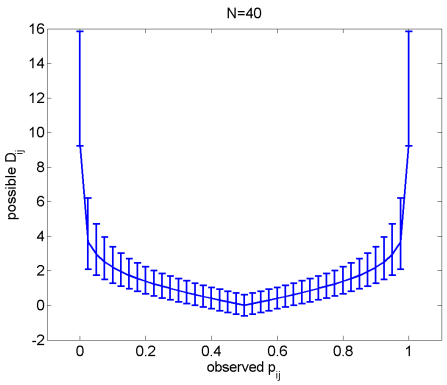
(e) $N = 30$, distribution of P_{ij}



(f) $N = 30$, distribution of D_{ij}



(g) $N = 40$, distribution of P_{ij}



(h) $N = 40$, distribution of D_{ij}

Figure 4.3: Confidence intervals of the true P_{ij} and D_{ij} value under different cases.

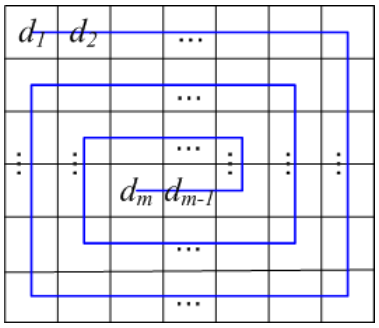


Figure 4.4: The design for rectangular matrix $\mathbf{R}_{ORD}$.

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:

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

$\mathbf{R}_{ORD} = (r_{pq})_{t_1 \times t_2}$, r_{pq} is the index of the stimulus in position (p, q). In this design, the matrix $\mathbf{R}_{ORD}$ doesn't change for all observers.

For better understanding, an example is given here. Supposing there are 12 test stimuli. In a pre-test, each pair was compared once. Thus, one whole observation was conducted (equals to the number of observations that one observer conducted a FPC test). According to this pre-test, the rank ordering of these stimuli is $\boldsymbol{d}$ = $(2, 5, 6, 1, 8, 9, 3, 10, 4, 11, 7, 12)$. The $\mathbf{R}_{ORD}$ is designed as follows:

$$\mathbf{R}_{ORD} = \begin{bmatrix} 2 & 5 & 6 & 1 \\ 11 & 7 & 12 & 8 \\ 4 & 10 & 3 & 9 \end{bmatrix}$$

In this way, the adjacent stimulus indices d_i and d_{i+1} are always arranged in the same column or row of the matrix $\mathbf{R}_{ORD}$ $(p = p' \vee q = q')$. In this example, $\boldsymbol{C}' = \{(2, 5), (2, 6), (2, 1), (5, 6), (5, 1), (6, 1), (11, 7), (11, 12), ...\}$. In the test, each participant compares the stimulus pairs whose indices belong to set $\boldsymbol{C}'$, i.e., stimuli $\{S_2S_5\}$, $\{S_2S_6\}$, etc. The number of appearance for each stimulus is five for each participant.

The accuracy of the ORD method is dependent on the accuracy of the pre-test, more analysis will be shown later in Section 4.4.

4.3.2 Adaptive Rectangular Design (ARD)

The "**Adaptive** Rectangular Design" is proposed in the way that the matrix $\mathbf{R}_{ORD}$ is updated for each observer. ASD (Adaptive Square Design) is a special case for ARD. This adaptive design is used for the conditions that previous estimates are not available. The detailed steps of this design are shown as follows:

1. For the $1st$ observer, the indices of the stimuli are randomly placed in $\mathbf{R}$. Run pair comparison experiment, only the pairs whose indices are in the same column or row of $\mathbf{R}$ are compared. This step is same as the original SD method.

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 the test). Based on the ordering vector $\boldsymbol{d}^{k-1}$, the matrix $\mathbf{R}_{ORD}^k$ and $\boldsymbol{C}'^k$ are constructed as shown in Figure 4.4, ($\mathbf{R}_{ORD}^k$ and $\boldsymbol{C}'^k$ represents $\mathbf{R}_{ORD}$ 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 the targeted accuracy on confidence intervals are obtained).

For better understanding, we still take the 12 stimuli as an example. As there is no pre-test for the test stimuli, for the first observer, the indices of the stimuli are randomly arranged in the matrix as follows:

$$\mathbf{R} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix}$$

Thus, for the first observer, there are in total 30 pairs to compare. After the first observer's test, the rank ordering of the quality of the stimuli is estimated as: $\boldsymbol{d}^1 = (3, 5, 1, 6, 9, 12, 2, 4, 8, 7, 10, 11)$. For the second observer, the matrix $\mathbf{R}_{ORD}$ is arranged according to this rank ordering, thus:

$$\mathbf{R}_{ORD}^2 = \begin{bmatrix} 3 & 5 & 1 & 6 \\ 7 & 10 & 11 & 9 \\ 8 & 4 & 2 & 12 \end{bmatrix}$$

Then, for the third observer, the matrix $\mathbf{R}_{ORD}^3$ is updated based on all previous 2 observers' pair comparison results. The same to the remaining observers until the test being finished.

Table 4.1: Design of the Monte Carlo simulation experiments for evaluation of the ARD methods

	Exp1		Exp2	
Exp. Name	Exp-1a	Exp-1b	Exp-2a	Exp-2b
Method	ASD, FPC, SD, SA		ASD, FPC, SD, SA	
Number of stimuli	25		36	
Number of obs.	10,20,30,40,50	-	10,20,30,40,50	-
Number of trials.	-	(3,6,9,12,15)$\times 10^3$	-	(6.3,12.6,18.9,25.2,31.5)$\times 10^3$
	Exp3		Exp4	
Exp. Name	Exp-3a	Exp-3b	Exp-4a	Exp-4b
Method	ARD, FPC, RD, SA		ARD, FPC, RD, SA	
Number of stimuli	20		30	
Number of obs.	10,20,30,40,50	-	10,20,30,40,50	-
Number of trials.	-	(1.9,3.8,5.7,7.6,9.5)$\times 10^3$	-	(4.35,8.7,13.05,17.4,21.75)$\times 10^3$

4.4 Monte Carlo simulation experiments

To evaluate the performance of the proposed designs, a group of Monte Carlo simulation experiments is designed and conducted. Firstly, the ARD method will be compared with the FPC, the original Rectangular and Square design, and the sorting algorithm based design (SA). Secondly, the ORD method will be compared with the FPC and its corresponding ARD method. In particular, the impact of the accuracy of the prior knowledge on the final estimation will be analyzed for the ORD method. Finally, all these designs are compared in a typical number of observers.

4.4.1 Evaluation of the ARD method

The design of the Monte Carlo experiments is shown in Table 4.1. As shown in this table, the factors considered in this study include the test methods (SD or RD), number of stimuli (25 and 36 for SD, 20 and 30 for RD), number of observers and number of trials. Please note that two kinds of comparison are conducted in terms of the number of observers and the number of trials. For example, in Experiment 1a, the performance of different designs are compared in the condition of with the same number of observers. In Exp-1b, the performances of these designs are compared in the condition of with the same number of trials. The number of trials selected in the study equals to the number of comparisons in a FPC test when the number of observers are 10, 20, 30, 40, 50. For example, in Exp-1b, the selected numbers of trials are $(3, 6, 9, 12, 15) \times 10^3$, which are corresponding to the number of comparisons for 10, 20, 30, 40, 50 observers by using FPC method.

The scores of all test stimuli were randomly selected from a uniform distribution on the interval of [1 5]. This design corresponds to the ACR-5-point MOS scale of the video quality assessment experiments. The simulation was conducted by the following assumptions:

1. Each stimulus has a single score;

2. In each observation, the observed value follows a gaussian distribution, the mean value is the stimulus score and the standard deviation is 0.7, which is obtained from the subjective scores of VQEG HDTV Final Report[137];
3. Each observer has a 5% probability to make a mistake on an observation, i.e., inverting the vote;
4. Each comparison is independent.

The Bradley-Terry model was used to convert the raw data to scale scores. The RMSE and ROCC between the estimated scores and the designed scores were calculated. The simulation was run 100 times for each case, thus, the mean and confidence intervals of the RMSE and ROCC can be obtained.

The performances of ASD method are shown in Figure 4.5 and Figure 4.6. The results indicate that to achieve the same accuracy of the estimates as the FPC methods with 20 observers (a typical number of observers for paired comparison test), the requited number of observers for ASD method is the least comparing with other methods, which is about 40-50. If comparing the performances of different designs with the same number of trials, for example in Figure 4.5(d), for the ASD method, the RMSE between the estimates and the ground truth value is decreased approximately 10% compared with the FPC method. Contrarily, for the SD and the SA design, the RMSE is increased approximately 10% and 39% respectively. Thus, the results demonstrated that though the SD and the SA design performed reliably under perfect assumption conditions in [113][32], in real subjective experiments, due to the influence of observation errors, these designs might not be as reliable as the FPC method. The proposed ASD method is proved not only efficient and robust to observation errors, but also more accurate in generating estimates than the FPC method.

The performances of ARD method with 20 and 30 stimuli are shown in Figure 4.7 and Figure 4.8. Similar conclusions are drawn. The required number of observers for ARD method to achieve the accuracy produced by 20 observers using FPC method is approximately 40 to 50. With the same number of trials, the ARD method performs the best.

4.4.2 Evaluation of the ORD method

The performance of the ORD method is dependent on the accuracy of the estimated ranking order or prior knowledge of the test stimuli. To evaluate its performance, three levels of prior knowledge on the estimated ranking order are considered. Assuming that in the pre-test, 1, 3, and 6 observers participate in the test by using the FPC method, and then the estimated ranking order is obtained for the arrangement of the matrix **R**. The ORD methods based on 1, 3, 6 observers' whole

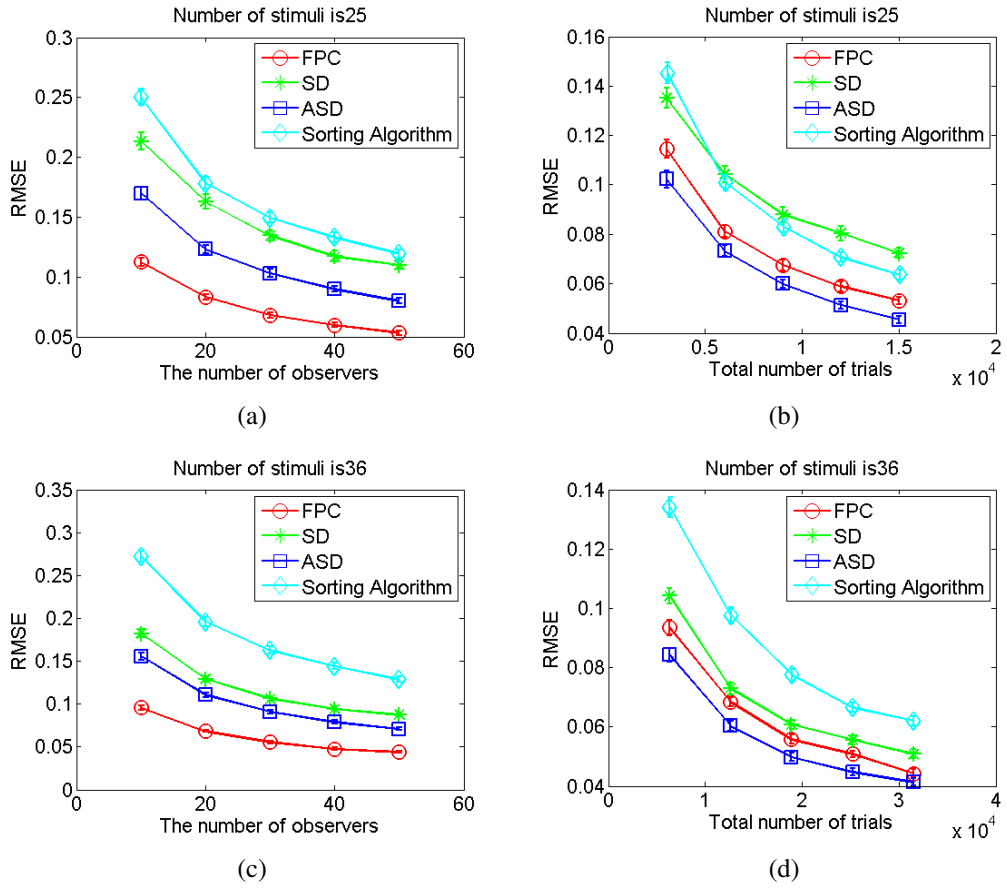


Figure 4.5: Comparison of RMSE between ASD and other designs at various test scenarios. The number of stimuli in (a) and (b) is 25, while in (c) and (d) is 36. In (a) and (c), the X-axis represents the number of observers. In (b) and (d), the X-axis represents the total number of comparisons. The Y-axis is the RMSE. The error bars are the confidence intervals of the estimated scores.

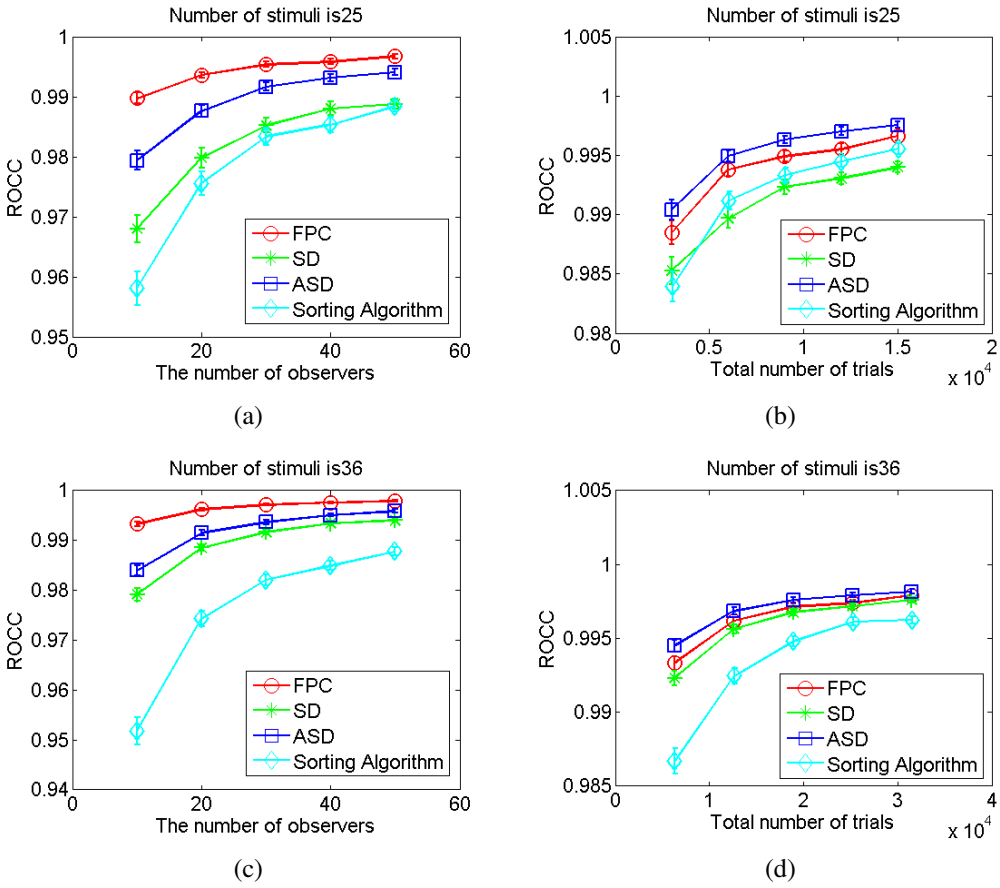


Figure 4.6: Comparison of ROCC between ASD and other designs at various test scenarios. The number of stimuli in (a) and (b) is 25, while in (c) and (d) is 36. In (a) and (c), the X-axis represents the number of observers. In (b) and (d), the X-axis represents the total number of comparisons. The y-axis is the ROCC. The error bars are the confidence intervals of the estimated scores.

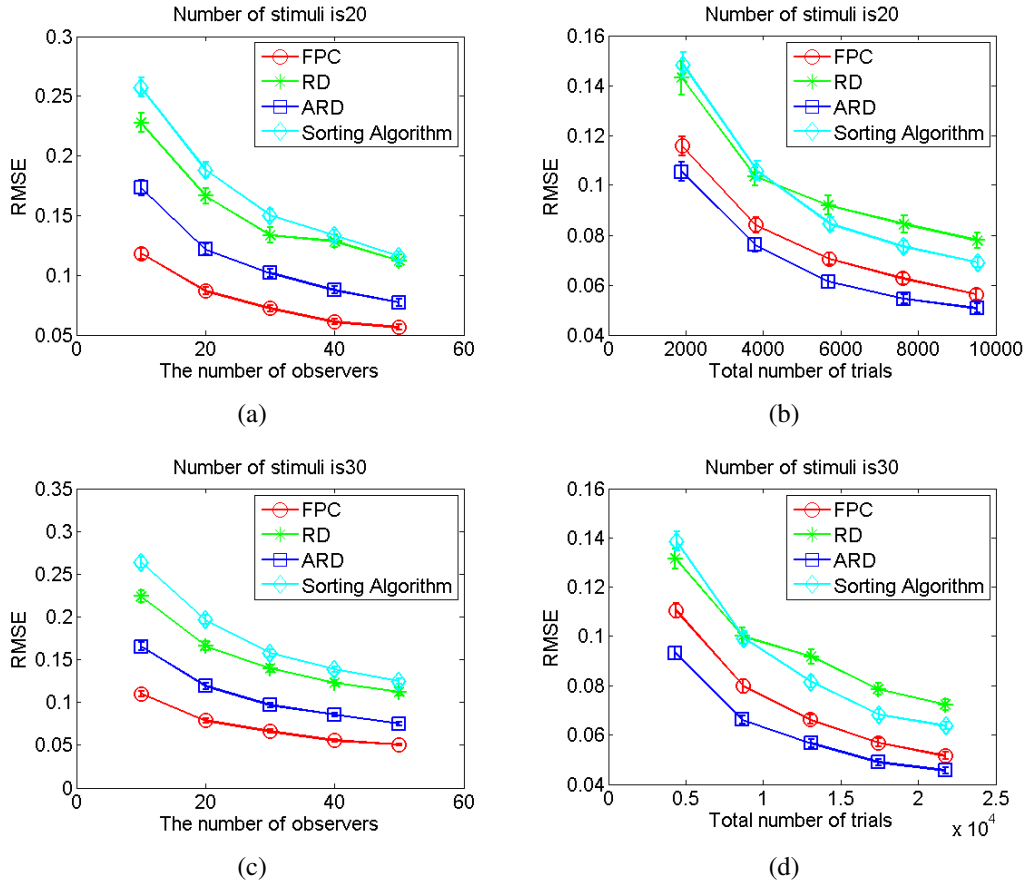


Figure 4.7: Comparison of RMSE between ARD and other designs at various test scenarios. The number of stimuli in (a) and (b) is 20, while in (c) and (d) is 30. In (a) and (c), the x-axis represents the number of observers. In (b) and (d), the x-axis represents the total number of comparisons. The y-axis is the RMSE. The error bars are the confidence intervals of the estimated scores.

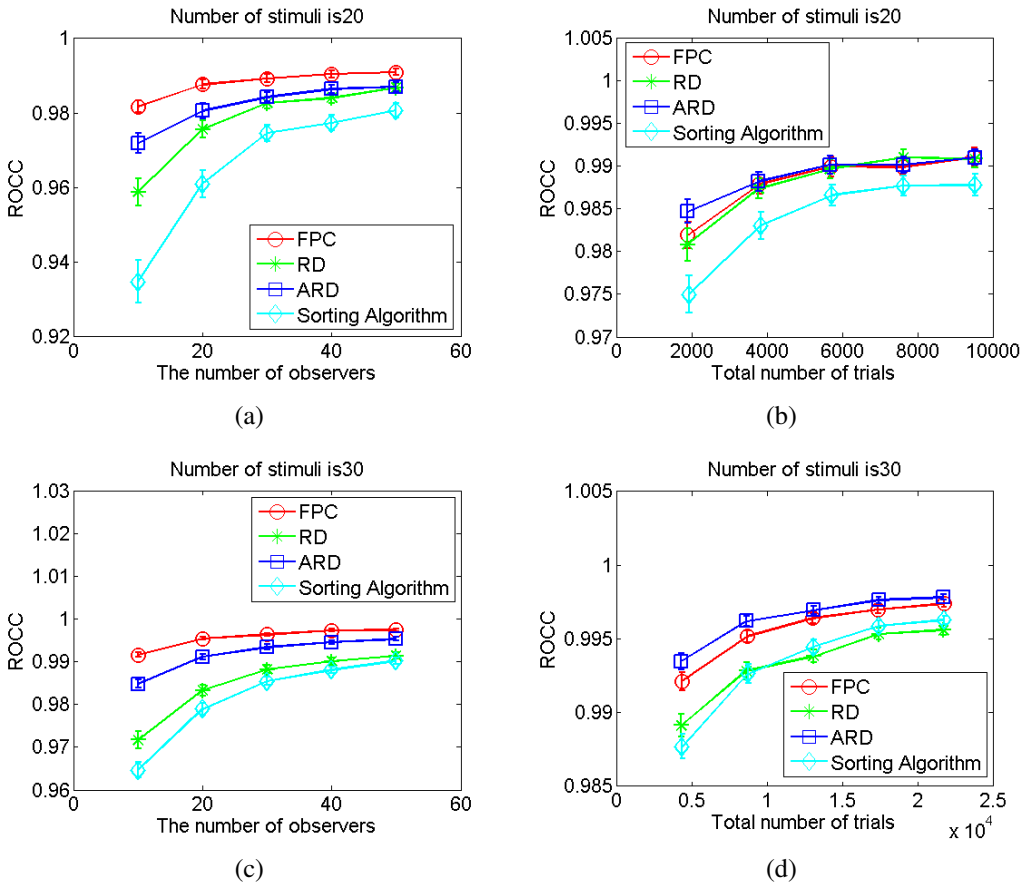


Figure 4.8: Comparison of ROCC between ARD and other designs at various test scenarios. The number of stimuli in (a) and (b) is 20, while in (c) and (d) is 30. In (a) and (c), the x-axis represents the number of observers. In (b) and (d), the x-axis represents the total number of comparisons. The y-axis is the ROCC. The error bars are the confidence intervals of the estimated scores.

Table 4.2: Design of the Monte Carlo experiments for evaluation of the OSD methods

	Exp5		Exp6	
Exp. Name	Exp-5a	Exp-5b	Exp-6a	Exp-6b
Method	ASD, FPC, OSD-1, OSD-3, OSD-6		ASD, FPC, OSD-1, OSD-3, OSD-6	
Number of stimuli	25		36	
Number of obs.	10,20,30,40,50	-	10,20,30,40,50	-
Number of trials.	-	$(3,6,9,12,15)\times10^3$	-	$(6.3,12.6,18.9,25.2,31.5)\times10^3$

Table 4.3: The accuracy of the pre-test estimation on the scores of the stimuli

Evaluation Method	Number of stimuli	Mean			Std.		
		OSD-1	OSD-3	OSD-6	OSD-1	OSD-3	OSD-6
RMSE	25	0.28	0.17	0.12	0.0042	0.003	0.0015
	36	0.24	0.15	0.11	0.0038	0.0009	0.0014
ROCC	25	0.95	0.98	0.99	0.0029	0.0013	0.0003
	36	0.96	0.98	0.99	0.0013	0.0005	0.0004

observations using the FPC method are referred as ORD-1, ORD-3 and ORD-6, respectively.

In this study, we take the OSD as an example to evaluate the ORD as the performance of ORD should be similar as OSD, as shown in the previous section (Section 4.4.1). The original SD and the SA (Sorting algorithm based design) methods are not taken into consideration in this study. As shown in the previous session, the performances of these two methods are not as reliable as the FPC and ASD method. Thus, in this study, we only compare the performance between FPC, ASD and OSD at different "prior-knowledge" levels, e.g., OSD-1, OSD-3, OSD-6. Details can be found in Table 4.2. Two kinds of comparison are conducted in terms of the number of observers and the number of trials, which is the same as the Exp1 and Exp2. In Exp-5a and Exp-6a, for example, in the condition of 25 stimuli, each observer needs to compare 300 pairs using the FPC method but 100 pairs for the ASD and OSD methods. In the condition of 36 stimuli, each observer needs to compare 630 pairs using the FPC method but 180 pairs for the ASD and OSD methods.

The design of the Monte Carlo simulation experiment were similar as the evaluation on ARD design. The scores of all test stimuli were randomly selected from a uniform distribution on the interval of [1 5] with standard deviation of 0.7. The simulation was run 100 times for each case.

As the performance of the ORD method is dependent on the accuracy of the pre-test. Thus, firstly, the accuracy of the estimation on the quality value of the stimuli in pre-test is shown in Table 4.3. Two evaluation methods are used, RMSE is used to evaluate the accuracy, ROCC is for the evaluation of the consistency.

The comparison on the Monte-Carlo test results of the FPC, ASD and OSD methods are shown in Figure 4.9 and Figure 4.10.

It is assumed that the performance of the OSD method is dependent with the accuracy of the pre-test estimation. This hypothesis is verified by the results shown

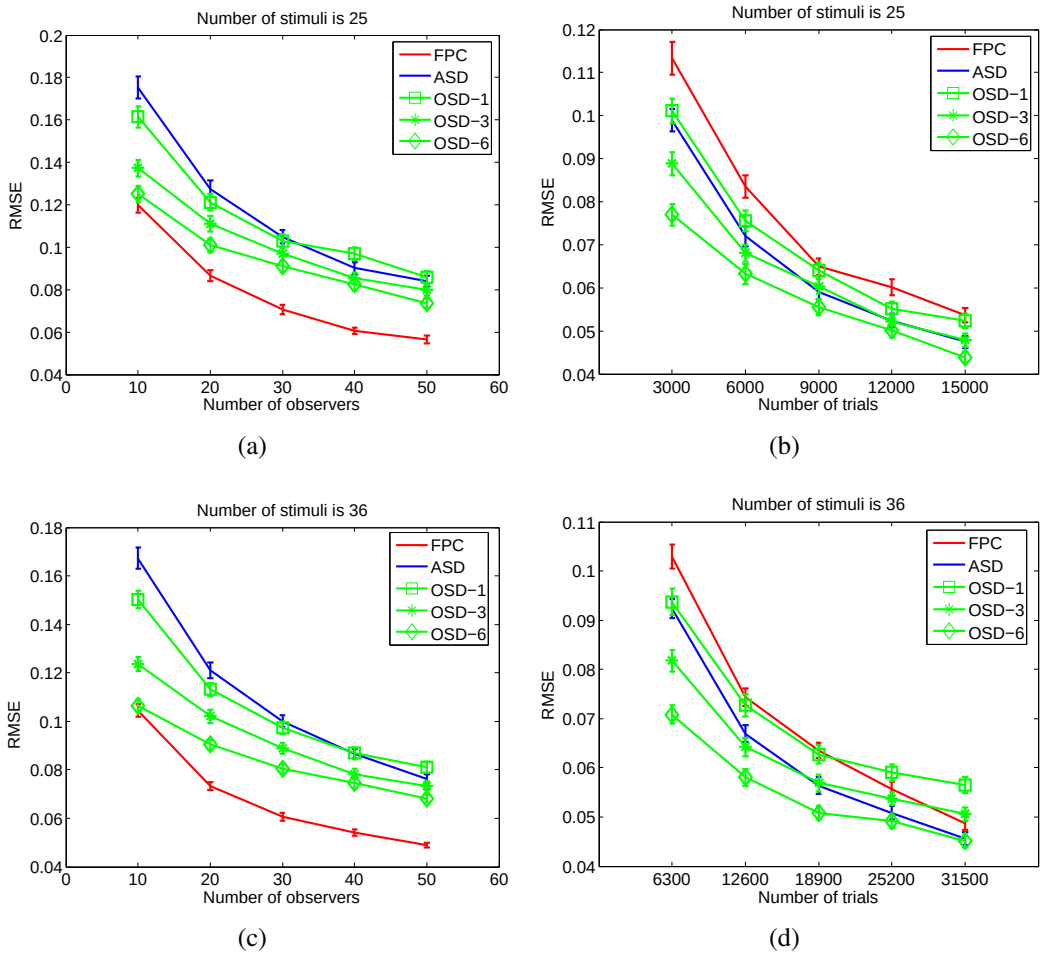


Figure 4.9: Comparison of RMSE between OSD and other designs at various test scenarios. The number of stimuli in (a)(b) are 25, and in (c)(d) are 36. In (a)(c), the x-axis represents the number of observers in the test. In (b)(d), the x-axis represents the number of comparisons which are determined by the total number of trials by 10, 20, 30, 40, 50 observers using the FPC method. The y-axis is the RMSE. The error bars are the confidence intervals of the estimated scores.

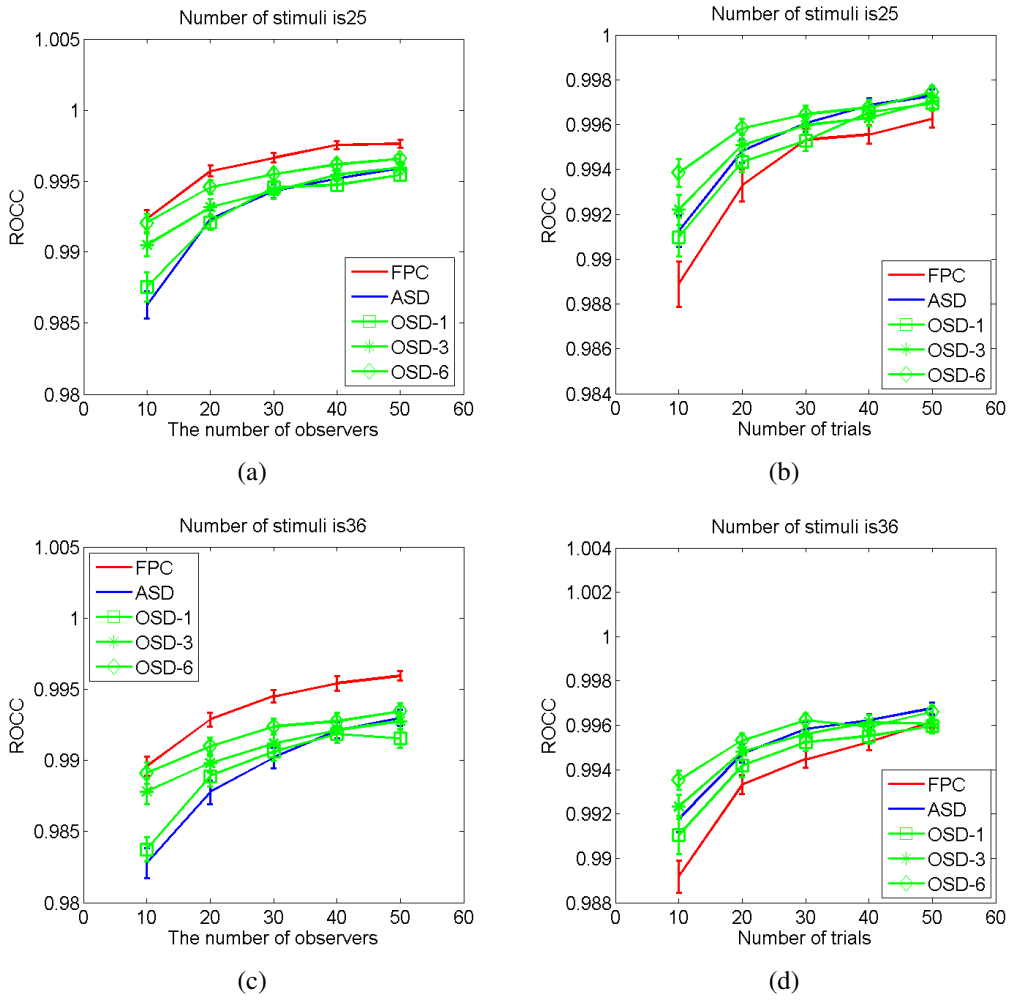


Figure 4.10: Comparison of ROCC between OSD and other designs at various test scenarios. The number of stimuli in (a)(b) are 25, and in (c)(d) are 36. In (a)(c), the x-axis represents the number of observers in the test. In (b)(d), the x-axis represents the number of comparisons which are determined by the total number of trials by 10, 20, 30, 40, 50 observers using FPC method. The y-axis is the ROCC. The error bars are the confidence intervals of the estimated scores.

in Figure 4.9, the performance of the OSD method is increasing with the accuracy of the pre-test estimation.

From Figure 4.9(a) and Figure 4.9(c) it could be found that with the same number of observers, the OSD method is more accurate in estimates than the ASD method in the condition that the number of observers is less than 30. With the increase of the number of observers, the performance of the OSD is getting closer to the ASD method. Furthermore, in the condition of small number of observers, e.g., 10, the performance of the OSD-6 is comparable with FPC. By comparing the influence of the number of observers, it is indicated that the OSD-6 method with 50 observers can generate a comparable result with the FPC method of 30 observers.

If take a look at the ROCC of these designs shown in Figure 4.10(a) and Figure 4.10(c), it could be found that the OSD method is generally more consistent with the ground truth in rank ordering.

In the condition of with the same number of trials, as shown in Figure 4.9(b) and Figure 4.9(d), the OSD-6 performs the best, then follows the OSD-3, the ASD method is comparable with the OSD-1 method when the number of trials is not quite large. With the increase of the number of trials, the performance of the ASD method is getting better and better and finally comparable with the OSD-6 method. It should be noted that in Figure 4.9(d), with the same number of trials, the estimation accuracy of the OSD-1 method does not converge with other designs. This might be due to the singularity of the paired comparison matrix, which will affect the estimation accuracy of the paired comparison model when the number of comparisons is large.

4.4.3 Performance analysis under different numbers of test stimuli

In the previous sections, the proposed designs are evaluated by comparing them with existing designs under different test scenarios, i.e., different number of observers, and different number of trials. In this section, the number of observers is fixed, the performances of different designs under different number of stimuli are evaluated.

In this study, the number of observers is fixed to 40. The number of stimuli is ranged from 9 to 36. Similar Monte-Carlo simulation tests were conducted for all test scenarios. RMSE is used to evaluate the performance. The results are shown in Figure 4.11. The results indicate that, generally, the original RD method performs the worst comparing with our proposed designs. With the increase of the accuracy of the pre-test results, the performance of the ORD is increasing as well. The ARD method is comparable with the ORD-3 method.

The performances of the proposed designs are dependent on the total number of comparisons. More comparisons will generate more accurate results. The rela-

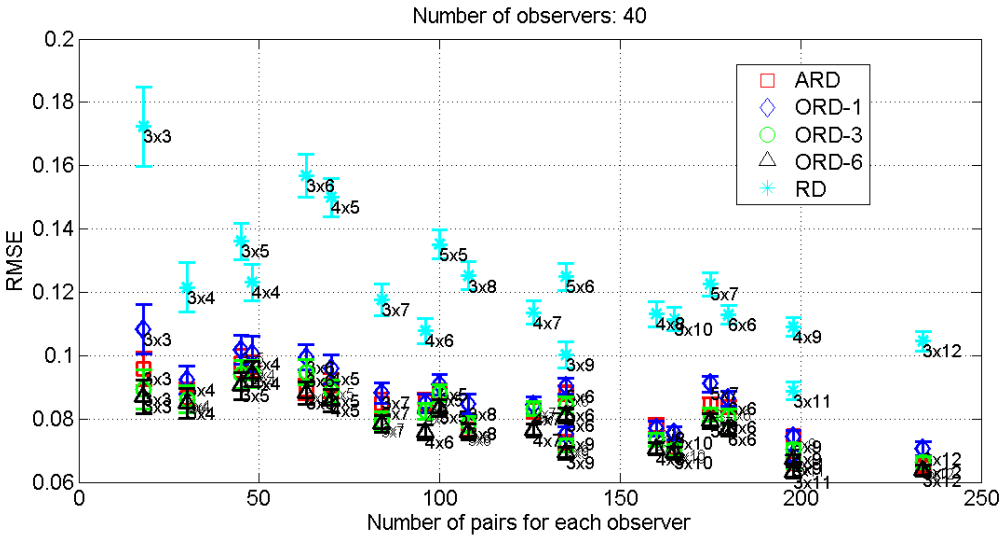


Figure 4.11: Comparison of different designs under different numbers of stimuli.

tionship between the number of comparisons and RMSE is not monotonic in Figure 4.11 due to the influence of the designed scores of the stimuli. In our simulation experiment, the scores of the stimuli are different for different test scenarios, which are randomly generated. According to the analysis in this chapter, the accuracy of the estimates is dependent on the original distances between the test stimuli. Thus, the randomly generated scale values would affect the performances of the test designs.

4.5 Constraints on Pair Comparison test

Two proposed optimization implementations for the balanced sub-set paired comparison designs are evaluated in previous sections, which showed higher correlation with the ground truth than the original designs and other efficient designs. As only parts of the whole pairs are compared in the test, there should be some constraints on the test procedure to avoid any bias from the video contents or presentation order of the stimuli.

4.5.1 Number of observers

Generally, in a paired comparison test using FPC method, 10 is the minimum required number of observers [13], generally, 20 observers are required. To achieve the same level of accuracy, it's necessary to evaluate the required number of observers by using an efficient pair comparison design. According to the Monte-Carlo simulation results, 40 observers are necessary for the ARD or ORD methods.

4.5.2 Number of stimuli

Considering the most disadvantages of the paired comparison, i.e., time consuming, the number of test stimuli should be within a certain range. Furthermore, the number of video contents should also be taken into account to avoid any visual fatigue induced by watching same video content. Thus, the guideline for the selection of the number of stimuli is proposed as below:

1. For each observer, the duration of the whole test should be within 30 to 60 minutes [56]. The number of stimuli can be estimated based on the test method and the duration of the stimuli. For example, if the duration of the test stimulus is 10s, and between each sequence there will be a gray image which lasts 3s, adding 5 seconds for voting, one pair will cost $10 + 3 + 10 + 5 = 28$ seconds using time sequential paired comparison, and $10 + 5 + 3 = 18$ seconds using time parallel paired comparison (the test stimulus pairs are shown on two displays simultaneously). The targeted number of pairs in the test is 64 to 128 pairs for time sequential method, and 100 to 200 pairs for side by side method.
2. Generally, in a image/video quality assessment experiment, there will be several video contents (SRC) with different types of distortions, or different levels of degradations (HRC). In this case, only the video sequences with same content are compared in the test. The number of video contents and the number of degradation types should be estimated to obey the rule 1. For example, if 8 video contents are selected, and the planned test duration is approximately 60 mins, the side by side pair comparison is selected, then, for each viewer, the maximum number of pairs is 200 which leads to $200/8 = 25$ pairs/SRC. According to Figure 4.12, the number of HRCs is approximately 10, e.g., the matrix $\mathbf{R}$ can be 2×5 or 2×4 or 3×3.
3. The selection of the number of SRC should obey the rule that observers would not feel fatigue, impatient or annoyed with the repeated contents, i.e., the number of contents should not be too limited.

4.5.3 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 as follows in this study:

1. The presentation of the sequence content should be as random as possible, no observer watches the same content in two consecutive presentations.
2. 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

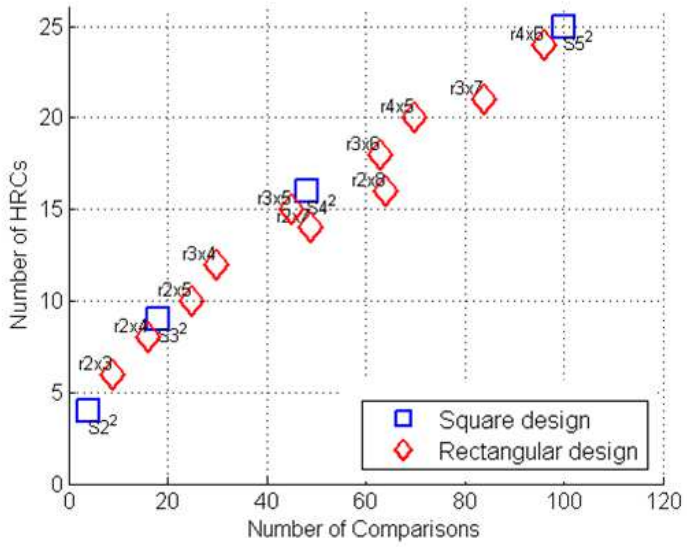


Figure 4.12: Relationship between the number of HRCs and the number of comparisons in Rectangular design.

stimulus S_A, half of the pairs should show S_A firstly, the rest should show S_A secondly.

3. For all observers, all the pairs of stimuli must be displayed in both orders. For example, if one observer watches $\{S_A S_B\}$, there must be another observer who watches $\{S_B S_A\}$.

4.6 Statistical test on preference

The traditional method to analyze the paired comparison results are using the paired comparison models, i.e., Bradley-Terry model, Thurstone-Mosteller model. However, when the paired comparison model fails in model fit, the converted scale values cannot used to explain the raw data. In this case, a statistical tool to analyze the raw data is necessary. In this section, some statistical tools are introduced. Furthermore, a Monte-carlo analysis method is proposed for the detection of significant influence factors.

4.6.1 Conditional and unconditional tests for 2 $\times$ 2 comparative trials

It is important to distinguish two proportional values statistically. The conditional and unconditional tests are frequently used methods in this scenario and they are usually applied on the food taste related area. A contingency table as shown in Table 4.4 is used here to help illustrate the objectives of this section. Supposing in a paired comparison test for the pair $\{S_1, S_2\}$, in observer Group 1, m_1 out of N_1

Table 4.4: An example of 2 × 2 contingency table

	Group 1	Group 2	Total
Choose S_1	m_1	m_2	$m = m_1 + m_2$
Choose S_2	$N_1 - m_1$	$N_2 - m_2$	$N - m$
Total number	N_1	N_2	N

participants prefer S_1 over S_2 while in Group 2 this ratio is m_2/N_2. m_1 and m_2 are two independent binomial variables, $m_i \sim B(N_i, \theta_i)$, i = 1, 2. θ_i denotes the proportion of observers choosing S_1 in Group i, i.e., $\theta_i = m_i/N_i$. The null hypothesis H_0 and alternative hypothesis H_a are :

$$H_0 : \theta_1 = \theta_2$$
$$H_a : \theta_1 \neq \theta_2$$

Basically, there are two fundamentally different exact tests for the null hypothesis, namely conditional and unconditional exact tests. Fisher's exact test [37] and Barnard's exact test [6] are two typical conditional and unconditional exact tests, respectively. In the condition of small sample size (e.g., $N_i < 50$), no matter whether the sample sizes are balanced or not (e.g., $N_1 \approx N_2$ or $N_1 = 4N_2$), the Barnard's exact test is more powerful than Fisher's exact test [6]. However, with the increase of the sample size, the Fisher's exact test becomes more powerful. For the explanation of "powerful" and more details about the comparison on these two tests, the reader are referred to [94][93].

For the condition of with large sample size, e.g., $N_i > 200$, the Barnard's exact test cannot be applied. The alternative is to use asymptotic tests, e.g., χ^2 type, arc sine or Fisher's mid-p-value test [90]. The optimal choice of the asymptotic tests is dependent on the real p-value, the unbalance of samples, etc. Generally, the Fisher's mid-p-value based methods are more reliable than others [90].

In conclusion, in paired comparison data analysis, these methods 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. The output of the Barnard's test is a p-value. On a 95% confidence level, 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. Otherwise, there is no significant difference.

4.6.2 Monte Carlo significant test

After applying the conditional or unconditional test on all pairs of the two conditions, the number that in total a out of N pairs are significantly different can

be obtained. To check if the test conditions have an influence on the results (if $r_t = a/N$ is statistically large), a Monte-Carlo simulation experiment can be conducted by randomly permuting the observers in two test scenarios. The analysis of the scenario as an influencing factor can be achieved by comparing the ratio of significantly different pairs in Monte-Carlo simulation test with the ratio of the test observer groups (r_t). The details are illustrated in Algorithm 1.

Algorithm 1: Monte-Carlo simulation experiment

Require: $Loop\_num$ The number of trials;
N The number of stimulus pairs;
$\mathbf{A}_i$ The pair comparison matrix of observer i;
k = 0
while $k < Loop\_num$ **do**
$Group_1, Group_2 \leftarrow$
Divide observers indices into two groups randomly
$(\mathbf{A}_{group1})_{m\times m} \leftarrow \sum_{i\in Group_1} \mathbf{A}_i$
$(\mathbf{A}_{group2})_{m\times m} \leftarrow \sum_{j\in Group_2} \mathbf{A}_j$
$e(k) \leftarrow$ Number of sig. different pairs on $\mathbf{A}_{group1}, \mathbf{A}_{group2}$
$r(k) \leftarrow e(k)/N$
$k \leftarrow k + 1$
end while
$H(r) \leftarrow$ Probability distribution of $\boldsymbol{r}$
$\mu \leftarrow$ Mean of $\boldsymbol{r}$
$\boldsymbol{\sigma} \leftarrow H(\sigma) < 0.05$
return $\mu, \boldsymbol{\sigma}$

If $Loop\_num$ is sufficiently large (which also depends on the number of observers), e.g., 1000, the probability distribution of $\boldsymbol{r}$ can be estimated by the histogram. The mean value μ and the set of small probability events $\boldsymbol{\sigma}$ with probability of less than 5% can be calculated. If $r_t \notin$ set $\boldsymbol{\sigma}$, the test conditions may not have a significant influence on the results.

4.7 Conclusions

Paired comparison methodology might be a reliable subjective method for the 3DTV related psychophysical study. However, due to the drawback of the FPC method, it is often not feasible and applicable in a subjective experiment. Thus, some efficient pair comparison designs have been proposed. Most of the existing efficient designs are based on the assumption that there is no observation error, system error or other errors during the test while in fact they inevitably occur in real subjective tests.

Thus, in this chapter, based on the possible system errors, observation error and

the sample size induced errors, we analyzed the possibility of optimizing the selection of test pairs which are robust to all these errors in real subjective experiments. Two optimization implementations on the balanced sub-set pair comparison designs are proposed. One is the ORD method, the other is the ARD method.

ORD is used for the condition that the pre-test or prior knowledge on the test stimuli is available, in particular, the ranking order of the test stimuli. ARD is used for the conditions where the prior information is not available. We used the Monte Carlo simulation experiments to evaluate the performances of the FPC, Sorting Algorithm based design, the original balanced sub-set design and the proposed designs under the condition that the observer has 5% probability to make an error on voting, i.e., inverting the selection. The simulation results showed that the proposed designs are more robust than other designs with the same number of trials. To achieve the accuracy level of 20 - 30 observers using FPC methods, approximately 40 - 50 observers are needed for the proposed designs. Using a typical number of observers, the proposed designs are compared under different number of test stimuli. The results showed that the shape of the matrix **R** is not a significant factor for the performance of the design. For example, the matrix **R** with size of 4×9 and 3×12 will not generate significant different results. In fact, the total number of comparisons in the test is a key factor for the accuracy of the results.

The paired comparison models, which are used to convert the paired comparison data to scale values for the stimuli, may not work in some cases, i.e., the model fit fails to explain the raw data. In this case, some statistic tools to analyze the raw paired comparison data is necessary. In this chapter, some novel practices for statistical analysis of the paired comparison results are introduced. We refereed it as "novel" because they are usually used in other community but not in image/video quality assessment domain. We analyzed these methods and adapted them to our domain for different test scenarios. Furthermore, a Monte-carlo statistic analysis method is proposed to evaluate the significant factors of the data.

This chapter mainly focused on the mathematical analysis of the proposed designs. Some real subjective experiments should be conducted to verify their performances, which will be presented in the next chapter.

5

Evaluation of the Adaptive Square Design in subjective experiments of 3DTV

In the previous chapter (Chapter 4), we proposed a set of optimized balanced sub-set pair comparison designs which can reduce the number of trials and be more robust to the possible errors in real subjective experiments. These designs were evaluated by Monte Carlo simulation experiments. In this chapter, we select one of these optimized designs, i.e., the ASD method, to apply to the subjective visual discomfort experiments in 3DTV to verify its performance.

5.1 Introduction

Visual discomfort induced by watching 3D images or videos is getting more and more attention recently [150][149][118][84] [88][87][77] as it decreases the viewing experience of the viewers severely. As we already explained in Chapter 2, Section 2.4. For most viewers, they are not used to the 3D display technology, thus, the subjective assessment on the degree of visual discomfort is a challenging work due to the observer context dependency and the attribute selection issues.

Pair comparison is proposed as a solution to these issues as we already discussed in Chapter 2. In the previous chapter, we proposed a set of optimized balanced sub-set designs which were designed to be robust to possible errors in subjective experiments. Their performances were validated by the Monte-Carlo simulation

experiments and the results showed that they are more robust than the original designs and the existing efficient method (i.e., Sorting Algorithm based method).

In this chapter, the performances of the three pair comparison design methods, i.e., FPC (*Full Paired Comparison*), SD (*original Square Design*) and ASD (*Adaptive Square Design*) are compared by subjective visual discomfort experiments in S-3DTV. Due to the fact that the viewer's vote on paired comparison may be influenced by observation errors or the interaction of the votes on stimuli, five subjective experiments were thus 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.

5.2 Experiment

5.2.1 Experimental setup

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, which is approximately 3 times of the screen height. All environmental conditions were in line with ITU-R BT.500 [58].

There are in total 36 video sequences in the test. The stereoscopic sequences consist of a left-view and a right-view image which were generated by the MATLAB psychtoolbox [105]. Each sequence contain a fixed background and a moving black Maltese Cross. The background was generated by adding salt and pepper noise to a black image of Full HD resolution, and then filtered by a circular averaging filter. All the background and foreground have no quality degradation. Thus, these stimuli only induce visual discomfort in this study. No visual quality degradation was perceived.

These sequences contain different features to generate different degree of visual discomfort, e.g., disparity, moving velocity. More details can be found in Chapter 8. Here we only use their index, i.e., stimulus 1, 2, 3, ..., 36 to represent these sequences.

Table 5.1: 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

5.2.2 Experimental design

Five experiments were conducted. The summary of the experiments is shown in Table 5.1. The details are illustrated in the following sections.

Experiment 1: FPC method for establishing ground truth

Fifteen stimuli (Stimulus 1 to 15) were tested in the Experiment 1. By using the FPC method, there were in total $15 \times 14/2 = 105$ pairs presented in each individual subjective experiment (for each observer). The results were used as the ground truth for the visual discomfort of the 15 stimuli.

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 (Stimulus 1 to 16) were tested in both experiments. The reason why one extra stimulus was added in this test is that sixteen is the minimum required number to arrange all previous 15 stimuli of Experiment 1 in a square matrix. It is assumed that this extra stimulus would not generate significant influence on the final results. Thus, the main difference between this experiment and Experiment 1 is the observation errors in the two experiments 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 for SD method and the first observer of ASD method, as shown in the upper left 4×4 matrix in Figure 5.1. 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. Thus, the pairs for each observer may be different.

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 5.1: The layout of the stimulus indices 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.

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 (stimulus 1 to 36) were tested in both experiments.

For Experiment 4, the upper left 4×4 matrix stays the same as in Experiment 2. All the other positions were randomly placed by the remaining 20 stimuli as shown in Figure 5.1. 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 stimuli (Stimulus 1 to 15) can be analyzed by comparing the results of Experiment 2, Experiment 4 and the ground truth (Experiment 1).

For Experiment 5, the initial positions of all stimuli were the same with the Experiment 4. As the ASD method was used, after each observer's test, the positions of the stimuli were updated according to the rule of ASD.

According to the SD and ASD method, there are 180 pairs to be compared for each observer in both experiments.

5.2.3 Observers

The number of observers for each experiment is shown in Table 5.1. The observers are all non-experts in psychophysical studies on 3D, image processing or 3D related fields. 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.

It should be noted that the observers in Experiment 2 also participated Experiment 3. Half of the observers conducted Experiment 2 first and the remaining half conducted Experiment 3 first.

5.2.4 Procedures

The subjective experiment contained a training session and a test session. The task for the observers is that in each pair, they should select the one which they feel more uncomfortable, concerning e.g., difficulty to fuse, eye strain, headache.

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 for each observer. To avoid visual fatigue caused by long time watching to affect the experimental results, 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.

In all experiments, the presentation order for voting the whole set of pairs was randomly permuted for each viewer. The temporal presentation order of each pair of stimuli was balanced for all viewers, e.g., for stimulus pair $\{S_A, S_B\}$, half of the viewers watched $\{S_A\}$ first, half of the viewers watched $\{S_B\}$ first.

5.3 Experimental 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 [143].

The input of the Bradley-Terry model is the paired comparison matrix **A** with size of the number of stimuli, i.e., 15 $\times$ 15 for Experiment 1, 16 $\times$ 16 for Experiment 2 and 3, and 36 $\times$ 36 for Experiment 4 and 5. We didn't take only the 15 $\times$ 15 paired comparison matrix as the input of the Bradley-Terry model for all experiments is because in Experiment 5 using ASD method, the 15 $\times$ 15 matrix is too sparse, it is not reasonable to use only the sub-matrix of the whole matrix to estimate the final scale values for the 15 stimuli.

The output of the Bradley-Terry model is the scale values for all stimuli. Based on the objectives of this study, only the Bradley-Terry scores of Stimuli 1 to 15 are analyzed. The goodness of model fit p-values for the five experiments indicate that the scale values are able to explain the raw data. In this study, the Bradley-Terry score represents the degree of the experienced visual discomfort. The higher the Bradley-Terry score, the more visual discomfort was perceived.

5.3.1 Comparative analysis

In this section, the general performances of the SD method and the ASD method are compared through the correlations between the experimental results and the ground truth. The scatter plot of the test results and the ground truth are shown in Figure 5.2.

The line in Figure 5.2 provides a reference of gradient equaling to 1. Generally,

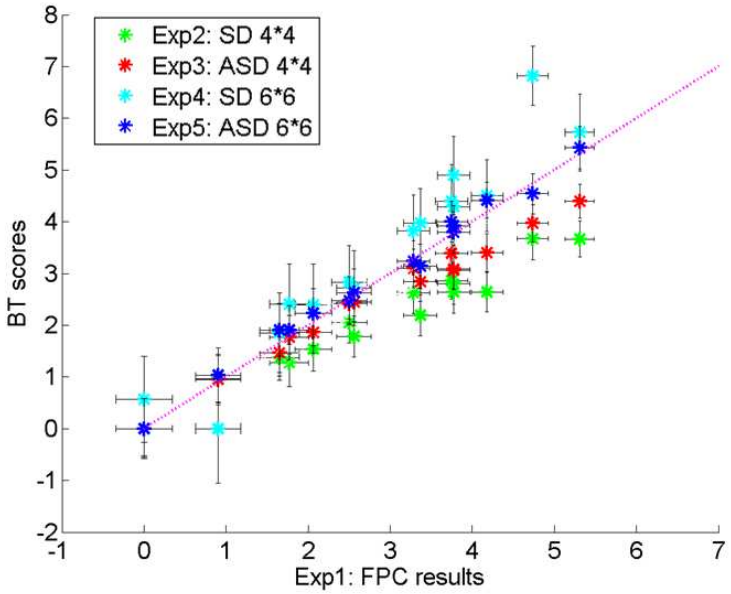


Figure 5.2: Scatter plot of the BT scores between the ground truth (Exp1) and other test results. The pink line is used as a reference with slope = 1. The error bars represent the confidence intervals of the model fit for the test results.

Table 5.2: 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

the experimental results showed high correlation with the ground truth. The results of Experiment 5 show the highest consistency, then follows the results of Experiment 3. The results of Experiment 2 and 4 showed lower consistency. This figure indicates that the ASD method performs better than the original SD method.

To evaluate their correlations with the ground truth more precisely, the Pearson Linear Correlation Coefficient (CC), Spearman Rank-order Correlation Coefficient (ROCC) and the Root Mean Square Error (RMSE) are used as the criterions. In particular, the RMSE was calculated directly on the tested data and the ground truth data. The results are shown in Table 5.2.

According to the CC, ROCC and RMSE values, it is 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 kept robust in this conditions indicating its efficiency and reliability.

It is interesting to find that the results of Experiment 5 have even higher correlation with the ground truth than the results of Experiment 3 while there are more influence factors in Experiment 5. The explanation is that the results of Experiment 5 are obtained with more information of other stimulus comparisons, i.e., the comparisons between other stimuli also provide information on the estimates. The accuracy of the estimation scales is highly dependent on the number of comparisons (see [10]) and the selection of the pairs.

5.3.2 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, where $D(i,j) = V_i - V_j$. 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.

As we already introduced in Chapter 4 Section 4.2, observation errors come from two aspects: one is from observers' attentiveness, the other is from the reduced number of observations. Firstly the influence from observation errors is analyzed. Then, the influence from the dependency of the voting on irrelevant stimuli is investigated.

Influence from observation errors

According to the Bradley-Terry model, the distance between the two stimuli $V_i - V_j$ is related to P_{ij}, where

$$V_i - V_j = \log \frac{P_{ij}}{1 - P_{ij}} \tag{5.1}$$

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

$$\begin{aligned} C_{gt,sd4\times4}(i,j) &= D_{gt}(i,j) - D_{sd4\times4}(i,j), i < j \\ C_{gt,asd4\times4}(i,j) &= D_{gt}(i,j) - D_{asd4\times4}(i,j), i < j \end{aligned} \tag{5.2}$$

The histogram of the errors C are shown in Figure 5.3, 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 the same, thus, it is assumed that there is no influence of the observers' characteristics (e.g., gender distribution, age, 3D viewing experience) on the two experimental results. As shown in Figure 5.3,

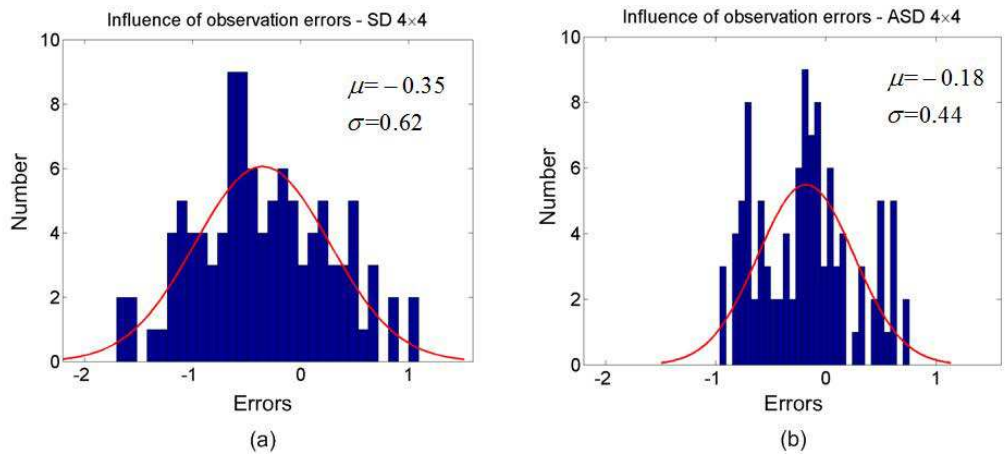


Figure 5.3: The histograms of $C_{gt,sd4\times4}$ and $C_{gt,asd4\times4}$. The red curves are fitted gaussian curve with mean values and variances. (a) is the results of Experiment 2. (b) is the results of Experiment 3.

the mean shift, the variance of the histogram for the SD method are larger than for the ASD method, which indicates that when the observation number is small and the raw pair comparison data is influenced by the observers' "wrong" selections, the ASD method is more reliable than the SD method.

Influence from irrelevant stimuli

In this section, we analyze another factor that may affect the results, i.e., the influence from irrelevant stimuli. In Experiment 4 and 5, besides Stimuli 1 to 15, 21 other stimuli were added. These 36 stimuli were arranged in a 6×6 matrix with the upper left sub-matrix being exactly the same as in Experiment 2 and 3. Thus, in this test, both the observation errors and the influence of the added stimuli would affect the results. Assuming the observation errors can be eliminated by subtracting the results from Experiment 2 and 3, then, the matrices $C_{sd6\times6-sd4\times4}$, $C_{asd6\times6-asd4\times4}$ represent the influence from other stimuli, as shown in the following Equation (5.3).

$$\begin{aligned} C_{sd6\times6-sd4\times4}(i,j) &= D_{sd6\times6}(i,j) - D_{sd4\times4}(i,j), i<j \\ C_{asd6\times6-asd4\times4}(i,j) &= D_{asd6\times6}(i,j) - D_{asd4\times4}(i,j), i<j \end{aligned} \tag{5.3}$$

The same processing as the previous section was performed, the histogram of the errors C are calculated and shown in Figure 5.4, 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 increases 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.

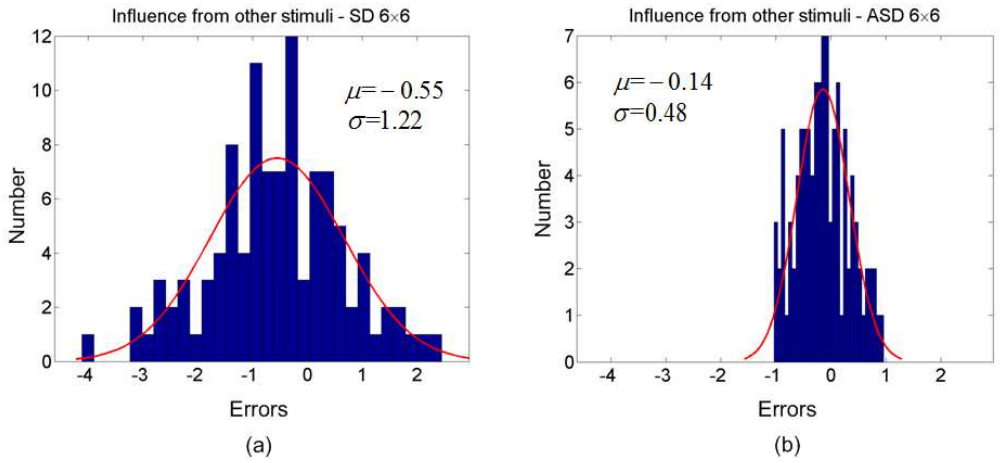


Figure 5.4: The histogram of $C_{sd6\times6-sd4\times4}$ and $C_{asd6\times6-sd4\times4}$. The red curves are fitted gaussian curve with mean values and variances. (a) is the results of Experiment 4. (b) is the results of Experiment 5.

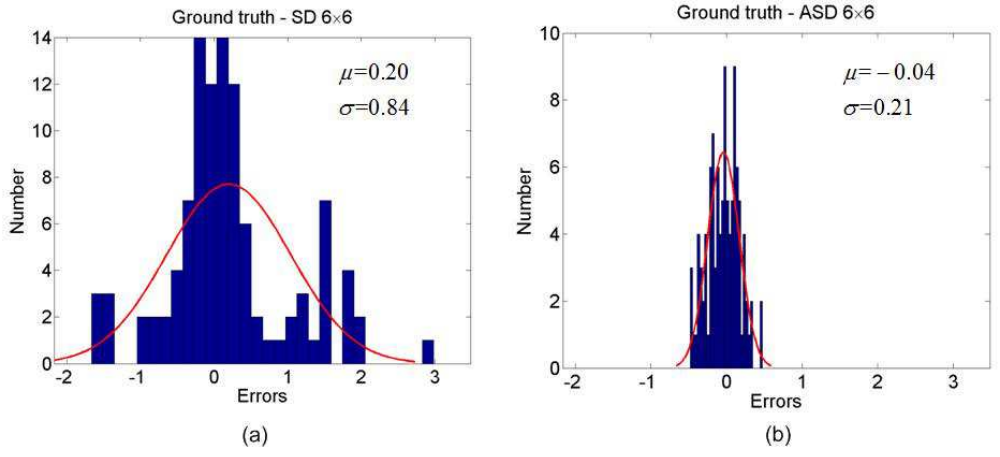


Figure 5.5: The histogram of $C_{gt,sd6\times6}$ and $C_{gt,asd6\times6}$. The red curves are fitted gaussian curve with mean values and variances. (a) is the results of Experiment 4. (b) is the results of Experiment 5.

Compare with the ground truth

If compare the results of Experiment 4 and 5 with the ground truth, the influence from both the observation errors and the irrelevant stimuli can be estimated, as shown in Equation 5.4.

$$\begin{aligned} C_{gt,sd6\times6}(i,j) &= D_{gt}(i,j) - D_{sd6\times6}(i,j), i < j \\ C_{gt,asd6\times6}(i,j) &= D_{gt}(i,j) - D_{asd6\times6}(i,j), i < j \end{aligned} \tag{5.4}$$

The histogram of the errors C are shown in Figure 5.5. The results indicate that the ASD method is robust to influences in the subjective experiments than the original SD method. The estimation accuracy of Experiment 5 is higher than Experiment 3 is due to the large number of total observations in the test, where the relationship of these 15 stimuli with other stimuli also generate information for the estimations.

5.4 Conclusions

In this study, the proposed ASD method was evaluated by a set of subjective visual discomfort experiments in S-3DTV. The performances of the ASD method are evaluated 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, and it also showed higher robustness against observation errors and dependence of comparisons. Due to the efficiency and robustness of the ASD method, paired comparison experiments become feasible with a reasonably large number of stimuli for the assessment of visual discomfort in 3DTV. As visual discomfort is one important dimension of QoE in 3DTV, the proposed efficient paired comparison method is also supposed to be applicable to subjective assessment on QoE.

6

Application of the OSD method in evaluation study of the Preference of Experience in 3DTV

The performances of the proposed efficient paired comparison designs have been verified by subjective experiments in Chapter 5. Thus, in this chapter, it was used for the subjective evaluation study of QoE in 3DTV. In particular, the proposed design is used to analyze how the influence factors affect the QoE, for example, the stereoscopic display technology, test environment, individual differences on age, gender, 3D viewing experience, etc.

6.1 Introduction

As we introduced in the very beginning of this thesis (Chapter 2), QoE may be influenced by many factors. Let's take the broadcasting chain as an example. The shooting of the source video sequence would induce different kinds of geometric distortions, or the non-optimized distribution of the disparity, which would induce visual discomfort or even visual fatigue issues. Due to the limitation of the transmission bandwidth, the video sequences have to be encoded. Different coding schemes would generate different image quality or depth quality. After the transmission, the rendering format of the video sequences, i.e., video format, would also have influence on QoE. For example, the resolution of the Side by Side format would be halved compared to the Full HD format, thus, the perceived image quality or depth quality would be influenced. Furthermore, the influence from display technology

should also be taken into account as studies already showed that different display technologies performed differently in terms of visual quality, depth rendering ability or visual comfort.

3DTV for home entertainment is the mainstream for the near future. An evidence is the setup of DVB 3DTV Phase 2, which aims at providing 3D services but compatible with 2D STB (*Set-top box*). However, most of the studies about 3D QoE nowadays were conducted in controlled lab environments where the observer's experience may differ from watching TV at home. To conduct a more reliable and systematical study on QoE, the home-like test environment is selected in this study, and the objective of this study is the evaluation on the influence factors (IF) of typical broadcasting chain on QoE by using the paired comparison method. The IFs considered in the study are video content, display technology, video encoders, image format, observers, test environment, etc.

In this study, we propose using PoE (*Preference of Experience*) to specify the outcome of the QoE assessed by paired comparison. As observers only provide their binary preference on each pair, a mathematical analysis model for pair comparisons, e.g., Bradley-Terry model [10], is needed to convert this binary data to scale values for all stimuli. Thus, PoE is a scale value after data conversion from the paired comparison data. It serves the same purpose and is comparable to the Mean Opinion Scores (MOS) obtained by, for example, the Absolute Category Rating (ACR) method.

Two experiments were designed in this study. One aimed at evaluating the influence of different video contents (SRC). The other aimed at evaluating the influence of typical Hypothetical Reference Circuits (HRC) in broadcasting. Both experiments were conducted in two labs. One lab used the polarization-multiplexed display technology and the other lab used the time-multiplexed technology. Thus there were in total 4 experiments in this study. As this study focused on the study of influence factors in the home environment, the test environments of the two labs were designed as close to living rooms as possible, and differences were accepted between the installation of the test rooms in the two labs, which is considered as a Context IF.

Six SRCs and twenty HRCs, including different video encoders, bit rates and image formats were selected in this study. Pair comparison was used in the subjective test. To reduce the number of comparisons, the ORD method was adopted which have been evaluated by Monte Carlo simulation experiments in Chapter 4.

Table 6.1: List of source video sequences.

SRC	Barrier	Castle	Rome	Soccer	Tree branches	Umbrella
Preview						
Description	A car is approaching a barrier gate, then the barrier gate opens	A camera is panning behind some old arches which overlooks a castle	Camera pan showing a fountain (Fontana di Trevi)	Scene 1: two players are passing the ball and score. Scene 2: the goal keeper fails to catch the ball	Tree branches and leaves are moving with the wind	A man is playing with an umbrella under a tree
SI	59	59	57	89	101	74
TI	21	15	22	38	14	19
DSI	20.42	5.75	5.1	24.7	23.02	17.02
DTI	15.43	0.9	3.74	18.08	13.63	15.24
D+	6	18	12	7	3	5
D-	9	16	17	10	9	17

6.2 Test Materials

6.2.1 Source Video Sequences

Six stereoscopic Full HD (1920×1080) video sequences were used as SRCs. The duration of the video sequences is 16 seconds except for one (*Umbrella*) which is 13 seconds. The frame rate is 25fps for all video sequences. They were chosen in such a way that they feature as many characteristics as possible, including spatial properties (textured versus uniform areas, contours and gradients, etc.), temporal properties (amount and type of parallax and scene motion, etc.), and depth properties (small and large depth budget, distribution of the depth budget, pop out effects, etc.). The scenes are summarized in Table 6.1. The video sequences *Barrier*, *Soccer*, *Tree branches* and *Umbrella* are from the NAMA3DS1 database [136]. SI and TI in the table are Spatial perceptual Information and Temporal perceptual Information respectively, as described in ITU-T Recommendation P.910 [59]. DSI and DTI are SI and TI calculated on depth maps which were generated by a disparity estimation algorithm based on a first order primal-dual convex optimization algorithm proposed by Chambolle *et al.* [16]. The maximum crossed (D+, objects projected in front of the screen) and uncrossed (D-, objects projected behind the screen) disparities are provided as well. More details can be found in [136].

6.2.2 Hypothetical Reference Circuits

In order to study the influence factors of typical broadcasting system on PoE, twenty HRCs were considered in the tests, including different encoders, allocated bitrates, and image formats (2D/3D, Full HD, SBS, and FCC). The list of the HRCs is shown in Table 6.2.

The first three HRCs are reference conditions, while the remaining seventeen were generated using five different commercial encoders (E0 to E3 and JM ver-

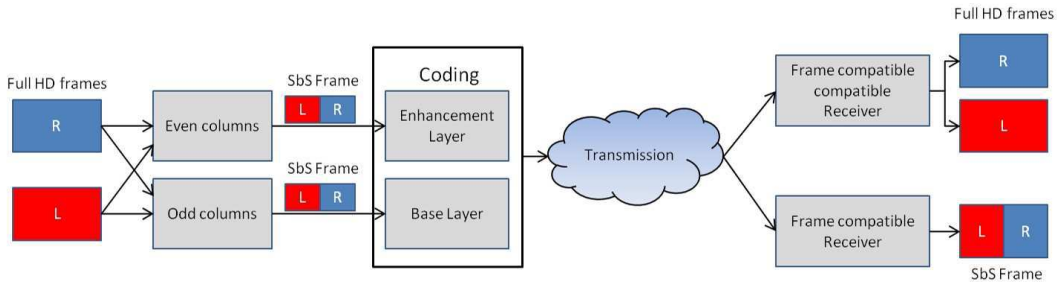


Figure 6.1: The flowchart of processing FCC format

sion 18.2), thus, being considered as distorted sequences. E0 to E3 are different encoder providers. Three video formats are considered which are Full HD format, SBS format and FCC format. "Full HD" indicates that each view has a full HD resolution. SBS is a frame compatible format, in which the spatial resolution of the left and right views are reduced by half in the horizontal direction, and both views are joined together conforming to a conventional 2D frame that can be processed by traditional 2D Full HD broadcasting chains. FCC provides full resolution for both views, while keeping backward compatibility with current generation of decoding devices limited to SBS format. The processing flowchart is shown in Figure 6.1. For each frame of the stereo views, the even columns are taken to form one SBS frame (with half the original columns), and the odd columns are taken to form another SBS frame. Then two individual SBS frames (with half horizontal size) are obtained, so one of them is used in the base layer, and the other one in the enhancement layer. This way the frame-compatible systems are able to obtain a valid SBS frame, and the new systems will be also able to get both stereo views in Full HD.

It should be noted that in this study, while the image resolution was Full HD, when it was prepared for the polarized display, each second line was removed according to the spatial position of the polarization filter in front of the display. Therefore, the vertical resolution was only half HD.

Both 2D and 3D videos were considered in this study. Except for the three reference HRCs, the 3D SBS and 2D Full HD video sequences were encoded using a conventional H.264/AVC (advanced video coding) hardware codec, a H.264/MVC (multi-view video coding) codec, and the JM v18.2 H.264/AVC codec. The bitrates chosen in this study are 6 Mb/s, 9 Mb/s, 12 Mb/s and 20 Mb/s which are typical bitrates for satellite transmission [2] and result in minor visual degradations.

Table 6.2: List of processing conditions (HRCs)

Index	Encoder	Standard	Image format	Bitrate
1	Ref.	-	3D Full HD	-
2	Ref.	-	3D SBS	-
3	Ref.	-	2D Full HD	-
4	E0	H.264/MVC(Hardware)	3D Full HD	20 Mb/s
5	E1	H.264/MVC	3D Full HD	12 Mb/s
6	E1	H.264/MVC	3D Full HD	9 Mb/s
7	E1	H.264/MVC	3D Full HD	6 Mb/s
8	E2	H.264/AVC(Hardware)	3D SBS	12 Mb/s
9	E2	H.264/AVC(Hardware)	3D SBS	9 Mb/s
10	E2	H.264/AVC(Hardware)	3D SBS	6 Mb/s
11	E3	H.264/MVC	3D Full HD	12 Mb/s
12	E3	H.264/MVC	3D Full HD	9 Mb/s
13	E3	H.264/MVC	3D Full HD	6 Mb/s
14	E3	H.264/MVC	3D FCC	12 Mb/s
15	E3	H.264/MVC	3D FCC	9 Mb/s
16	E3	H.264/MVC	3D FCC	6 Mb/s
17	JM18.2	H.264/AVC	3D Full HD	9 Mb/s
18	E2	H.264/AVC(Hardware)	2D Full HD	12 Mb/s
19	E2	H.264/AVC(Hardware)	2D Full HD	9 Mb/s
20	E2	H.264/AVC(Hardware)	2D Full HD	6 Mb/s

6.3 Experimental design

6.3.1 Experiment 1

The objective of Experiment 1 is to analyze the influence of video content on PoE with different display technologies. All SRCs were considered in this experiment. To make the pair comparison test feasible, 9 out of 20 HRCs (HRC1, 2, 3, 5, 7, 10, 16, 18, 20) were selected which led to 18 pairs for each SRC using the OSD method. So, each observer evaluated 108 pairs. The 9 HRCs were selected in such a way that different encoders (references, E1, E2, E3), different bitrates with the same encoder (HRC5 and 7, HRC18 and 20), and different image formats (Full HD, SBS, FCC, 2D/3D) were included.

The ordering of the PoE of all stimuli was visually estimated by experts in the field of 3D quality assessment resulting in the estimated descending index vector $\boldsymbol{d}$ = $(1, 5, 2, 7, 16, 10, 3, 18, 20)$. In addition, some particular HRC pairs were included for comparison in this experiment, e.g., HRC{1, 2} for the comparison of reference videos, HRC{7, 10} for the comparison of different image format with the same bitrate and HRC{10, 20} for the comparison of 2D and 3D with the same encoder. Based on all these requirements and constraints, the ordering indices $\{d_i, d_{i+1}\}$ cannot strictly satisfy the requirements that they should be placed in the same column or row, thus, it has been liberalized to that $\{d_i, d_{i+n}\}$, $n \leq 2$ are in the same column

or row. The arrangement of the 9 HRCs was designed as follows:

$$\begin{matrix} 1 & 2 & 5 \\ 20 & 7 & 10 \\ 3 & 16 & 18 \end{matrix}$$

Following the rules of the OSD, pairs which are in the same column or row are compared. Thus, each stimulus is compared with 4 other stimuli, which leads to 18 pairs for one SRC.

An imbalance of the randomization of the stimuli would affect the paired comparison results, thus, restrictions for the stimulus randomization are defined as follows:

1. The presentation order for each HRC should be as balanced as possible avoiding, for example, that HRC1 is always presented on the left.
2. The presentation order of the video sequences should be as random as possible. In particular, no observer watches the same SRC in two consecutive presentations.

6.3.2 Experiment 2

Experiment 2 focused on the comparison between the PoE of different HRCs, thus, all 20 HRCs were included in this test. The ORD method was used here to place the 20 HRCs into a matrix of size 4×5. According to the visual verification by the same experts in Experiment 1, the estimated rank order of the HRCs was $\boldsymbol{d}$ = $(1, 4, 5, 11, 14, 6, 12, 15, 17, 2, 8, 9, 7, 13, 16, 10, 3, 18, 19, 20)$. In addition, some particular pairs are required in this study, for example, HRC{1, 2, 3, 4} for high quality conditions, HRC{2, 5, 11, 14} and HRC{15, 9} for different encoders and image formats at a bitrate of 12Mb/s and 9Mb/s, and HRC{6, 12, 17} for the performance of different encoders. To satisfy the requirements and constraints of the experimental objectives (as in Experiment 1), the restriction is the same as Experiment 1, i.e., $\{d_i, d_{i+n}\}$, $n \leq 2$ are placed in the same column or row of the matrix. Thus, the HRC layout is designed as follows:

$$\begin{matrix} 1 & 4 & 3 & 16 & 13 \\ 2 & 5 & 11 & 14 & 9 \\ 8 & 17 & 12 & 6 & 15 \\ 18 & 20 & 19 & 10 & 7 \end{matrix}$$

Following the rules of the ORD, pairs which are in the same column or row are compared. Thus, each stimulus is compared with 7 other stimuli, which leads to 70 unique pairs for one SRC.

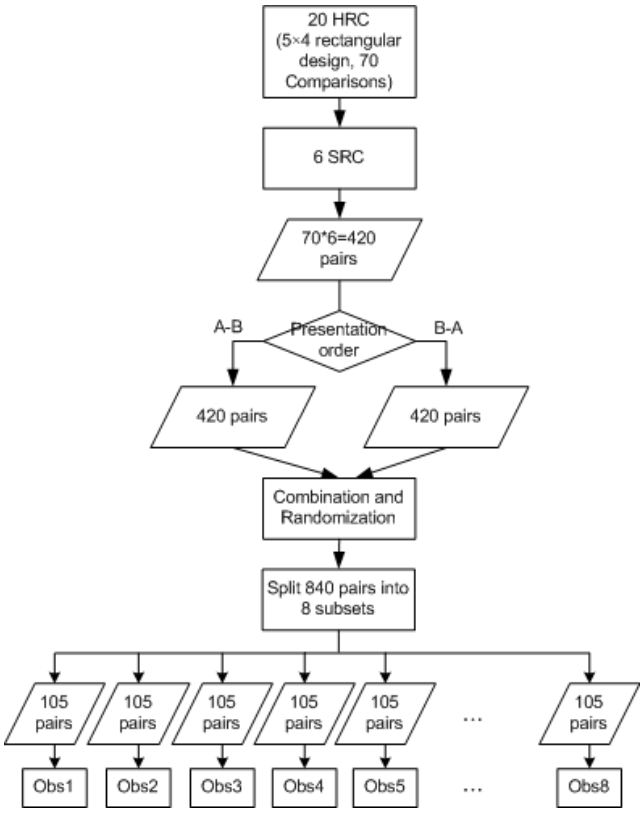


Figure 6.2: The process of distributing one whole observation to 8 observers.

There were 6 SRCs in the tests, thus, the total number of pairs is $6 \times 70 = 420$. As each pair $\{S_A, S_B\}$ should be presented in a different order, e.g., S_A on the left screen, S_B on the right screen $\{S_A S_B\}$ and vice versa $\{S_B S_A\}$ to avoid the impact of dissimilarities between both displays, there would be 840 votes for one complete observation. Since it is infeasible to conduct such a test by only one observer as the required time is approximately 840× (16s displaying + 5s voting) = 3.5 hours, we decided that each group of 8 observers conducted one complete observation (20 HRCs × 6 SRCs), which led to 105 pairs for each observer as shown in Figure 6.2. In this way, the viewing time of the test for one observer is approximately 40 minutes (105 pairs× (16+5)s = 36.7 minutes) excluding breaks, which is a reasonable duration for subjective quality assessment tests. This solution is under the assumption that the voting is independent for all observers. This assumption can be verified by comparing the results of Experiment 2 with Experiment 1, since the latter one is a subset of Experiment 2, in which one observer conducted the whole subset test.

For the randomization of the presentation order for each stimulus pair, due to the fact that one complete observation was conducted by 8 observers, besides the restrictions in Experiment 1, some additional restrictions for the 8 observers are defined as follows:

1. The number of SRC for each observer should be as balanced as possible, e.g., no observer watches only two SRCs in the test.

Table 6.3: Overview of the experiment setup in two labs.

Experimental Setup			UPM		IVC	
Observers	Experiment		Exp.1	Exp.2	Exp.1	Exp.2
	Number of observers (number of faculties in university)		30 (0)	32 (24)	30 (0)	32 (0)
	Gender (m/f)		19/11	26/6	15/15	17/15
	Age range (mean)		23-52 (31.5)	21-49 (29.9)	20-53 (32)	20-66 (29)
Equipment	Display	Model	LG DM2350D-PZ		Philips 46PFL9705H	
		Size	23"		46"	
		Resolution	1920*1080		1920*1080	
		Refresh rate	60Hz		400Hz	
		3D technology	Polarized		Active shutter glasses	
	Luminance		Living room condition		Living room condition	
	Viewing distance		90cm (3.1H)		172cm (3H)	
	Voting interface		Keyboard (left and right arrow)		Three buttons (left, right, validate) on a touch screen.	

2. The frequency of occurrence of each HRC should be as balanced as possible for each observer, e.g., HRC1 is not balanced if observer 1 watches HRC1 only once but observer 2 watches HRC1 20 times.

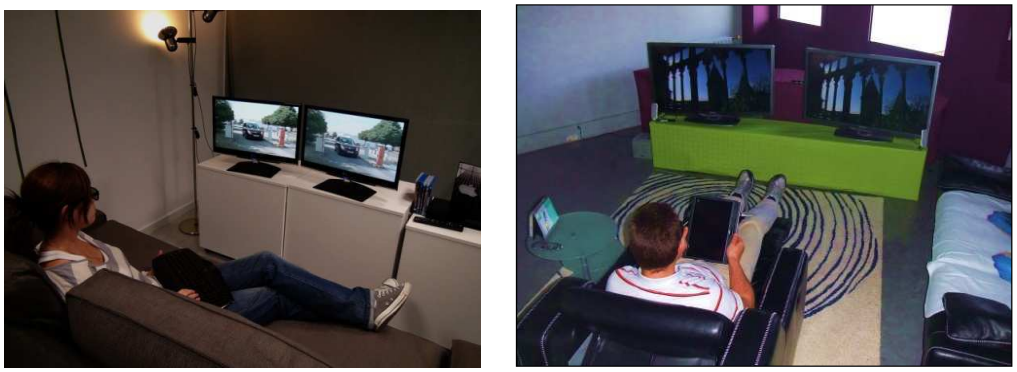
6.4 Experimental setup

The experiments in this study were conducted in two labs. One was the Image and Video Communication (IVC) lab of the University of Nantes in France using 3D displays with time-multiplexed technology, and the other was the Lab-3DTV of the Campus of Montegancedo in the Universidad Politécnica de Madrid (UPM) in Spain, using polarization-multiplexed display technology.

6.4.1 Equipment and environment

Viewing environment is one of the Context IF which may affect the QoE [18]. To better understanding the viewer's experience in home environment, the home-like viewing environment was selected. The test rooms in the two labs were set up as close to a typical living room as possible. Differences were accepted between the installation of the two test rooms. For example, in UPM, the illumination of the test room was constant. In IVC, as there were two windows behind the screens, the illumination was changing with the sun light. The pictures of the two test rooms are shown in Figure 6.3. The observer characteristics and equipments used in the tests in each lab are listed in Table 6.3.

Side by side pair comparison method was used as it's easier for observers when stimulus qualities are close. The technical solution for synchronizing the two active 3D displays is as follows: A 3D playout software has been specifically developed to synchronize two high performance PCs featuring a Blackmagic extreme3d+ graphics card. They were synchronized by a genlock generator such that uncompressed Full HD 3D video playback at exactly the same refresh rate could be achieved.



(a) UPM room (b) IVC room

Figure 6.3: The test environment of UPM and IVC labs.

The Philips displays were driven by a customized firmware to allow for phase synchronous display of the left and right view and the corresponding emission of the infrared signal to the shutter glasses.

6.4.2 Observers

All of the observers in Experiment 1 are naive observers in the sense that they are not directly concerned with television picture quality as part of their normal work, and are not experienced assessors. In Experiment 2, observers in IVC are naive observers, while in UPM, some of the observers are faculty members in the university who are more used to 3D technologies and have more experience on 3D visualization than naive observers although their work is not related to quality evaluation. The detailed information of all observers is listed in Table 6.3. 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 observers in this study passed the pre-test vision check.

6.4.3 Test Process

The question for observers after watching each pair of sequences was “Which one do you prefer?”, thus, it is an overall preference on each stimulus pair. The whole test included a training session and a test session. There were five training sequences in the training session. After viewing each pair of video sequences there was a message shown on the screen to ask the observers to start voting. Observers needed to select the video they preferred by pressing the corresponding selection button on the keyboard or touch screen as shown in Figure 6.4. During the training session, all questions of the viewers were answered. It was ensured that after the

(a) UPM voting keyboard (b) IVC touch pad voting interface

Figure 6.4: The voting interface of UPM and IVC labs.

training session, all of the viewers understood the process and task of this experiment.

The test session was split into two sub-sessions with a similar duration. There was a break of about 5 minutes between the two sub-sessions. The duration of each experiment was approximately one hour.

6.5 Results of Experiment 1

6.5.1 Analysis on the influence of video content on PoE by the Bradley-Terry model

The Bradley-Terry (BT) model [45][10] was used in this study to generate PoE values for all stimuli. Based on the Bradley-Terry model, PoE is a relative scale value. To facilitate the comparison, the HRC with the lowest PoE score is usually set as a reference with PoE = 0. All other HRCs' PoE scores are thus estimated afterwards. In this study, HRC16 is set as a reference for all SRCs (HRC16 does not always generate the lowest PoE in different SRCs, but its PoE across SRCs is the lowest as shown in Figure 6.6).

Intra-lab analysis

In this part, the PoE scores of each SRC for different HRCs are compared within each lab to evaluate the influence of video content on PoE. The PoE of each SRC are shown in Figure 6.5.

To evaluate the correlation between SRCs within each lab, the Pearson Linear Correlation Coefficient (PLCC) is calculated as shown in Table 6.4 and Table 6.5. Both of the results indicate that *Castle*, *Rome*, *Tree branches* and *Umbrella* have higher correlations with each other. The *Barrier* and *Soccer* sequences correlate poorly with other SRCs in terms of PoE with respect to the evaluated HRC conditions. Considering the distribution of the PoE in Figure 6.5, the absolute PoE for

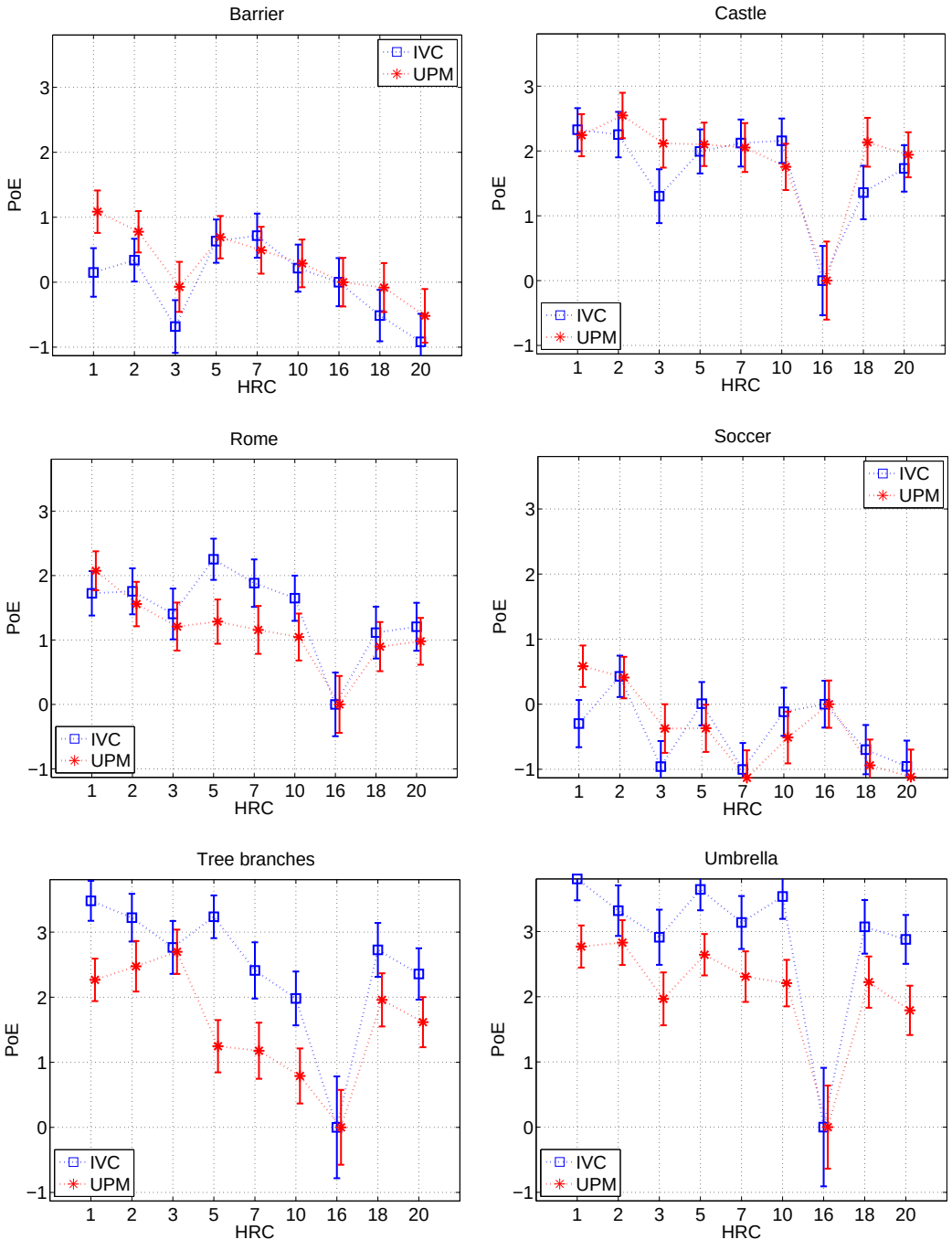


Figure 6.5: The comparison results between IVC and UPM for different SRCs. HRC16 is set as reference with PoE = 0. The error bars represent the confidence intervals of the Bradley-Terry model fit.

Table 6.4: PLCC matrix for the SRCs in IVC, correlations higher than 0.8 are marked in bold

	Barrier	Castle	Rome	Soccer	Tree-b	Umbrella
Barrier	-	0.38	0.48	0.51	0.11	0.19
Castle	0.38	-	**0.91**	0.05	**0.82**	**0.93**
Rome	0.48	**0.91**	-	0.03	**0.84**	**0.91**
Soccer	0.51	0.05	0.03	-	-0.05	-0.09
Tree-b	0.11	**0.82**	**0.84**	-0.05	-	**0.91**
Umbrella	0.19	**0.93**	**0.91**	-0.09	**0.91**	-

Table 6.5: PLCC matrix for the SRCs in UPM, correlations higher than 0.8 are marked in bold

	Barrier	Castle	Rome	Soccer	Tree-b	Umbrella
Barrier	-	0.35	0.68	0.66	0.18	0.58
Castle	0.35	-	**0.84**	-0.04	**0.81**	**0.95**
Rome	0.68	**0.84**	-	0.38	0.73	**0.90**
Soccer	0.66	-0.04	0.38	-	0.22	0.10
Tree-b	0.18	**0.81**	0.73	0.22	-	0.67
Umbrella	0.58	**0.95**	**0.90**	0.10	0.67	-

Barrier and *Soccer* are much smaller than for the other SRCs, which means that the degradations due to these HRCs did not affect the preference as much as for the other SRCs.

A possible explanation for this results is that in *Castle*, *Rome*, *Tree branches* and *Umbrella*, the motion and depth changes are slow, while *Barrier* and *Soccer* contain slightly faster in-depth motion objects which move from background to foreground (the car in *Barrier*, the ball and the player in *Soccer*). The fast motion may attract the observer's attention and blur his perception on distortions. In addition, the in-depth motion may have influence on visual discomfort which would affect the PoE. The scene cuts in *Soccer* may also affect the PoE. Further study is required.

Inter-lab analysis

In this section, the results from the two labs are compared for each SRC. PLCC, Spearman's rank correlation coefficient (SROCC) and Root Mean Square Errors (RMSE) after fitting the IVC PoE values to UPM are used to evaluate the correlations between the two labs. The results are shown in Table 6.6.

As shown in Table 6.6, the PLCC for the SRC *Umbrella* is the highest. However, it could be found in Figure 6.5 that in the lab IVC using shutter glasses, the preference of each stimulus to the PoE reference (HRC16) is higher than in UPM using polarized display. A similar phenomenon can be found in *Tree branches*. In these two sequences, most of the screen area shows tree branches and leaves which con-

Table 6.6: Comparison of PoE scores between IVC and UPM for each SRC condition

	Barrier	Castle	Rome	Soccer	Tree-b	Umbrella
PLCC	0.7914	0.8523	0.7899	0.7265	0.8061	0.9650
SROCC	0.7500	0.4667	0.7667	0.7000	0.7000	0.8167
RMSE	0.2732	0.2020	0.2763	0.4041	0.4222	0.1750

tain very high spatial frequencies. Due to the characteristics of polarized displays, i.e., the horizontal resolution is halved, the high spatial frequency components may decrease the discrimination of slightly different degradations.

In *Barrier*, *Castle* and *Soccer*, there is no significant difference for most of the HRCs. But for HRC3, the PoE in UPM is significantly higher (according to confidence intervals) than in IVC. This means that observers showed higher preference on 2D video sequences when using polarized displays than using shutter glasses.

6.5.2 Analysis on the influence of video content on PoE by Barnard's exact test

In this analysis, the Barnard's exact test is used to examine the significant differences between the pair comparison data of the two display technologies. Through the comparison of all 18 pairs tested in the experiments, the significantly different pairs are detected on a significance level of 0.05. The results are shown in Table 6.7.

The number of significantly different pairs in *Tree branches* is the largest among all SRCs. For *Barrier* and *Rome*, they show higher correlation between the results of the two labs as only one pair is significantly different. It should be noted that in the case of *Tree branches*, there are 3 out of 6 pairs belonging to the condition of 2D compared with 3D (HRC{1,3}, HRC{5,18}, HRC{7,20}). The raw paired comparison data showed that most observers preferred 2D conditions when using polarized displays (UPM) but preferred 3D conditions when using shutter glasses (IVC).

If taking the occurrence frequency of each HRC pair into account, it could be found that the pair HRC{1, 5} and pair HRC{3, 20} occur more often than other HRC pairs. In detail, as shown in Table 6.7, in the video sequences of *Barrier*, *Rome* and *Soccer*, the observers preferred 3D coded Full HD video (HRC5) to the 3D reference Full HD video (HRC1) when using shutter glasses in the IVC lab. However, the opposite preference was found in the UPM lab. In *Tree*, there is no significant difference for HRC1 and HRC5 in IVC, but the preference on HRC1 is significant in UPM.

For the condition of 2D reference video (HRC3) and 2D distorted video (HRC20),

Table 6.7: Barnard's exact test results for each SRC condition

SRC	Sig. diff. pairs HRC{A,B}	Type	p-value	IVC vote A:B	UPM vote A:B
Barrier	HRC{1,5}	Ref.3D vs 3D	0.01	8:22	18:12
Castle	HRC{1,2}	Ref.3D vs Ref.3D	0.03	16:14	9:21
	HRC{3,20}	Ref.2D vs 2D	0.00	9:21	20:10
	HRC{10,18}	3D vs Ref.2D	0.01	21:9	12:18
	HRC{10,20}	3D vs Ref.2D	0.00	23:7	12:18
Rome	HRC{1,5}	Ref.3D vs 3D	0.00	9:21	20:10
	HRC{2,5}	Ref.3D vs 3D	0.02	10:20	19:11
Soccer	HRC{1,5}	Ref.3D vs 3D	0.00	12:18	23:7
	HRC{3,20}	Ref.2D vs 2D	0.00	11:19	24:6
Tree	HRC{1,3}	Ref.3D vs Ref.2D	0.02	20:10	11:19
	HRC{1,5}	Ref.3D vs 3D	0.02	17:13	24:6
	HRC{3,20}	Ref.2D vs 2D	0.01	18:12	27:3
	HRC{5,10}	3D vs 3D	0.03	25:5	18:12
	HRC{5,18}	3D vs 2D	0.02	19:11	10:20
	HRC{7,20}	3D vs 2D	0.02	19:11	10:20
Umbrella	HRC{2,5}	Ref.3D vs 3D	0.01	8:22	17:13
	HRC{7,10}	3D vs 3D	0.03	10:20	18:12

the results in *Castle*, *Soccer* and *Tree branches* indicate that the preference on 2D reference videos in UPM using polarized display is higher than that in IVC.

Both of the results showed that the artifacts in HRC5 and HRC20 have positive influence on PoEs when using shutter glasses.

6.5.3 HRC analysis

Correlation analysis on the Bradley-Terry scores

The paired comparison matrix of HRCs can be obtained by combining the paired comparison matrices of all SRCs. The PoEs across the HRCs are calculated for each lab. The results are shown in Figure 6.6(a).

In the condition of HRC1, 2, 16, 18, and 20, the PoEs of the two labs correlate well. For the condition of HRC3, 5, 7 and 10, the PoEs have slight offset. The PLCC, SROCC and RMSE between the two datasets are 0.8467, 0.7833 and 0.2389, respectively. Thus, in general, the performance of the two display technologies in measuring PoE correlate well.

Comparison of raw data by Barnard's exact test and Monte-Carlo simulation

To evaluate whether the differences between the two test results are significantly, the raw paired comparison data are analyzed. The Barnard test is applied on the

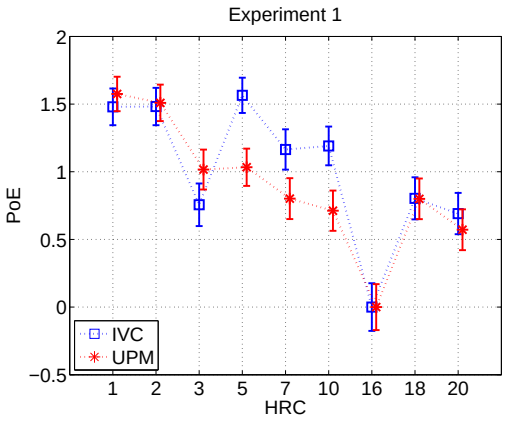
paired comparison data of HRCs of both labs to make a comparison. The results show that 5 pairs are significantly different, $r_t = 5/18 = 0.28$. In order to evaluate if r_t is statistically large (or if the display technology has influence on the results), a Monte-Carlo experiment based on the observer's data is performed as introduced in Chapter 4. We randomly divide the observers of the two labs into two groups (the number of observers of each group is equal) and run the Barnard test. The histogram of $\boldsymbol{r}$ is calculated after 1000 trials of simulation as shown in Figure 6.6(b).

According to the results, the mean of $\boldsymbol{r}$ is 0.1478, the bound of the 95% cumulative probability is $\sigma = 6/18$, and the maximum value in $\boldsymbol{r}$ is $10/18 = 0.5556$. $r_t = 5/18$ is not a low probability event in this case which means the display technology may be not a main factor in relation to the differences between the results from both labs. To inspect in which case $\boldsymbol{r}$ reaches the maximum, the PoE scores of the two groups are calculated and shown in Figure 6.6(c).

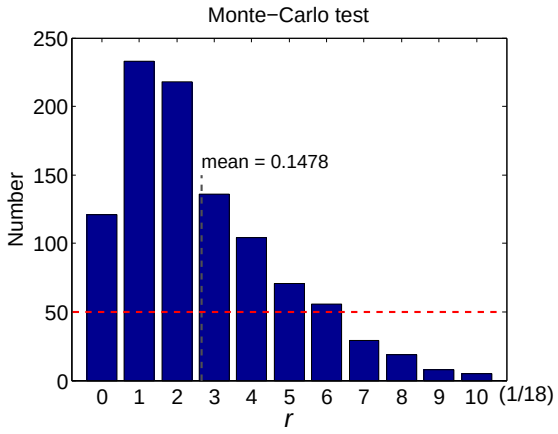
As shown in the figures, the PoEs of 2D conditions (HRC3, HRC18 and HRC20) do not correlate well between the two groups. Observers in Group1 show higher preference in 2D conditions than Group2. As there are 15 IVC observers and 15 UPM observers in Group1, the remaining 30 observers are in Group2, the results indicate that observer may be a predominant influence on the final results which will be studied in Experiment 2.

6.5.4 Discussion

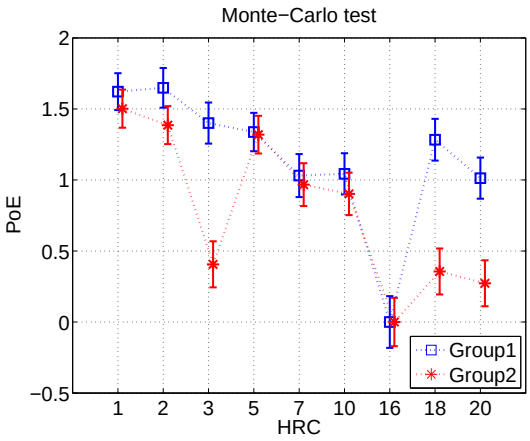
Video content is an important factor in PoE in this study. For instance, motion (planar motion and in-depth motion) might be an important factor in video content as it may attract observer's attention and mask the visibility of distortions in other areas. Moreover, the high spatial frequency components (e.g., tree leaves in *Tree branches* and *Umbrella*) may reduce the discriminability on preference when using polarized displays due to the reduction on resolution. In this study we also found that in some particular video sequence (*Tree branches*, small local motion, small depth indicator, high spatial frequency), the preference on 2D or 3D videos is quite different for the two locations. 2D video presentation is preferred in UPM using polarized displays while 3D is preferred in IVC using shutter glasses. The Monte-Carlo experimental results indicate that human factors might affect the results significantly. In addition, some video contents with particular degradations may enhance the visual perceptual experiences when using shutter glasses. One possible explanation is that the blur introduced by coding may have a similar effect as motion blur: in shutter glasses, the temporal succession of the video frames displayed with temporal left/right eye views, sampling on the shutter glasses is perceived more smoothly due to the removal of high frequency components.



(a)

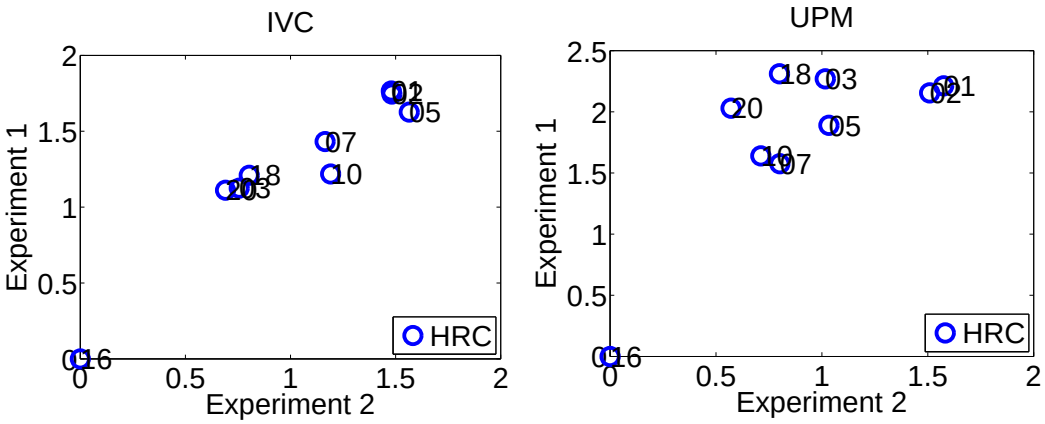


(b)



(c)

Figure 6.6: PoE of Experiment 1 for HRCs. (a) The PoE across HRC of the two labs. The error bar shows 95% confidence intervals for Bradley-Terry model fit. (b) The Monte-Carlo experimental results: the histogram of $\boldsymbol{r}$. (c) An example of the Monte-Carlo experimental results: the PoEs of the two groups with the maximum $r = 10/18$.



(a) (b)

Figure 6.7: The scatter plots of PoE scores of the two experiments. (a) IVC results (b) UPM results

6.6 Results of Experiment 2

Experiment 2 aims at studying the influence of different encoders, video formats, and display technologies on PoE. In this experiment, one complete observation was conducted by eight observers based on the assumption that the voting is independent for all observers. Before further data analysis, this assumption should be verified.

6.6.1 *Observation independency* analysis

Experiment 1 can be considered as a sub-test of Experiment 2. In Experiment 1, one complete observation was conducted by one observer. However, in Experiment 2, one whole observation was conducted by eight observers. To evaluate the *Observation independency* assumption in Experiment 2, the results from both experiments should be compared. The scatter plot of the PoEs of the two experiments are shown in Figure 6.7. The PLCC, SROCC and RMSE of the PoEs of Experiment 1 and Experiment 2 in IVC are 0.9520, 0.9167 and 0.1089. For UPM, they are 0.7637, 0.4667 and 0.2889, respectively. The results in IVC show higher correlation than in UPM.

The Fisher's-mid-p-value test is applied on the common-set of the pair comparison data of the two experiments as the sample size is unbalanced and large. There are in total 8 pairs compared in both experiments, only one of them shows significant difference in IVC (HRC{10, 18}) and two in UPM (HRC{1, 3} and HRC{10, 20}).

The *Observation independency* assumption is verified by the results of IVC as the results from Experiment 1 and Experiment 2 show quite high correlation. However, for the results of UPM lab as shown in Figure 6.7(b), the correlation between

Experiment 1 and 2 is not as high as in IVC. Due to the fact that the difference between Experiment 1 and 2 in UPM may be mostly related to the observers, the factors that may affect the results of UPM in Experiment 2 will be analyzed later in this chapter.

6.6.2 Influence of HRCs on PoE

The PoEs of Experiment 2 across HRCs are shown in Figure 6.8. The PLCC, SROCC and RMSE of the PoEs of both labs are 0.8244, 0.4857 and 0.3338, respectively. According to the results the following conclusions may be extracted:

- Full HD format performs better than FCC format with the same encoder E3. The FCC format with 9Mb/s (HRC15) generate similar PoE (slightly less, but not significantly different) compared to the Full HD format at 6Mb/s (HRC13) which indicates that about 50% more bitrates is necessary for FCC compared to Full HD. One cause of the poor performance of FCC format might be the downsampling process in the generation of the FCC sequences in which no spatial filtering was applied.
- When comparing the reference video sequences, it is very difficult for the observers to vote their preference between SBS and Full HD stereo views (HRC{1, 2}).
- Using the same standard H.264/MVC and the same image format of 3D FullHD, encoder E1 performs significantly better than E3 in both labs (HRC5, 6, 7 vs HRC11, 12, 13).
- Generally, there is no significant difference between the PoE values of HRCs in UPM and IVC except the 2D format (HRC3, 18, 19, 20) and HRC17.
- In UPM using polarized display, the performance of the encoder JM.18.2 (HRC17) is generally worse when compared with the other MVC encoders. However, in IVC this conclusion is inverted. This fact might be caused by the different display technology used in both labs, although further analysis is needed.
- The preference on 2D format in UPM is significantly higher than in IVC. As this phenomenon also occurs in Experiment 1, influence factors will be analyzed in the following section.

6.6.3 *2D/3D Preference*

According to Figure 6.8, the PoEs in HRC3, 18, 19 and 20 do not correlate well between the two labs. As all these four HRCs are 2D conditions, this may indicate that the preference of 2D to 3D is different between the two display technologies. In particular, the results indicate that the 2D video sequences were preferred to 3D

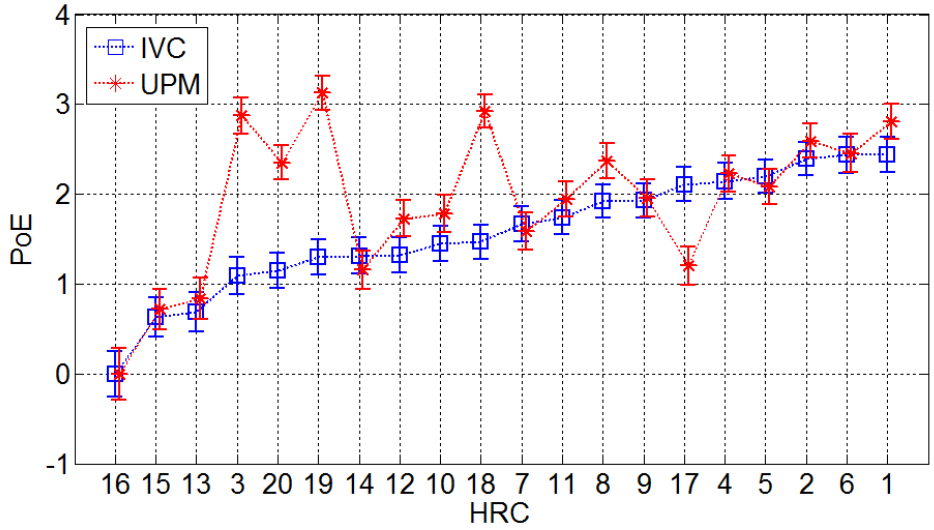


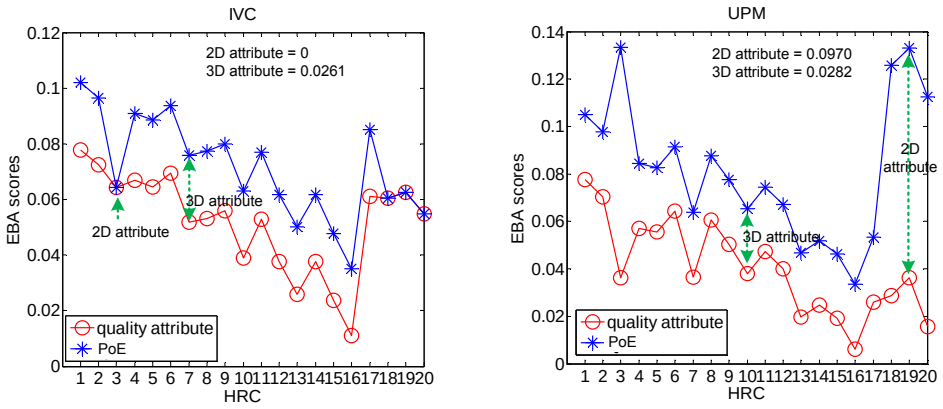
Figure 6.8: Results of Experiment 2: The PoEs across HRCs of the two labs. The error bar shows 95% confidence intervals of the BT model fit. The indices are sorted in ascending order according to the PoEs of IVC.

in the UPM lab and vice versa in IVC lab. Therefore, a hypothesis is developed: "2D is preferred to 3D in UPM lab using polarized displays and 3D is preferred to 2D in IVC lab using shutter glasses". We will refer to it as the *2D/3D Preference*.

EBA analysis

To evaluate and better understand the "2D/3D Preference" in our study, the EBA (*Elimination By Aspects*) model [132] [133] was employed here which has been introduced in Chapter 3 section 3.4.3. EBA allows for a different analysis on paired comparison data. According to EBA, 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 [143]. In our case, we assume that each HRC has its own attribute on PoE, we call it "quality attribute". The 2D and 3D condition can be considered as a second attribute. The sum of the "quality attribute" and the "2D/3D attribute" is the PoE score of each stimulus. According to the EBA model, the "quality attribute" value and the PoE for each HRC are shown in Figure 6.9, the distance between PoE and the "quality attribute" corresponds to the "2D/3D attribute" and it is marked with green arrows in the figure.

As shown in Figure 6.9, when using shutter glasses to watch video sequences, the 3D attribute is larger than the 2D attribute, which means that 3D mode contributes more to PoE than 2D. For polarized screens, the 3D attribute is smaller than the 2D attribute. All of the analysis above verified the "2D/3D Preference", that



(a) (b)

Figure 6.9: The EBA results of Experiment 2. Red point represents quality attribute, Blue point represent PoE, which is the sum of quality attribute and 2D/3D attribute. The 2D/3D attributes for two display technologies are marked in the figure. (a) The EBA scores across HRC of IVC lab. (b) The EBA scores across HRC of UPM lab.

is, observers prefer 2D video sequences to 3D when using the polarized display in UPM. However, 3D is preferred to 2D when using shutter glasses in IVC.

Barnard's exact test results

To analyze which pairs in "2D vs 3D" conditions show significant difference between the IVC and UPM labs, the raw paired comparison data is investigated. The Barnard test is used to evaluate the significant differences between the "2D vs 3D" paired comparison data of the two labs. The significantly different pairs are shown in Table 6.8. The first two columns are the HRC pairs. p_{2D-3D} is the ratio of voting that HRC-2D is preferred in the pair of HRC{2D, 3D}, p-value is the Barnard's exact test result (p <0.05 represents significant difference between the two labs at the significance level of 0.05) except for the last row which is conducted by Fisher's-mid-p-value test due to the large sample size.

All the analysis above verified the results that in our study, the 2D format is preferred to 3D format in the condition that the 3D display is polarized while the conclusions is opposite in the condition of using 3D displays with shutter glasses. Due to the fact that there are many different factors in the two labs except the display technology, the real influence factors for this conclusion will be investigated in the following part.

6.6.4 Possible causes of 2D/3D preference

To investigate the cause of *2D/3D Preference*, besides the most important differences in the setup of the two labs, i.e., display technology, some other influence factors should be analyzed. As this 2D preference also occurs when comparing the

Table 6.8: Significantly different "2D vs 3D" pairs for Experiment 2 of two labs

HRC-2D	HRC-3D	p_{2D-3D}(IVC)	p_{2D-3D}(UPM)	p-value
3	11	18/48 (0.38)	29/48 (0.60)	0.0158
3	12	25/48 (0.52)	35/48 (0.73)	0.0187
3	13	30/48 (0.62)	38/48 (0.79)	0.0411
18	2	15/48 (0.31)	26/48 (0.54)	0.0127
18	7	20/48 (0.42)	33/48 (0.69)	0.0042
19	7	25/48 (0.52)	34/48 (0.71)	0.0318
19	10	19/48 (0.40)	33/48 (0.69)	0.0023
19	12	21/48 (0.44)	34/48 (0.71)	0.0040
20	5	18/48 (0.38)	31/48 (0.65)	0.0052
20	7	19/48 (0.40)	29/48 (0.60)	0.0260
20	10	21/48 (0.44)	31/48 (0.65)	0.0229
20	17	18/48 (0.38)	36/48 (0.75)	0.0001
All 2D	**All 3D**	**249/576 (0.43)**	**389/576 (0.68)**	**0.0000**

Table 6.9: Significant test: comparison of the influence of 3D experience on PoE of Experiment 2 in two labs. * represents there is significant difference within the lab. ** represents there is significant difference between the labs.

2D	3D	IVC (3D experience)	UPM (3D experience)		
		$p_{2D\text{-}3D}$ (naive)	$p_{2D\text{-}3D}$ (faculty)	$p_{2D\text{-}3D}$ (naive)	p-value
3	11	18/48(0.38)	18/36(0.50)	11/12(0.92)	0.0097*
3	12	25/48(0.52)	22/34(0.65)	13/14(0.93)	0.0140*
3	13	30/48(0.62)	32/36(0.89)	6/12(0.5)	0.0058*
18	2	15/48(0.31)	23/42(0.55)	3/6(0.5)	0.6538
18	7	20/48(0.42)	31/46(0.67)	2/2(1)	0.3171
19	7	25/48(0.52)	25/36(0.69)	9/12(0.75)	0.4759
19	10	19/48(0.40)	22/34(0.65)	11/14(0.79)	0.1989
19	12	21/48(0.44)	22/28(0.71)	14/20(0.70)	0.4772
20	5	18/48(0.38)	24/40(0.60)	7/8(0.88)	0.0881
20	7	19/48(0.40)	15/26(0.58)	14/22(0.64)	0.3914
20	10	21/48(0.44)	27/44(0.61)	4/4(1)	0.1206
20	17	18/48(0.48)	18/26(0.69)	18/22(0.82)	0.2317
All 2D	**All 3D**	249/576(**0.43**)**	277/428(**0.65**)	112/148(**0.76**)**	**0.0127***

Table 6.10: Significant test: comparison of the influence of gender on PoE of Experiment 2 in two labs. * represents there is significant difference within the lab. ** and *** represents there is significant difference on males and females between the labs.

HRC-2D	HRC-3D	IVC (gender)			UPM (gender)		
		$p_{2D\text{-}3D}$ (female)	$p_{2D\text{-}3D}$ (male)	p-value	$p_{2D\text{-}3D}$ (female)	$p_{2D\text{-}3D}$ (male)	p-value
3	11	6/23(0.26)	12/25(0.48)	0.0666	8/8(1)	21/40(0.53)	0.0128*
3	12	10/26(0.38)	15/22(0.68)	0.0227*	10/10(1)	25/38(0.66)	0.0281*
3	13	14/25(0.56)	16/23(0.70)	0.2471	9/10(0.9)	29/38(0.76)	0.2118
18	2	6/24(0.25)	9/24(0.37)	0.2622	6/6(1)	20/42(0.48)	0.0511
18	7	5/22(0.23)	15/26(0.58)	0.0080*	6/6(1)	27/42(0.64)	0.0571
19	7	8/22(0.36)	17/26(0.65)	0.0252*	12/12(1)	22/36(0.61)	0.0097*
19	10	6/22(0.27)	13/26(0.50)	0.0688	10/10(1)	23/38(0.61)	0.0166*
19	12	9/26(0.35)	12/22(0.55)	0.0984	9/10(0.9)	25/38(0.66)	0.1097
20	5	7/20(0.35)	11/28(0.39)	0.3979	8/10(0.8)	23/38(0.61)	0.1441
20	7	7/22(0.32)	12/26(0.46)	0.2318	9/10(0.9)	20/38(0.53)	0.0281*
20	10	7/24(0.29)	14/24(0.58)	0.0260*	7/8(0.875)	24/40(0.60)	0.0882
20	17	7/20(0.35)	11/28(0.39)	0.3979	12/12(1)	24/36(0.67)	0.0265*
All 2D	**All 3D**	92/276(**0.33**)***	157/300(**0.52**)**	**0***	106/112(**0.95**)***	283/464(**0.57**)**	**0***

results of Experiment 1 and 2 in UPM, and the only difference between these two experiments were the observers, the influence factors from observers are analyzed based on the observers' pair comparison data. Besides, the possible influence from screen size is also analyzed.

Influence from 3D experience of observers

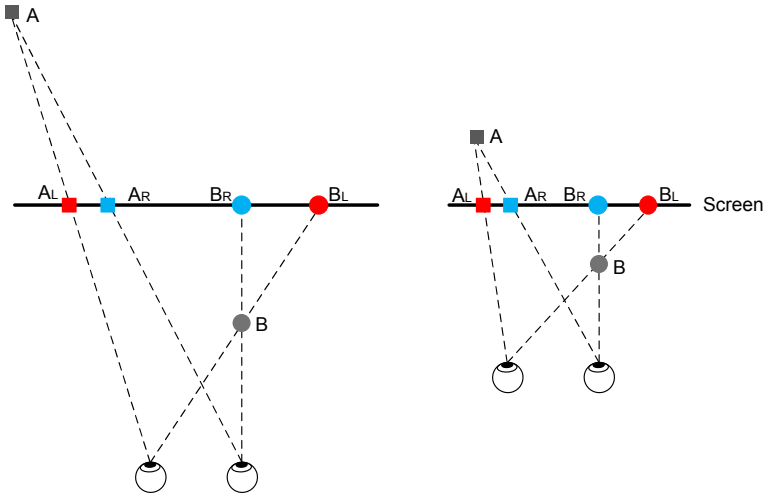
In Experiment 2 of UPM, 24 of 32 observers are faculty members of the university who are more used to 3D technologies than the others. Studies already showed that observers' experience on 3D may affect the subjective test results [117]. Therefore, in this section, the influence of the observers' 3D experience are evaluated.

The Barnard test is applied on faculty member's data and the naive observers' data of UPM. The results are shown in Table 6.9. The 2D preference of the naive observers is significantly higher than that of the 3D experienced observers. The naive observers seemed more critical to 3D videos than the 3D experienced observers. Thus, the observer's 3D experience is an important factor in PoE. Besides, the 2D preference for both faculty members (0.65) and naive observers (0.76) are significantly higher than the results in IVC (0.43)(using Fisher's-mid-p-value test, p-value < 0.05 in both cases) which indicates that besides 3D experience, there are some other factors that affect the PoE in this study.

Influence from the gender of observers

Studies already showed that observer's gender might affect the subjective test results [140][53]. Therefore, the influence of the gender on the *2D/3D Preference* will be evaluated.

As the experiment of UPM was not initially designed for gender analysis, the distribution of observers' gender is biased (6 female and 26 male), thus, the Barnard's



(a) (b)

Figure 6.10: The perceived depth changes due to the size of the screen and the viewing distance. (a) large screen. (b) small screen. Figures are copied from [25] and redrawn.

exact test is applied on the data. The test results according to observer's gender are shown in Table 6.10. p_{2D-3D}(female) = 9/10 means 9 out of 10 observations by female observers selected the 2D video sequence. The results indicate that in IVC, males chose more often 2D conditions than female observers. However, it was the opposite in UPM. Thus, we may conclude that gender is a main factor in PoE. When comparing the performance of the same gender in different labs, the test results indicate that the preference of both genders between the two labs are significantly different (p-value < 0.05). Observers in UPM showed more preference in 2D video sequences, from which we may conclude that there are still some other factors that may have significant impact on PoE.

Influence of screen size

Apart from the difference of the display technologies used in two labs, the screen size in IVC was larger than in UPM. The perceived depth in stereoscopic content is strongly linked to screen size and viewing distance as shown in Figure 6.10. Larger screens could be more impressive, or provide more immersiveness or naturalness than smaller screens. This conclusion is supported by the results of [25]. In [25], the majority of observers prefers to watch 3D content on large size displays, 83% of the observers explicitly said that they would not watch 3D content on a mobile device. In addition, in the test of UPM, some observers reported that the screens were too small after viewing the test environment. Thus, in our study, another possible explanation of the *2D/3D Preference* might be that due to the poor depth rendering ability of the small screen in the home environment, observers prefer

watching 2D videos rather than 3D videos.

According to the analysis on the *2D/3D Preference*, the possible influence factors are display technology, observer's 3D experience, gender, and screen size.

6.6.5 Discussions on other factors

The main difference between the two test labs is the installation of the test rooms. In both labs, the test rooms were set as close to a living room as possible, however, a lot of differences can be found as shown in Figure 6.3. Particularly, in IVC, due to the windows behind the screens, the illumination of the test room was changing with the sun light while in UPM it stayed almost constant. However, as shown in Figure 6.8, the influence from the test environments is not significant for 3D videos which is consistent with the conclusions from [89].

The observer is a very important factor in PoE. The influence from observer's native language, culture/country of origin may have influence on PoE [5]. In this study, the PoE is used only in the way as an overall and averaged experience of the observers. The *Observation independency assumption* in this study was verified by the experimental results in IVC, however, improvement on the test methods are necessary to maximize the probability of gathering the opinion of "all people".

6.7 Conclusion

How to measure QoE more reliably and how the influence factors affect QoE are two objectives of this study.

In this study, the proposed ORD method was adopted in the experiments, and the results are called PoE (Preference of Experience). Based on the two designed experiments conducted in the two labs at different locations, the influence factors of typical broadcasting systems on PoE are evaluated. Some novel statistical analysis methods which fit the 3D subjective measurement community are used as well, which could overcome the restriction of the traditional statistical analysis methods that they are not applicable to small-size sample.

In this study, we found a series of possible factors that would influence PoE in 3DTV. Video content affected the PoE significantly. FCC format performs worst when comparing with SBS and Full HD format. In addition, the observers' preference on 2D or 3D video were different when using different 3D display technologies. The possible influence factors, e.g., gender, 3D experience of the observers and screen size are investigated. When using shutter glasses, the preference of 2D degraded video sequences is higher than the 2D high quality video sequences.

This study provides some hints on the design of experiments in 3DTV. For example, when evaluating the performance of video codecs, the video content and

display technologies should be taken into account. In addition, the selection of the participants should be balanced in gender and 3D experience distribution.

II

Visual discomfort in 3DTV: subjective, objective prediction and modeling

7

Visual discomfort in 3DTV: State of the art

As one of the most important dimensions in QoE of 3DTV, visual discomfort is often complained by the viewers. Thus, it is quite necessary to investigate the possible causes of visual discomfort, and then, develop an objective prediction method to automatically monitor, adjust or optimize the related systems and thus, to minimize the possibility of visual discomfort. Before going into the details of the most challenging work in visual discomfort as mentioned above, this chapter provides the readers with a basic knowledge and the state-of-the-art research work on it.

7.1 Definitions

Visual discomfort and visual fatigue are two distinct concepts though they are often confused and interchangeably used in some papers. Visual fatigue and discomfort are notions that encompass medical, subjective and psychological aspects. Related studies similarly lies in between associated research fields. For this reason, terminologies and definitions may vary from one study to the other, and are rarely documented.

Before introducing the definitions of visual fatigue and discomfort, some terminations should be defined firstly.

– Symptom: A symptom is a subjective sensation reported by the patient, as an evidence of his perceived physical or mental condition. Symptom is subjectively, it cannot be measured directly [29].

- Clinical sign: A clinical sign is observed or measured by the medical examiner, thus is an objective evidence of a patient's condition [74].
- Syndrome: A syndrome is a set of (subjective) symptoms and (objective) signs that occur together, which is characteristic of a physical or mental condition [98].

7.1.1 Visual fatigue

Depending on the context, fatigue is either considered as a symptom of a medical condition, or a medical condition itself. In our context, the latter terminology applies.

Visual fatigue is caused by the repetition of excessive visual efforts, which can be accumulated, and disappears after an appropriate period of rest. Visual fatigue can be assessed by the presence of:

- zero, one or more symptoms reported by the patient, which may include the sensation of fatigue reported by the patient;
- zero, one or more clinical signs observed by the medical examiner or measured through experimental protocols.

Their nature, intensity, and temporal properties (time of appearance, duration, raise and fall time) may be used to assess the severity of visual fatigue.

7.1.2 Visual discomfort

Visual discomfort is a physical and/or a psychological state assessed by the patients themselves, as a presently perceived degree of annoyance. As such, it may be related to experienced symptoms, perceived difficulties when performing a visual task, or any negative sensation associated with this task. Visual discomfort appears and disappears with any of these negative associations, and is supposed to have a short raise and fall time contrary to visual fatigue. In other words, visual discomfort disappears rapidly when the visual task is interrupted, either by asking the observer to close his eyes or by terminating the visual stimulus. Thus, visual discomfort can be measured by asking the viewer to report its level.

In this thesis, we focus on the visual discomfort issues.

7.2 Main causes of visual discomfort

7.2.1 Vergence-Accommodation conflict

Vergence-Accommodation conflict is a well-known factor that would induce visual discomfort [50][72]. When viewing an object by means of a 3D screen, the

eyes will converge to the virtual object which is in front of or behind the screen plane. However, the accommodation has to be performed at the screen depth level, which is unnatural and will not happen in our daily life. The larger this discrepancy between the vergence and accommodation gets, the higher the possibility that observers will perceive visual discomfort.

To define the threshold of this discrepancy in which conditions viewers may not experience visual discomfort, i.e., the comfortable viewing zone, numerous studies have been conducted. Yano et al.[150] proposed that the depth of field (DOF), which refers to the range of distances in image space within which an image appears in sharp focus, can be used to define the comfortable viewing zone in terms of diopters (D). A value of $\pm$ 0.2 D is suggested [18][149]. Another definition on comfortable viewing zone is based on the results of empirical measurements, in which $\pm$1 arc degree of visual angle is used [76][118]. If considering the screen disparity, the comfortable viewing zone can be defined by a percentage of the horizontal screen size. For 3D television, values of $\pm$3% are suggested[121]. For cinema, 1% for crossed and 2% for uncrossed disparities are suggested [95].

The comparison of these different comfortable viewing zone is shown in Figure 7.1. Generally, these definitions generate similar comfort area [122].

7.2.2 Disparity distribution

In addition to the Vergence-Accommodation conflict, some studies also showed that the disparity distribution might introduce visual discomfort as well:

1. Excessive uncrossed disparity (behind screen) will induce less visual discomfort compared to the crossed disparity (in front of the screen) when the angular disparity magnitude is the same [54].
2. When most parts of an image are positioned behind the screen (or the averaged disparity is uncrossed), there will be less visual discomfort compared with the condition that they are distributed in front of the screen [99].
3. If the image is split into top and bottom parts, the stereoscopic image will be more comfortable to watch when the top part of the image is distributed behind the screen and the bottom of the image is in front of the screen [100].
4. In the condition of the same averaged value of disparity distribution, higher dispersion of the disparity would lead to more visual discomfort due to the Vergence Accommodation conflict [99].

7.2.3 Binocular distortions

Binocular distortions or binocular image asymmetries seriously reduce visual comfort if present to a sufficient extent [75]. Asymmetries can be classified into

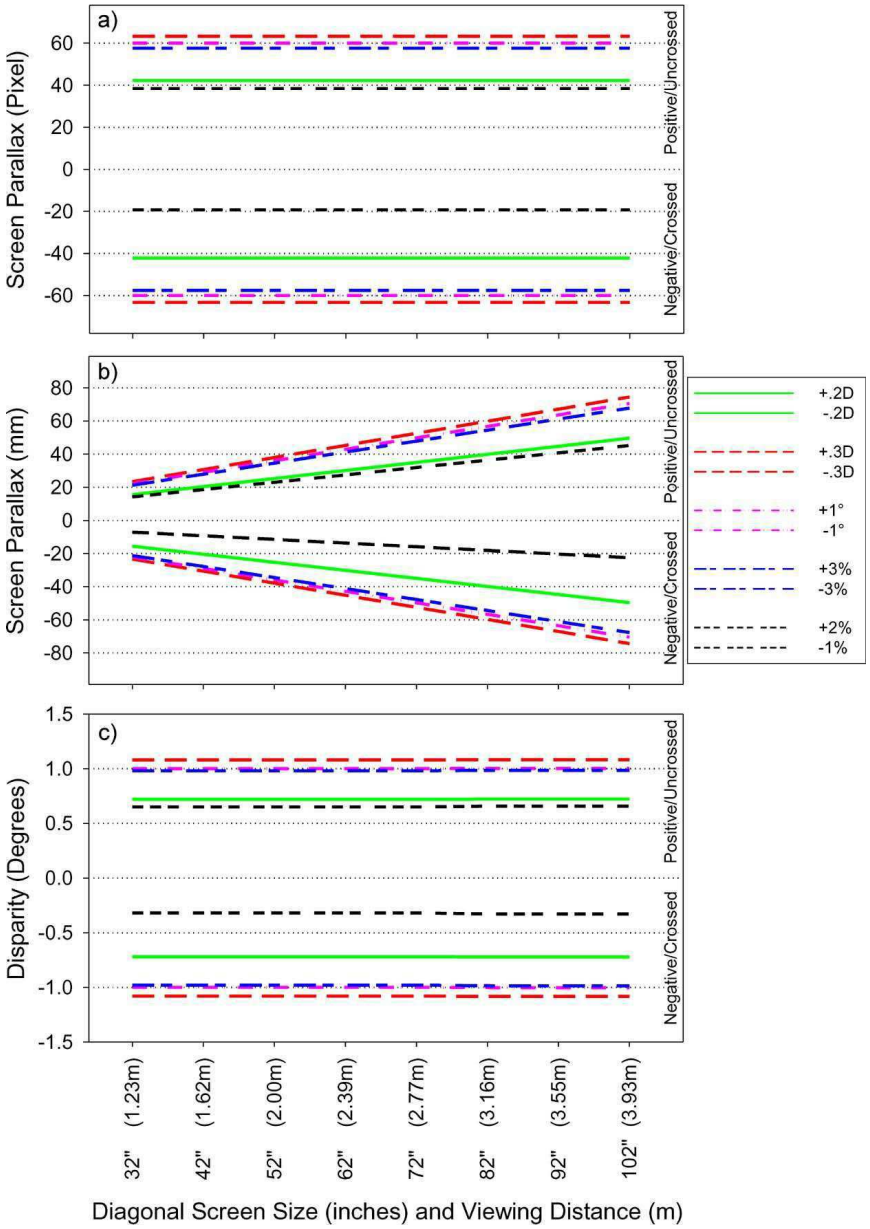


Figure 7.1: Comparison on different definitions on comfortable viewing zone. The viewing distance is 3 times of the screen height [122].

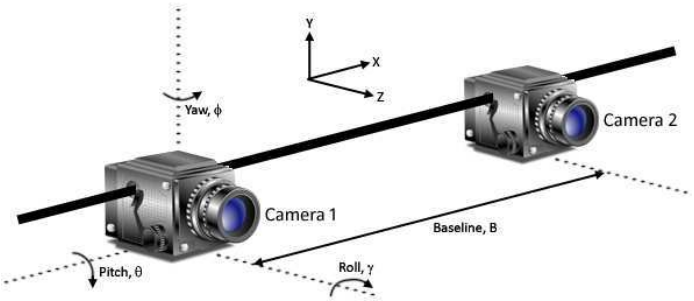


Figure 7.2: Depiction of the three axes of camera misalignment errors (the pitch, roll and yaw axis)[110].

optics related errors, filters related errors and display related errors. Optics errors are mainly geometry differences between the left and right images, e.g., size inconsistency, vertical shift, rotation error, magnification or reduced resolution. These errors usually occur when shooting or displaying stereoscopic images/videos. Filter related errors are mainly photometry differences between the two views, e.g., color, sharpness, contrast. The main error induced by display systems is crosstalk. Crosstalk produces double contours and is a potential cause of discomfort [103]. A study showed that vertical disparity, crosstalk, and blur are most dominant factors when compared with other binocular factors in visual comfort [75].

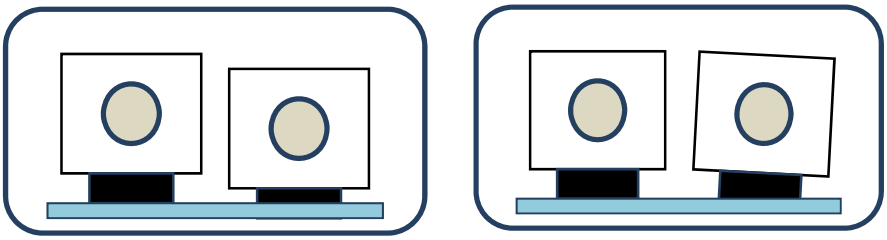
Optics related errors

The accuracy of the alignment between the two cameras during shooting determines the perceptual visual comfort to a large extent. Generally, the camera misalignments are divided into [134]:

- Horizontal misalignments;
- Vertical misalignments;
- Torsional misalignments;
- Size and keystone disparity fields.

The examples of the camera misalignment and the induced results are shown in Figure 7.2. The pitch and yaw axis horizontally and vertically through the picture plane, while the roll axis coinciding with the optic axis through the center of the lens. Vertical inconsistency of images caused by inconsistency of optic axes and errors in the rotational alignment between the two cameras are known to cause fatigue of eyes. Examples of these two errors are shown in Figure 7.3.

Studies in [134] indicated that the relationship between visual discomfort, vertical disparity (in unit of arcmin) and torsional disparity (degree) are that, in the condition of watching 3D stimuli for 2 seconds, vertical disparity about 60 arcmin will induce severe visual discomfort (the five-level rating scale is: 0- no discomfort, 1- mild, 2- moderate, 3- strong, 4- severe discomfort). For torsional disparity, 50



(a) Vertical adjustment error (b) Rotation adjustment error

Figure 7.3: Examples of the adjustment errors on different axis [57].

Table 7.1: Studies on the detection and tolerance limits of the geometric discrepancy on left and right views [148].

Geometric discrepancy	Detection limit	Tolerance limit	Remark
Size	1.2%	2.9%	Taking the size of one image as 100%
Vertical displacement	0.7%	1.5%	Taking the image height as 100%
Rotation	0.5 degree	1.1 degree	Angle of rotation about the image center

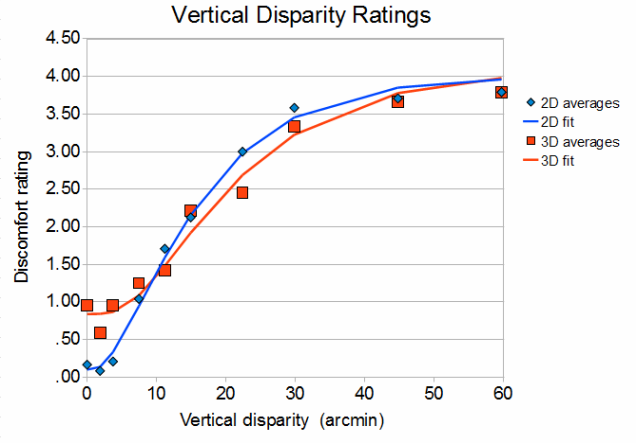
degree will induce strong discomfort. The results are shown in Figure 7.4.

In the converged camera configuration there is often a distortion called keystone distortion. It is the phenomenon that in one of the views, the image of the grid appears larger at one side than the other as shown in Figure 7.5[71]. Studies already showed that this distortion will induce visual discomfort or even visual fatigue [145][55].

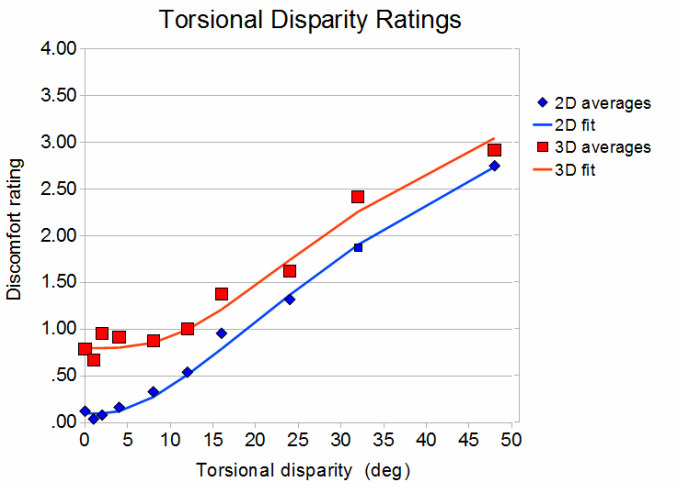
The tolerance of these optics related errors on visual discomfort have been investigated in [148] and the results are shown in Table 7.1.

Filters related errors

In [61], the Just Noticeable Difference (JND) thresholds of the visual comfort in 3DTV were investigated psychovisually for seven types of between-eye image differences, including luminance, gamma, contrast, color temperature, chroma, hue and random tone differences. The experimental results showed that: (1) the visual comfort threshold values are higher when increasing the luminance and color temperature differences between two-views, (2) decreasing the binocular differences on contrast or hue to zero will result in low threshold values on visual discomfort, which indicates this type of differences easily inducing binocular rivalry, and (3) luminance adaptation and chromaticity adaptation plays an important role on the variations of visual comfort thresholds.



(a)



(b)

Figure 7.4: (a) Discomfort ratings for the relative vertical displacement of the images to the two eyes, generating a vertical disparity. (b) Discomfort ratings for the relative rotation of the images to the two eyes around the optic axis, generating a torsional disparity. Data are the mean values of 9 subjects [134].

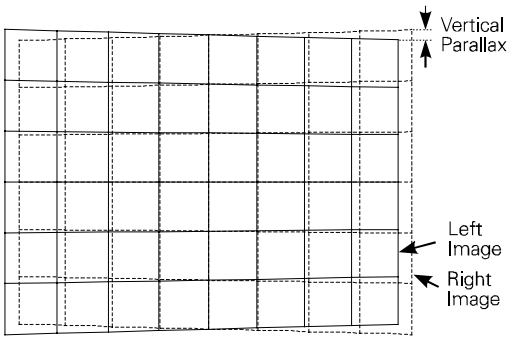


Figure 7.5: Keystone distortion.

Crosstalk

Crosstalk, or image ghosting, in stereoscopic displays refers to imperfect image separations where the left-eye view leaks to the right-eye view and vice versa. Crosstalk can be classified into system crosstalk and viewer crosstalk. System crosstalk is only induced by the display technology, it is independent of the quality of stereoscopic image pairs while the viewer crosstalk is dependent on the video contents.

Studies showed that visibility of crosstalk increases with increasing contrast and increasing binocular disparity (depth) of the stereoscopic image [112]. Even a small amount of crosstalk would induce visual discomfort or visual fatigue [103]. Studies on the thresholds of crosstalk level for the acceptance of the viewing experience and visual discomfort have been conducted. In [17], they found that "crosstalk between 2 and 6% significantly affected image quality and visual comfort". In [75], they found "crosstalk level of about 5% is sufficient to induce visual discomfort in half of the population". In [44], they reported that the crosstalk tolerance limit is 5% - 10%, and visual detection limit is 1% - 2%. In [141], it is shown with natural still images that the S-3D display technology with the lowest luminance and contrast level tolerates the highest level of crosstalk, while still maintaining an acceptable image-quality level.

7.2.4 Motion

Motion in 3DTV can be classified into planar motion (or lateral motion) and in-depth motion. Planar motion means that the object only moves in a certain depth plane perpendicular to the observer, and the disparity does not change temporally. In-depth motion, which is also called motion in depth or z-motion, is defined as object movement towards or away from an observer [47]. For planar motion, both eyes make the same conjunctive eye movements, called $version$ [21]. For in-depth motion, the eyes make opposite, disjunctive eye movement, called $vergence$ [48]. The eye movements for the planar motion and in-depth motion are shown in Figure 7.6. The speed of the planar motion and in-depth motion can be expressed by the change of distance per second or the change of the visual angle (version or vergence) per second.

Fast motion can induce visual discomfort even if the object is within the comfortable viewing zone [149][118]. Studies showed a consistent conclusion on the influence of motion velocity on visual discomfort, i.e., visual discomfort increases with the in-depth motion velocity [121][149][118][21], and for planar motion video sequences, visual discomfort increases with the planar motion velocity [121][84].

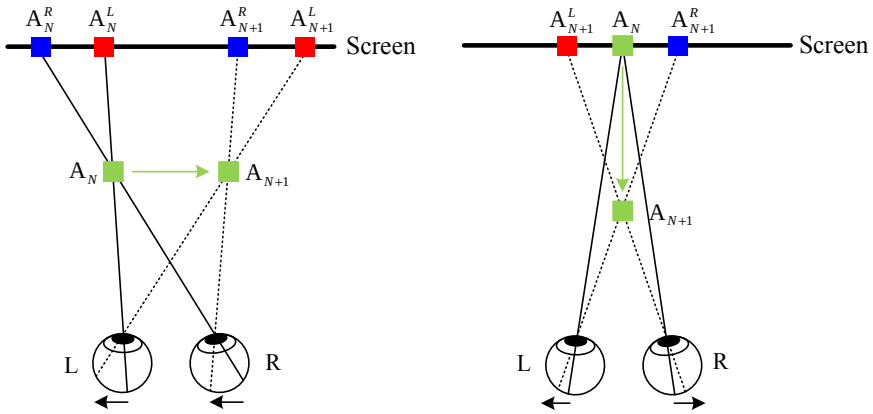


Figure 7.6: The left figure shows the eye movement of the planar motion object. Planar motion velocity is the amount of the change of the version per second. The right figure shows the eye movement of the in-depth motion object. In-depth motion velocity is the amount of the change of the vergence per second. A_N represents the perceived virtual object at frame N, A_N^L and A_N^R represent the left and right view images on the screen at frame N.

7.3 Subjective assessment methodology

In ITU-R BT.2021 [56], four assessment methods are recommended for measuring visual discomfort, which are a subset of the methods from Recommendation ITU-R BT.500 [58]. 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 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”, as shown in Figure 7.7.

These methods were already widely used in the community of S-3DTV. For example, in [150], the SSCQE method was used as it can measure the influence of stimulus duration on visual discomfort or visual fatigue. In [118][116][121] and [115], the five-grade SS method was used.

Questionnaires are also popular for the measurement of visual discomfort. The simulator sickness questionnaire (SSQ) is a well known and well established questionnaire which is used to evaluate motion sickness caused by motion images [68]. The test items in SSQ include “General discomfort”, “Fatigue”, “Headache”, “Eye strain”, “Difficulty focusing”, “Blurred vision”, etc. The participants are asked to

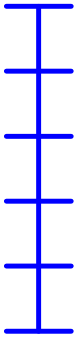
Discrete scale	Continuous scale
5 Very comfortable	 Very comfortable
4 Comfortable	Comfortable
3 Mildly comfortable	Mildly comfortable
2 Uncomfortable	Uncomfortable
1 Extremely uncomfortable	Extremely uncomfortable

Figure 7.7: Example of scale labels for subjective assessment of visual discomfort.

fill the questionnaire item by item and select the answers from “None”, “Slight”, “Moderate” and “Severe”. Ohno and Ukai [102] developed their own questionnaire based on SSQ and List of Symptoms of Visual Fatigue from [120]. In [76], the authors developed a new questionnaire which is used to subjectively assess visual fatigue caused by viewing various types of motion images.

7.4 Objective psychophysical prediction

Besides the subjective assessment methods, visual discomfort or visual fatigue induced by 3DTV may be predicted or measured by objective psychophysical devices. For example, Electroencephalography (EEG) and Functional Magnetic Resonance Imaging (fMRI) have been used to assess the brain activities which are related to the processing and reactions to the stimuli, e.g., emotion, visual fatigue. Electromyography (EMG) and Electrooculography (EOG) are used to detect the activities related to the eyes, e.g., electrical activity produced by skeletal muscles of the eyes and eye movement. Furthermore, eye blinking rate might be changed in different viewing conditions, which could be used to predict visual discomfort or visual fatigue.

7.4.1 Electroencephalography (EEG)

In clinics, EEG is often used to detect disorders of brain activity or to monitor certain procedures, e.g., the depth of anesthesia. Recently, it has been adopted in psychophysical studies on the relationship between brain activity and 3D QoE. It is shown that stereoscopic 3D videos would elicit responses from certain brain regions which relate to stereo perception processing, visual discomfort/fatigue, emotion, etc.

Brain waves can be classified into four basic groups according to the frequency

band: gamma (25-100Hz) [52], beta (12-30Hz) [107], alpha (8-12Hz), theta (4-7Hz) and delta (0.1-4Hz) bands [125][73]. Different brain activities are reflected on different bands. For example, alpha activity is induced by closing the eyes or by relaxation, and abolished by thinking or calculating. The gamma band is related to high cognitive processes. These selected responses of brain regions allow for discovering the relationship between the test stimuli and a certain attribute of the QoE. For example, in the study of [73], the authors used an EEG device to compare the brain activity between watching 2D and 3D video sequences. The results showed that the power of the EEG signals in beta frequency was significantly higher when watching 3D contents, which might be related to either visual discomfort or visual fatigue.

7.4.2 Functional Magnetic Resonance Imaging (fMRI)

fMRI is an MRI procedure that measures brain activity by detecting associated changes in blood flow. Compared to EEG, fMRI is more precise in understanding the human brain regions related with the stereoscopic perception due to its high spatial resolution.

A large amount of efforts have been dedicated to measuring human cortical activity when viewing stereoscopic stimuli [70]. It was discovered that while watching stereoscopic images, the stereoscopic shape recognition and the corresponding processing were probably performed in certain regions [3][130][40]. For example, in [70], the authors used fMRI to test visual fatigue when watching stereoscopic images. The results indicated that V3A (a cortical visual area, for more details please refer to [129]) is related to stereoscopic perception as the activation at V3A is much stronger when watching stereoscopic images rather than 2D images. In addition, the results showed that there were strong activities in the frontal eye field (FEF) when watching 3D images with large disparities. This conclusion is consistent with some previous EEG studies which showed that the areas near the prefrontal cortex (PFC) were related with 3D visual fatigue [34][86].

7.4.3 Electromyography (EMG)

Usually, EMG is used to analyze the neuromuscular activation of muscles within postural tasks, functional movements, work conditions and treatment or training regimes. EMG often measures not only at the extremes of the muscle but also along the muscle. As visual fatigue is defined as a decrease of the performance of the human vision system produced, it may be possible to use EMG to detect muscle activities around the eyes and to find a relationship between the muscle activities and the visual fatigue/discomfort.

Nahar et.al [97] studied the EMG response of the orbicularis oculi muscle to "low-level visual stress" conditions, where "low-level" means the work conditions in which muscles are activated at a level that can be maintained for a long period of time. The results showed that for conditions that visual fatigue may stem from eyelid squinting (e.g., refractive error, glare), the power of the EMG response increased with the degree of eyestrain.

7.4.4 Eye Blinking

Eye blinking rate is a possible indicator for predicting visual discomfort or visual fatigue. Studies showed that when in relaxed conditions, people would blink more often than in book reading and computer reading tasks [131]. In [82][151], the results showed that blinking rate was higher when watching 3D video than 2D videos. The study of [69] gives the conclusion that eye blinking rate increases with visual fatigue when watching 3D images. For the conditions of watching screens, the blinking frequency was significantly decreased when fatigue was reported (e.g., reading information from the screen for a long time) [30].

7.5 Conclusions

Visual discomfort is a very important dimension in QoE of 3DTV. In this chapter, we introduced the state-of-the-art studies on visual discomfort, including the definitions on visual discomfort and visual fatigue which is often interchanged in literature; the possible factors that may induce visual discomfort; the subjective assessment methodologies, and the psychophysical measurement methods on visual discomfort.

In the following chapters, the challenging issues in visual discomfort will be studied, i.e., how the 3D motion components affect visual discomfort; how individual differences affect the test results; what is the influence of the test methodologies on measuring visual discomfort; the development of an objective visual discomfort model for 3D videos; and a more accurate psychophysical prediction method for visual discomfort.

8

Subjective assessment on visual discomfort in 3D videos: influence of 3D motion

It is well accepted that large disparity and large amount of motion are two main causes of visual discomfort. To quantify this influence, three objectives are set in this chapter. The first one is the comparative analysis on the influence of different types of motion, i.e., static stereoscopic image, planar motion and in-depth motion, on visual discomfort. The second one is the investigation on the influence factors for each motion type, for example, the disparity offset, the disparity amplitude and velocity. The third one is to propose an objective model for visual discomfort. In addition, the influence from viewers' 3D experience, i.e., the differences between experts and naive viewers are studied.

8.1 Introduction

As we already introduced in Section 7.2, there are many possible factors that would induce visual discomfort. Most of the causes have been well studied for the case of still stereoscopic images. For stereoscopic 3D videos, as the only difference between stereoscopic image and video is the motion, the influence of motion on visual discomfort has been widely investigated recently.

In-depth motion is one of the significant factors that may cause visual discomfort. Studies already showed that visual discomfort increases with the in-depth motion velocity [121][149][118][84][21]. However, the influence from disparity am-

plitude (disparity range) and the disparity type (crossed or uncrossed) of in-depth motion on visual discomfort are still under study. In [118], the results showed that disparity amplitude of the moving object is not a main factor. However, in their recent study [121] it is shown that visual discomfort increases with the disparity amplitude. Furthermore, the results also showed that the in-depth motion with crossed disparity would induce significantly more visual discomfort than the uncrossed and mixed conditions. In [84], as they only analyzed the in-depth motion in the disparity range of ± 1 degree with different velocities, there is no conclusion about the influence of crossed or uncrossed disparity amplitude on visual discomfort.

The influence of the planar motion on visual discomfort was studied as well [121][88][84][87]. These studies showed high consistency on the conclusion that visual discomfort increases with the motion velocity. However, the influences of the disparity on visual discomfort led to different conclusions in these studies. In [84], the results indicated that the disparity type, i.e., crossed and uncrossed disparity, did not affect the visual discomfort thresholds. However, in [121], the results showed that the crossed disparity will generate more visual discomfort than the uncrossed disparity. A possible explanation for this inconsistency might be the position of the background. In [84], the background was positioned at the screen plane. In [121], the position of the background was not depicted but in their previous study [118], the background was positioned at a fixed place with the disparity of -2.6 degree. The impact of the position of background on visual discomfort may therefore require further study.

Most of the studies mentioned above investigated the influence of the in-depth motion and planar motion on visual discomfort individually. For quantifying the influence of static situations, planar and in-depth motion, it would be important to directly compare their impact on visual discomfort. Thus, this chapter is focusing on the influence of motion on visual discomfort of 3DTV, including the comparative analysis on the influence of different motion types on visual discomfort, the influence of disparity and velocity within a certain motion type and the proposal of an objective visual discomfort model based on the results. In addition, the influence of viewers' experience (experts and naive observers) on 3DTV is analyzed.

8.2 Experiment

Based on the two main objectives, i.e., analysis on the influence from motion and the study on the influence from human factors, two experiments were designed as shown in Table 8.1.

Table 8.1: Summary of the two experiments.

Item	Exp 1	Exp 2	
		Exp2-a	Exp2-b
Study target	Influence from motion	Human factors	
Num. of stimuli	36 (static+planar+in-depth)	15 (planar)	
Method	ASD	FPC	
Display technology	Shutter glasses		
Resolution	Full HD, 1920 × 1080		
Viewing distance	3H, 90cm		
Num. of observer	42	10	45
Observer type	Naive observer	Experts	Naive observer
Gender	21 males + 21 females	8 males + 2 females	21 males + 24 females
Age (mean)	19-48 (26.8)	24-43 (27)	18-44 (24)
Trials/obs	180 pairs	210 pairs	105 pairs
Votes/Stimulus Pair	42	15	45

8.2.1 Definitions

To analyze the influence of crossed and uncrossed disparity, and the disparity magnitude of the moving object on visual discomfort, in this study, we use *disparity amplitude* and *disparity offset* to define the motion in stereoscopic videos. The disparity amplitude d_a between the nearest point A and farthest point B can be expressed by Equation (8.1) which represents the range of the disparity of the moving object. ϕ_A represents the disparity of point A, and ϕ_B represents the disparity of point B. The disparity offset d_o between the two points A and B can be expressed by Equation (8.2) which represents the center of the angular disparity between the two points.

The static and planar motion stimuli can be characterized by the disparity offset and the planar velocity, where the disparity amplitude equals zero. The in-depth motion stimuli can be defined by the disparity amplitude, the disparity offset and the in-depth velocity.

$$d_a = |\phi_A - \phi_B| \tag{8.1}$$

$$d_o = \frac{1}{2}(\phi_A + \phi_B) \tag{8.2}$$

8.2.2 Experimental design

To avoid the complexity of the influence factors contained in 3D natural video sequences, the synthetic stimuli were decided to be used in this study allowing for precise control on the possible influence factors, including motion type, velocity, disparity offset and disparity amplitude. In Experiment 1, 36synthetic video stimuli were used, including 15 planar motion stimuli, 5 static stimuli and 16 in-depth motion stimuli.

For the planar motion stimuli, we selected five angular disparity offset levels

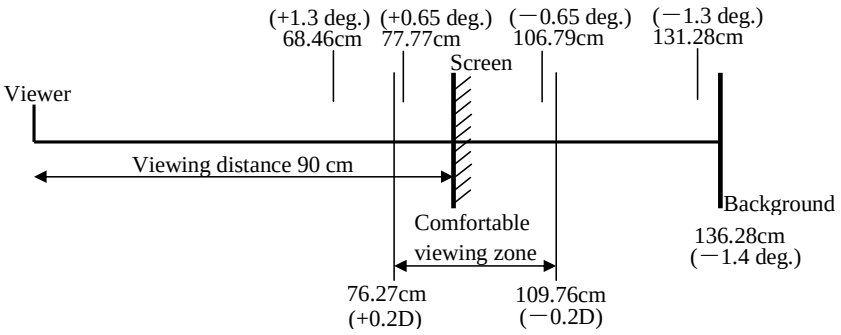


Figure 8.1: The relationship of the foreground and the background position and the comfortable viewing zone in planar motion stimuli.

(0, ± 0.65, and ± 1.3 degree) and three velocity levels (slow, medium, and fast with velocity of 6, 15 and 24 degree/s). A background is designed to be placed at a fixed position (-1.4 degree) which is consistent with a typical natural video content where the background is almost fixed and placed behind the screen. Figure 8.1 shows the disparities used in the planar motion stimuli and their relationship with the comfortable viewing zone.

For the static condition, five disparity offset levels were selected which were the same as the planar motion design. The foreground object stays fixed in the center of the screen at a certain disparity plane.

For the in-depth motion condition, four disparity amplitude levels (0.65, 1.3, 2 and 2.6 degree), three disparity offset levels (-0.65, 0, 0.65 degree) and three velocity levels (1, 2, and 3 degree/s, binocular angular degree) were selected. The reason for choosing binocular angular disparity speed was that the object's velocity appears visually constant which is not the case for a constant value in the unit of cm/s. The direction of the movement is inverted at the far or at the near end of the movement so the object in the experiment moved forth and back in an endless loop. The three velocity levels 1, 2 and 3 degree/s represent slow, medium and fast, respectively. Figure 8.2 shows the design of the disparity amplitude and disparity offset for the in-depth motion.

In experiment 2, only the 15 planar motion stimuli were used. Two types of viewers participated in the test, one being expert viewers who work in 3-D perception, coding, quality assessment and subjective experiments and know very well about the 3D depth perception, visual discomfort, etc. The others were naive viewers, who do not have experience in this domain.

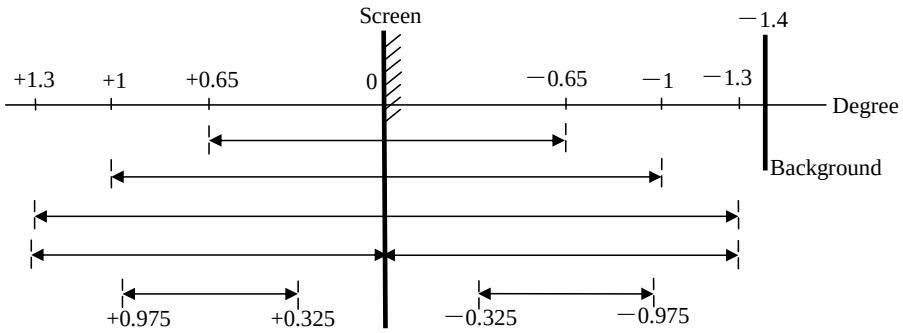


Figure 8.2: The disparity amplitude and offset design for in-depth motion stimuli. The arrows represent the depth interval in which the object moves.

8.2.3 Apparatus

The stereoscopic sequences were displayed on 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. The graphics card of the PC was an NVIDIA Quadro FX 3800. Stimuli were viewed binocularly through the NVIDIA active shutter glasses (NVIDIA 3D vision kit) at a distance of about 90 cm, which was approximately three times the picture height. The peripheral environment luminance was adjusted to about 44 cd/m^2. When seen through the eye-glasses, this value corresponded to about 7.5 cm/m^2 and thus to 15 % of the screen's peak brightness as specified by ITU-R BT.500 [58].

8.2.4 Stimuli

The stereoscopic sequences consisted of a left-view and a right-view image which were generated by the MATLAB psychtoolbox [12][105]. Each image contained a foreground object and a static background. A black Maltese cross which was frequently used in such kind of psychometric experiments [38][92] was used as the foreground object with a resolution of 440×440 corresponding to visual angle of 7.6 degree. As it contained both high and low spatial frequency components, it was supposed to limit the influence of one particular spatial frequency in the experiment [101].

The background was generated by adding salt and pepper noise to a black image of Full HD resolution, and then filtered by a circular averaging filter with radius of 5. The background is shown in Figure 8.3. The reason for using this kind of image as the common background of all stimuli was that it could preclude all of the monocular cues on stereopsis.

For the planar motion stimuli, the trajectory of the moving object is a circle with

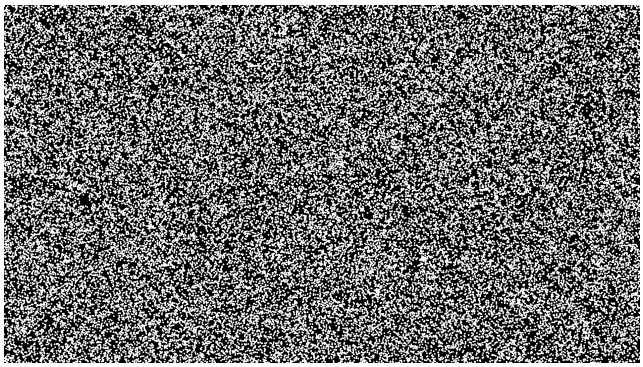


Figure 8.3: The background image of the synthetic stimuli.

center point at the center of the screen, and radius of 300 pixels, approximately 10 degree of visual angle. The motion direction of the object was anti-clockwise. An example of the stimuli is shown in Figure 8.4(a), in which the foreground object is placed in front of the screen with an angular disparity of 1.3 degree. For the static stimuli, the Maltese cross was positioned at the center of the screen. For the in-depth motion stimuli, the Maltese cross was positioned in the center of the screen and moved back and forth to the viewers. An example is shown in Figure 8.4(b), in which the foreground object is moving in the depth plane with disparity amplitude of 2.6 degree and offset of 0 degree.

The 15 planar motion stimuli, 5 static stimuli and 16 in-depth motion stimuli used in the subjective experiment are listed in Table 8.2 with their stimulus serial number, disparity offset d_o, disparity amplitude d_a, planar motion velocity v_p and in-depth motion velocity v_d.

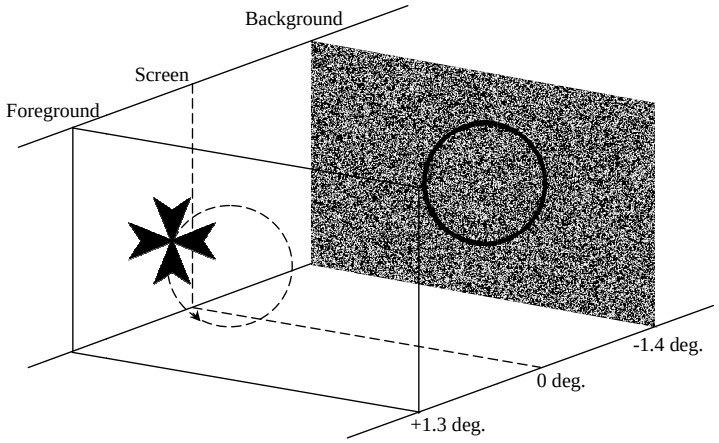
8.2.5 Viewers

42 viewers participated in Experiment 1. 21 are male, 21 are female. They are all non-experts in the domain of subjective experiments, image processing or 3D. Their age ranged from 19 to 48 with an average age of 26.8.

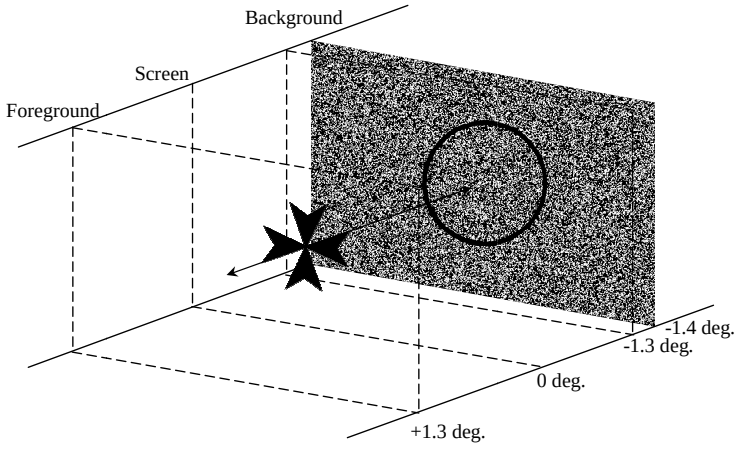
10 experts in 3-D perception, coding, quality assessment and subjective experiments participated in the Experiment 2-a. Eight experts are male, two are female. Their ages ranged from 24 to 43 with an average age of 27. As the number of viewers is too small, to generate a reliable result, 5 of them conducted the test twice but on a different day. Thus, for each pair, there are 15 observations in total.

45 naive viewers participated in the Experiment 2-b. Twenty-one are male, twenty-four are female. They are all non-expert in subjective experiment, image processing or 3D related field. Their ages ranged from 18 to 44 with an average age of 24.

All of the viewers 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.



(a)



(b)

Figure 8.4: (a) An example of stimulus with planar motion in the experiment. The foreground object is moving at the depth plane with a disparity of 1.3 degree. The background is placed at a fixed depth plane of -1.4 degree. The motion direction of the Maltese cross is anti-clockwise. (b) An example of stimulus with in-depth motion in the experiment. The disparity amplitude of the Maltese cross is 2.6 degree, offset is 0 degree. The foreground object is moving in depth between disparity +1.3 to -1.3 degree back and forth.

Table 8.2: All stimuli used in the experiment. d_o is disparity offset, d_a is disparity amplitude, v_p is planar motion velocity and v_d is in-depth motion velocity. The last two columns are BT score and confidence interval of the BT score (CI) which are discussed in Section 8.3.

Type	Number	d_o (deg.)	d_a (deg.)	v_p (deg./s)	v_d (deg./s)	BT score	CI
Planar motion	1	-1.3	0	6	0	0	0.42
	2	-1.3	0	15	0	1.70	0.33
	3	-1.3	0	24	0	3.56	0.27
	4	-0.65	0	6	0	1.03	0.37
	5	-0.65	0	15	0	2.09	0.31
	6	-0.65	0	24	0	3.40	0.27
	7	0	0	6	0	1.80	0.35
	8	0	0	15	0	2.44	0.31
	9	0	0	24	0	3.61	0.26
	10	0.65	0	6	0	2.36	0.31
	11	0.65	0	15	0	2.84	0.29
	12	0.65	0	24	0	3.85	0.27
	13	1.3	0	6	0	3.10	0.27
	14	1.3	0	15	0	3.58	0.27
	15	1.3	0	24	0	4.22	0.26
Static	16	-1.3	0	0	0	0.90	0.38
	17	-0.65	0	0	0	2.39	0.30
	18	0	0	0	0	3.89	0.26
	19	0.65	0	0	0	5.86	0.24
	20	1.3	0	0	0	7.39	0.25
In-depth motion	21	0	1.3	0	1	4.70	0.25
	22	0	1.3	0	2	5.35	0.24
	23	0	1.3	0	3	5.76	0.24
	24	0	2	0	1	4.82	0.25
	25	0	2	0	2	5.45	0.24
	26	0	2	0	3	6.14	0.24
	27	0	2.6	0	1	4.99	0.25
	28	0	2.6	0	2	5.95	0.24
	29	0	2.6	0	3	6.24	0.24
	30	-0.65	1.3	0	1	4.00	0.26
	31	-0.65	1.3	0	3	5.50	0.24
	32	-0.65	0.65	0	1	3.97	0.27
	33	-0.65	0.65	0	3	5.00	0.25
	34	0.65	1.3	0	1	5.34	0.24
	35	0.65	1.3	0	3	6.04	0.24
	36	0.65	0.65	0	3	5.99	0.24

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.

8.2.6 Assessment Method

The paired comparison method was used in this test. As there are 36 stimuli in Experiment 1, to reduce the number of comparisons, the ASD method was used which has been introduced in previous chapters (Chapter 4 and 5). In Experiment 2, as only 15 stimuli were considered, the FPC method was used.

In Experiment 1, according to ASD method, 36 stimuli lead to a total of 180 pairs for each viewer. In Experiment 2-a, only 10 experts conducted the experiment. In order to produce more reliable results, each viewer conducted the test using the FPC method and both presentation orders for each pair were considered, i.e., each stimulus pair will be shown twice but with different order, the stimulus which is shown firstly in one trial will be shown secondly in another trial. Thus, there were in total $2 \times \begin{pmatrix} 15 \\ 2 \end{pmatrix} = 210$ pairs for each viewer in Experiment 2-a. In Experiment 2-b, a total of 105 pairs were presented to each viewer. The presentation order of the stimulus pair was different for odd numbered and even numbered viewers. For example, viewers with even numbers will watch stimulus A first, then stimulus B. For odd numbered viewers, this order is inverted. This is used to balance the presentation order.

The presentation order of each stimulus pair in all experiments was randomly permuted for each viewer.

8.2.7 Procedures

The subjective experiments contained a training session and a test session. Five pairs of stimuli were shown in the training session. The viewers were asked not to stare at the moving object but to watch the whole stereoscopic sequence. Then, they should select the one which is more uncomfortable, concerning e.g., mental uneasiness. The viewers used two distinct keys to select one of the two stimuli in a pair for immediate visually check on the screen. There was a minimum duration for the display of each stimulus before making a decision by pressing a specific third button. The minimum duration is defined as either 5 seconds or the duration of a complete cycle of movement (the moving object went back to its start point) whichever was longer. During the training session, all questions of the viewers were answered. We ensured that after the training session, all of the viewers knew about the process and task of this experiment clearly.

In the main test session, 180 pairs were compared for Experiment 1, 210 pairs for Experiment 2-a, and 105 pairs for Experiment 2-b. As the duration of each test was different due to the number of pairs and individual differences of each viewer, and to avoid visual fatigue caused by long time watching affecting the experimental results, the Experiment 1 and Experiment 2-b were split into two sub-sessions. Each session contained half of the total number of stimulus pairs. There was a 10-min break between the two sub-sessions. For Experiment 2-a, the viewers were asked to have a 15 minutes break after each 30 minutes of the test.

8.3 Results of Experiment 1: Influence of motion

The Bradley-Terry (BT) model [10][11] is used to convert the pair comparison data to psychophysical scale values for all stimuli, more details about BT model can be found in Chapter 3. The BT scores and confidence intervals of all stimuli are shown in Table 8.2. For easier comparison, the lowest BT score is set to 0.

The static condition can be considered as either a special case of the planar motion or in-depth motion, both with the motion velocity of 0. Thus, in this section, the static condition is analyzed in both conditions.

8.3.1 Planar motion and static conditions

The BT scores for the planar motion stimuli and static stimuli are shown in Figure 8.5 where the static condition can be considered as a special case of the planar motion.

The experimental results on the planar motion stimuli showed that:

- Visual discomfort increases with the planar motion velocity;
- The vergence-accommodation (VA) conflict might not significantly affect the visual discomfort. As shown in Figure 8.5(a), the visual discomfort neither reaches the minimum at the screen plane nor increases with the absolute value of disparity offset.
- The relative disparity r_o between the foreground and the background ($r_o = d_o + 1.4$ in this study) determines the visual discomfort, i.e., visual discomfort increases with the relative disparity.

A possible explanation of the influence of VA conflict and relative disparity on visual discomfort in our study might be the existence of the background. During the test, the viewers switched their attention between the background and the foreground, the larger of this distance, the larger of the change of VA conflict, which may lead to more visual discomfort.

For the static stimuli as shown in Figure 8.5(a), the visual discomfort increases with the relative disparity as well. Under the condition of small relative disparity,

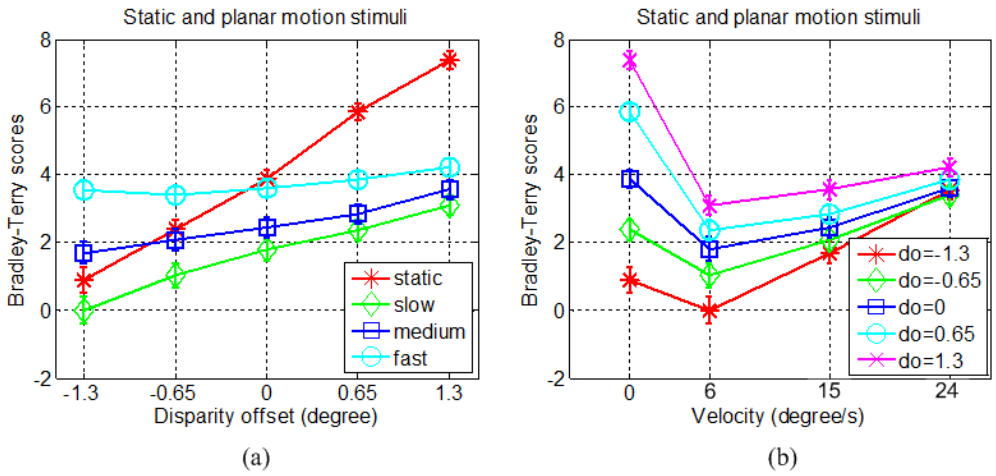


Figure 8.5: The Bradley-Terry scores of the static and planar motion stimuli. (a) Different lines represent different velocity levels, where static, slow, medium and fast represent 0, 6, 15 and 24 degree/s. (b) Different lines represent different disparity offset levels. Bradley-Terry scores represent the degree of visual discomfort. Error bars are 95% confidence intervals of the BT model fit.

i.e., $r_o = 0.1$ degree ($d_o = -1.3$ degree), the visual discomfort induced by static stimuli is less than the planar motion stimuli with medium or fast velocities. However, the gradient of the curve for the static case is steeper than the planar motion conditions. Thus, the visual discomfort increases faster with the disparity offset for the static stimuli than for the planar motion stimuli. By interpolating the static stimulus curve, when the static stimuli is very close to the background, i.e., with disparity offset close to -1.4 degree, the generated visual discomfort might be similar to the condition where the stimulus with similar disparity offset but slow planar motion. In the condition of disparity offset equals to zero degree ($r_o = 1.4$ degree), the static stimuli would generate similar visual discomfort as the fast planar motion stimuli. When the disparity offset is larger than 0 degree, the visualization of static objects seems to induce more visual discomfort than planar motion stimuli.

A different interpretation of this results is that when the relative disparity between the foreground and the background is increasing and the disparity offset becomes crossed, the planar motion seems to help to reduce visual discomfort when compared to the static condition. In our experiment, the viewers explained that when watching the planar motion stimuli, it was easier to fuse the Maltese cross compared to the static conditions, in particular when the disparity is crossed.

As shown in Figure 8.5(b), for the planar motion condition, there might be a minimum in the curve of visual discomfort that would be located at some velocity in between static and the slowest velocity that was included in this study.

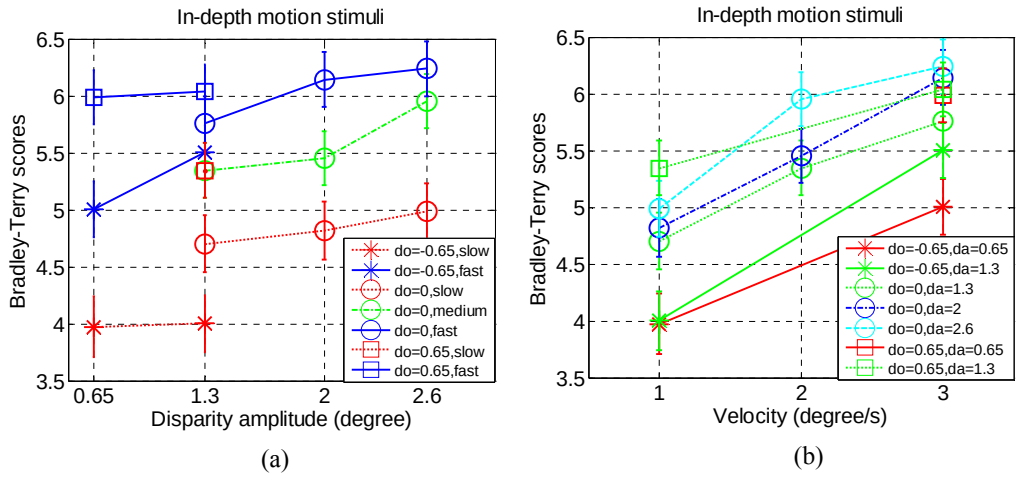


Figure 8.6: The Bradley-Terry scores of the in-depth motion stimuli. (a) The x-axis represents the disparity amplitudes, different lines represent different disparity offsets (d_o) and velocities. (b) The x-axis represents disparity velocities, different lines represent different disparity offsets (d_o) and disparity amplitudes (d_a).

8.3.2 In-depth motion and static conditions

The Bradley-Terry scores for in-depth motion stimuli are shown in Figure 8.6. According to the results, we may draw the following conclusions:

- As shown in Figure 8.6(a), in general, disparity amplitude may not affect the visual discomfort significantly. For example, in the condition of d_o = -0.65 degree and slow velocity, the visual discomfort induced by the stimulus with disparity amplitude of 0.65 degree is not significantly different from the stimulus with disparity amplitude of 1.3 degree. However, for the fast motion conditions, disparity amplitude may influence visual discomfort.
- Visual discomfort increases with the disparity offset as shown in Figure 8.6(a). This results is similar as the effect of relative disparity offset on planar motion stimuli, i.e., the relative disparity might be a main factor in this case.
- As shown in Figure 8.6(b), visual discomfort increased with the in-depth motion velocity.

The static condition can be considered as a special case of the in-depth motion as well. The BT scores of the static stimuli were compared with the in-depth motion stimuli which are shown in Figure 8.7.

As shown in Figure 8.7(a), the gradient of the curve for in-depth motion is much flatter than for the static conditions. When the disparity offset is less than 0.65 degree, the in-depth motion will generate more visual discomfort than the static stimuli. For example, when compared with the static stimuli with the disparity of 0 degree, all the in-depth motion stimuli in our study generated more visual discomfort. However, when the relative disparity is larger than 2.05 degree (d_o =

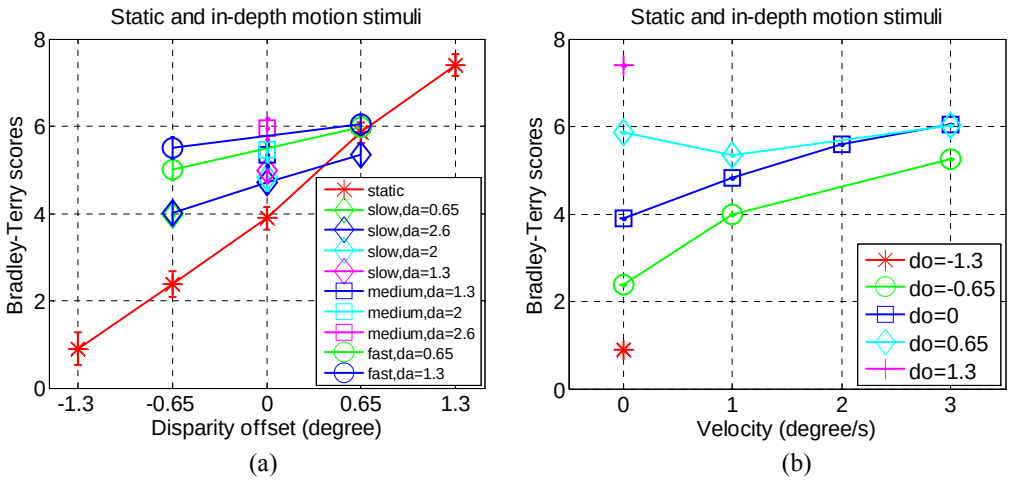


Figure 8.7: The Bradley-Terry scores of the static and the in-depth motion stimuli. (a) The x-axis represents the disparity offsets, different lines represent different disparity amplitudes (d_a) and velocities. (b) The x-axis represents velocity. Different lines represent different disparity offset (r_o). Note that the BT score is the mean scores of the stimuli with same offset and velocity but different disparity amplitudes.

0.65 degree), we may extrapolate that the visual discomfort induced by the static stimuli would be higher than the in-depth motion stimuli.

Considering the velocity, as shown in Figure 8.7(b), when the relative disparity is less than 2.05 degree (d_o = 0.65 degree), the visual discomfort increases with the velocity. However, if the relative disparity is larger than 2.05 degree, the static stimuli might generate more visual discomfort than the in-depth motion stimuli with slow velocity.

8.3.3 Discussion

In literature it is often mentioned that the motion in stereoscopic videos would induce more visual discomfort than static conditions. However, in this study, a counter-indication was found.

All three motion types showed that the relative disparity between the foreground and the background is a main factor in visual discomfort, i.e., visual discomfort increases with the relative disparity. The gradient of visual discomfort with relative disparity is highest for the static stimuli, followed by in-depth and then planar motion stimuli. This implies that static stimuli induce more visual discomfort when the relative disparity exceeds a certain value. This value is approximately 1.4 degree for planar motion and 2.05 degree for in-depth motion.

The gradient analysis also reveals that there is no “crossing point” between the planar motion and the in-depth motion in the positive three-quarters of the disparity

Table 8.3: The linear regression analysis results for all stimuli

Motion type	Factor analysis			Model analysis
	factor	coefficient	p-value(t-test)	D = Intercept + $\sum$ coefficient × factor
Planar motion	r_o	1.45	0.0000	Intercept = -1.11
	v_p	0.18	0.0000	R^2=0.98, RMSE=0.17
	$r_o \times v_p$	-0.04	0.0000	F=221.474, p-value=4.10×10^{-10}
Static stimuli	r_o	2.53	0.0000	Intercept = 0.54, R^2=0.99, RMSE=0.15 F=1176.37, p-value=5.45×10^{-5}
In-depth motion	r_o	1.23	0.0000	Intercept = 2.51
	v_d	0.31	0.0000	R^2=0.98, RMSE=0.11
	$d_a \times v_d$	0.45	0.0001	F=147.18, p-value=1.8×10^{-9}
	$r_o \times d_a \times v_d$	-0.21	0.0031	

space, i.e., d_o from -0.65 to 1.3 degree. The in-depth motion stimuli are always more uncomfortable than the planar motion stimuli in this study. However, for the condition that the disparity offset is less than 0.65 degree, we may extrapolate that the slow in-depth motion stimuli might generate less visual discomfort than the fast planar motion stimuli. However, further studies are required.

8.4 Linear regression analysis: towards an objective visual discomfort model

To investigate the influence factors of each motion type, multiple linear regression analysis is used in this study which attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data.

For the static situation, there is only one possible factor which is the relative angular disparity. For motion stimuli, the relative disparity offset, disparity amplitude, planar motion velocity, in-depth motion velocity, and their interactions are possible factors. The stepwise regression function in Matlab was used to select the most significant factors or remove the least significant factors [31]. The output of the stepwise regression includes the estimates of the coefficients for all potential factors, with confidence intervals, the statistics for each factor and for the entire model. To avoid over fitting of the model, the Leave-one-out Cross Validation (LOOCV) method was used to all possible models to find the model with the minimum averaged RMSE. The selected models are shown in Table 8.3.

All the factors shown in Table 8.3 are statistical significant factors with p-value of the student's-t-test < 0.05. The *coefficient* in the table is the coefficient of the corresponding factor in the linear model. The *model analysis* shows the linear model for each motion type. The R^2, RMSE, the F-statistic and its p-value are provided

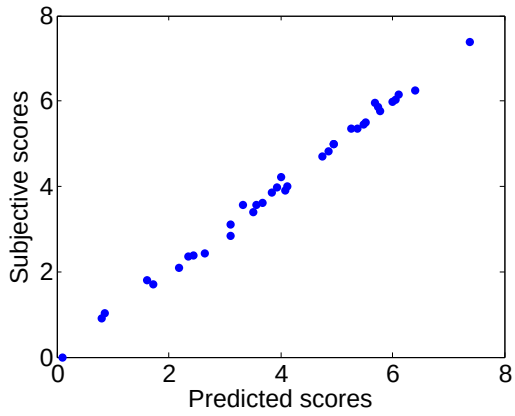


Figure 8.8: The scatter plot of the predicted scores and the BT scores.

as the evaluation results of this model. D represents visual discomfort score. It is shown that for the planar motion stimuli, the relative disparity, planar motion velocity and their interaction term are important factors for visual discomfort. For the static stimuli, the relative disparity offset in this study shows its predominant effect. For the in-depth motion stimuli, the disparity amplitude is not a main factor which is consistent with the conclusions of Section 8.3.2 and [118]. However, the interaction term of the disparity amplitude and the velocity, and the combination of the three factors (velocity, disparity amplitude, relative disparity offset) plays an important role in determining visual discomfort.

According to the regression analysis results, an objective model for comparing visual discomfort of still stereoscopic images, planar motion stimuli and in-depth motion stimuli is developed. All disparity and velocity values are measured in visual angular degree. Here we rewrite it as:

$$D = \begin{cases} 2.53r_o + 0.54 & \text{static condition} \\ 1.45r_o + 0.18v_p - 0.04r_o v_p - 1.11 & \text{planar motion} \\ 0.31r_o + 1.23v_d + (0.45 - 0.21r_o)d_a v_d + 2.51 & \text{in-depth motion} \end{cases} \tag{8.3}$$

The scatter plot of the objective and subjective results is shown in Figure 8.8. The Pearson Linear Correlation Coefficient (PLCC), Spearman's Rank Correlation Coefficient (ROCC) and Root Mean Square Error (RMSE) are used to evaluate the correlation between the objective scores and the subjective results, they are 0.9976, 0.9967, and 0.1198, respectively.

As this model is based on the paired comparison results, the D can be used to compare the degree of visual discomfort between the stimuli. The difference can be interpreted as the probability that one condition is preferred to another.

This model works for the condition of a single moving object. However, for natural content conditions, how to combine the visual discomfort induced by sev-

eral different moving objects and how to integrate the visual discomfort scores of different scenes will be presented in Chapter 10.

8.5 Results of Experiment 2: Influence of human factors on visual discomfort

8.5.1 Comparison between experts and naive viewers

The BT scores of the Experiment 2 from experts and non-experts data are shown in Figure 8.9. Both the experts and non-experts BT scores for the 15 planar motion stimuli provide the same conclusion as found in Section 8.3.1. The consistency of the experts and naive viewers' test results are: CC = 0.9688, ROCC = 0.9357, RMSE = 0.2737.

The Barnard's exact test is applied on the raw pair comparison data of the experts and naive viewers results, and there are in total 21 pairs significantly different (p<0.05), which corresponds to 20% of the whole pairs. Thus, in general, the two experimental results are well correlated.

8.5.2 Classification of observers

When considering the influence from relative disparity and velocity, people may have different sensitivity on them. Thus, it may be interesting to classify them into different groups and analyze the different influences of relative disparity and velocity on different observers.

The relative disparity and the planar velocity are two factors that may induce visual discomfort in our study. Thus, the analysis which factor is dominant in determining the visual discomfort is conducted on each observer. There are two hypotheses in this analysis:

- $Hypothesis$ 1: the relative disparity is predominant;
- $Hypothesis$ 2: the velocity is predominant.

The methods to measure which factor is more predominant are based on the $p1$ and $p2$ values, where:

- $p1$: the proportion of each observer voting for the stimulus whose relative disparity is larger;
- $p2$: the proportion of voting for the stimulus whose velocity is faster than the other one.

Each observer's opinion on these two hypotheses can be reflected by $(p1, p2)$ which can be expressed by a point in a two-dimensional space. According to these points, the observers can be classified as different groups. In our study, the K-means clustering method was used. For better illustration, we define the term G-H1(Group

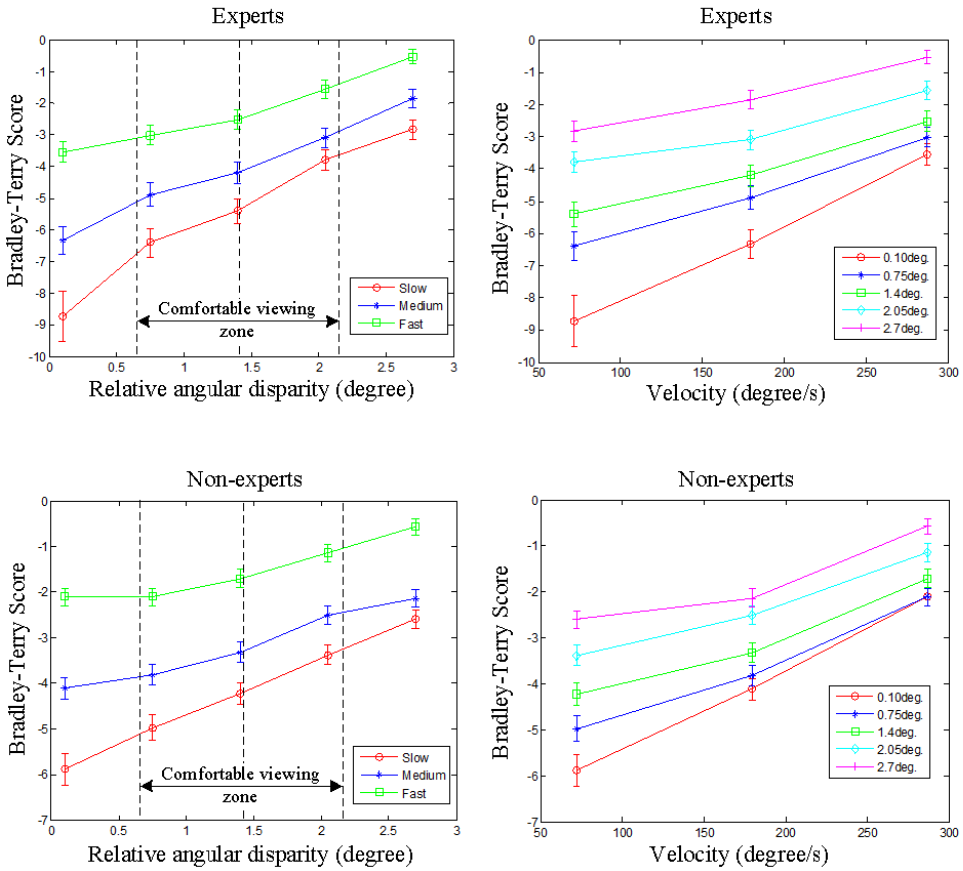


Figure 8.9: BT scores for visual discomfort. The top two figures are experts results. The bottom two figures are non-experts results. The different lines in the left figures represent the different velocity levels. The vertical two dashed lines represent the upper and lower limits of the comfortable viewing zone, which are at 0.66 and 2.14 degree. The dashed line in the middle represents the position of screen plane. The different lines in the right figures represent the different relative angular disparity levels. The error bars are the 95% confidence intervals of the BT model fit.

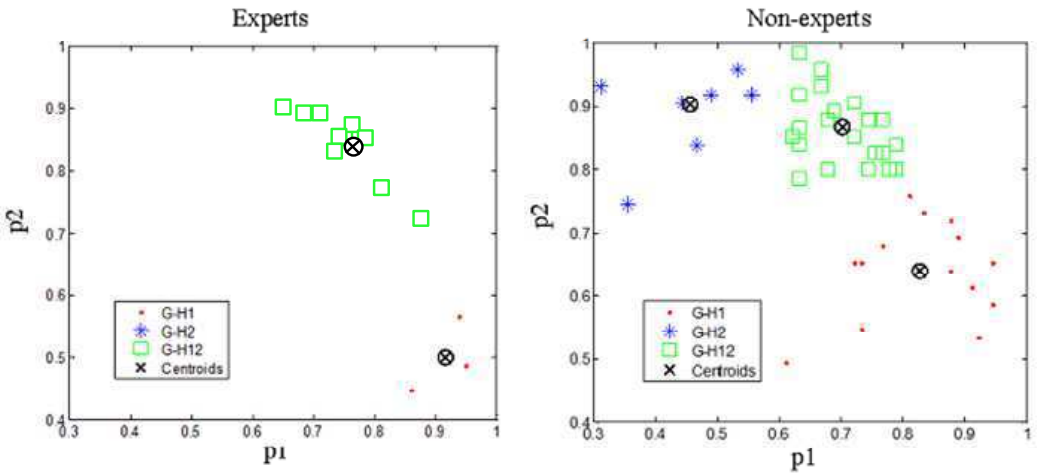


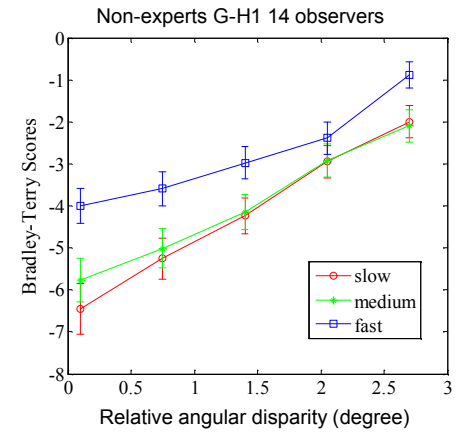
Figure 8.10: The clustering results for experts and non-experts observers. X-axis represents the agreement on "relative disparity is the predominant factor" and y-axis represents the agreement on "velocity is the predominant factor".

of Hypothesis 1) to represent the observer group who voted more according to $Hypothesis$ 1, which means relative disparity is predominant in determining visual discomfort. A similar definition is used for G-H2. G-H12 is for the group who are equally sensitive to relative disparity and velocity, like the global subjective results. The clustering results are shown in Figure 8.10. The BT scores for all stimulus generated by each observer cluster are shown in Figure 8.11.

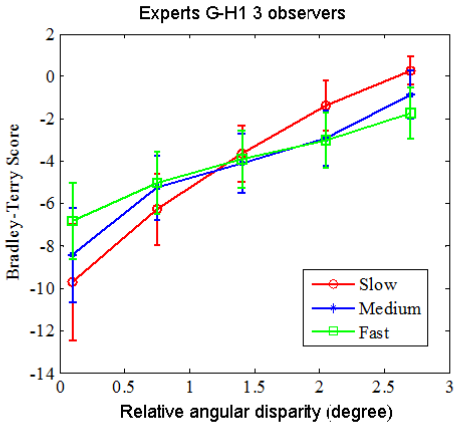
According to Figure 8.11 it is showed that most of the viewers agree with the global subjective experiment results, i.e., visual discomfort increase with the motion velocity and relative disparity. There are small number of viewers who have totally different sensitivity on relative disparity and motion velocity. It should be noted that for the results of Experts group G-H1, as the number of viewers is too small and for each pair, there are only 3 observations, thus, this results might not be reliable.

8.6 Conclusions

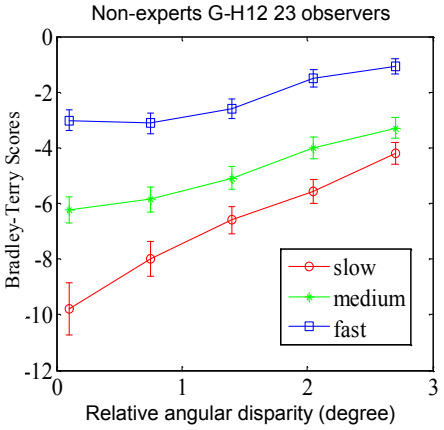
In this chapter, we used paired comparison method to evaluate visual discomfort induced by motion of stereoscopic videos. Different motion types, i.e., static condition, planar motion and in-depth motion conditions are compared. The results showed that motion does not always induce more visual discomfort when compared with static stereoscopic images. In particular, in the condition that the moving object is far from the background, static objects would induce more visual discomfort than the moving conditions. Generally, in-depth motion stimuli would generate more visual discomfort than the planar motion stimuli, which might be opposite in



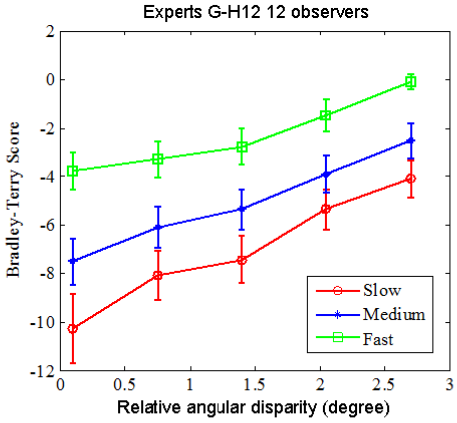
(a)



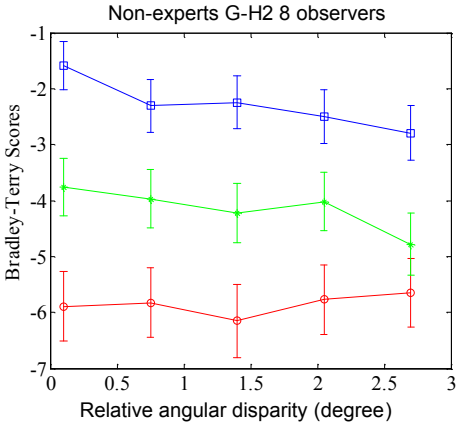
(b)



(c)



(d)



(e)

Figure 8.11: BT scores of the different classes of the viewers. Naive viewers' results are in the first column. Experts' results are in the second column. The rows represent group G-H1, G-H12, and G-H2 respectively.

the condition of very small disparity offsets.

The velocity and the relative disparity are main factors for visual discomfort in stereoscopic videos. The disparity amplitude did not affect visual discomfort significantly which is consistent with the conclusion of [118]. However, the interactions between disparity amplitude and in-depth motion velocity, and the three-way interaction of the relative disparity, disparity amplitude and in-depth motion velocity showed significant effect on visual discomfort.

According to the regression analysis, an objective model which can be used to compare the visual discomfort of different types of motion was proposed. In Chapter 10, this model will be extended and optimized by natural content video sequences.

In addition, we classified the observers as different clusters according to which factor is predominant in determining their feeling of visual discomfort. The clustering results showed that most of the observers agreed with the global subjective experiment results. However, there were small number of observers who considered either the relative disparity or velocity as the predominant factor in inducing visual discomfort while the other factor has small influence on their perceptions.

9

Comparison of test methodologies on assessing visual discomfort in 3DTV

An objective visual discomfort model is proposed in Chapter 8. To evaluate this model on natural 3D video sequences, a ground truth of visual discomfort scores for this database is needed. In this chapter, two subjective experiments were conducted. One used the paired comparison method and the other used the ACR method. Thus, besides providing the ground truth data, a comparison study of the test methodology on the experimental results is conducted. This chapter shows a very important but usually ignored conclusion on the subjective assessment methodology, i.e., the test method would affect the experimental results significantly.

9.1 Introduction

For the study of visual discomfort induced by stereoscopic images or videos, most of the subjective experiments were conducted by different test methodologies, e.g., in [149], five scale based ACR methodology was used, where the score from 1 to 5 represents "I'm very tired" to "I am not tired"; and in [118], a continuous scale from 0 to 100 was used, where "0" represents "Extremely Uncomfortable" and "100" represents "Very Comfortable"; the test methodology in [65] is a 5-point ACR test, where the attributes were also selected from "very comfortable" to "extremely uncomfortable". However, there are few studies on the comparison of the visual discomfort results obtained by different test methodologies with the same test conditions. Due to the multi-dimensionality of the QoE, and the difficulties for the viewers to make judgement on unfamiliar and multi-dimension scales, it would be

interesting to know the influence of the test methodology on test 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, and the other was conducted by the ORD (*Optimized Rectangular Design*) paired comparison method. The remainder of this chapter is arranged as follows: In Section 9.2, the test stimuli are introduced, which are natural 3D video sequences. In Section 9.3, the two experiments are introduced. In this study, only one experiment was conducted in our IVC lab, the other experimental setup and results are available from the test database. In Section 9.4, the experimental results obtained from ACR and paired comparison are compared and analyzed by different statistical tools. Finally, Section 9.5 concludes this study.

9.2 Stimuli

In this study, the IVY Lab stereoscopic video database [60] 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 were estimated by an 8×8-pixel block matching method [60] and the depth estimation reference software (DERS from MPEG 3D video standardization) [124], 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 that only 36 stimuli were chosen are 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 paired comparison test, 36 stimuli is feasible for using the Square design method and it already reaches the maximum limit for test duration, approximately 1 hour (180 pairs = $180\times$(10+5)s = 45 minutes without break). 40 stimuli would make the test even longer (for $8\times$ 5 condition, the total number of pairs = 220 pairs = $220\times$(10+5)s = 55 minutes without break), which is not recommended.

The preview of the video sequences used in the subjective test are shown in

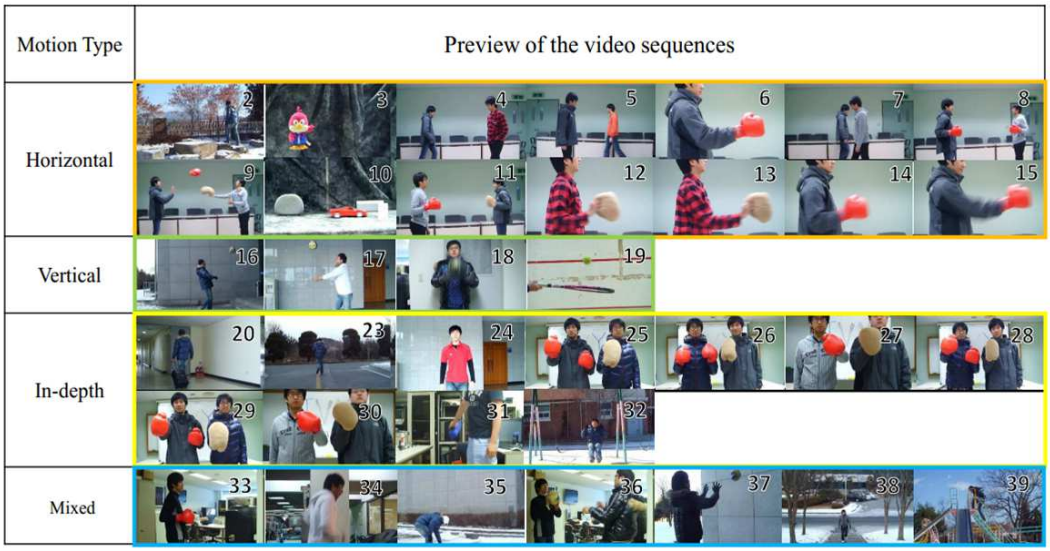


Figure 9.1: Preview of the test video sequences. They are captured from frame 100.

Figure 9.1. 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 missing in Figure 9.1 are the MPEG 3D video test sequences which were not chosen in this study. The corresponding disparity and motion information are listed in Table 9.1.

9.3 Experiment

9.3.1 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 on the IVY database. 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 $\times$ 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 $\times$ 720) was re-scaled to fit the full screen. The test environment was in line with the recommendations of ITU-R BT.500 [58].

Table 9.1: Information of the IVY stereoscopic video database [60]. $^{+}$ Maximum disparity extracted from salient regions. $^{++}$ For mixed motion type, the motion velocity (h, v, d) represents the horizontal, vertical and in-depth motion, respectively.

Content index	Motion type	The number of moving objects	Maximum disparity (degree)$^{+}$	Motion velocity (degree/s)$^{++}$	MOS	CI (95%)
2	Horizontal	1	0,66	3,41	4,06	0,31
3	Horizontal	1	0,17	3,45	3,85	0,37
4	Horizontal	2	0,24	7,29	3,71	0,30
5	Horizontal	2	0,30	7,32	3,50	0,25
6	Horizontal	1	0,95	7,49	4,03	0,33
7	Horizontal	2	0,34	7,80	3,56	0,41
8	Horizontal	2	0,14	9,50	4,06	0,30
9	Horizontal	2	0,48	10,36	3,74	0,38
10	Horizontal	1	0,17	10,71	3,15	0,29
11	Horizontal	2	0,19	12,72	3,65	0,35
12	Horizontal	1	0,89	12,91	3,59	0,26
13	Horizontal	1	0,89	16,40	3,15	0,42
14	Horizontal	1	0,95	17,51	3,29	0,45
15	Horizontal	1	0,95	25,55	3,06	0,42
16	Vertical	2	0,38	3,79	4,03	0,35
17	Vertical	2	0,53	6,44	3,79	0,29
18	Vertical	1	0,32	12,04	3,88	0,31
19	Vertical	1	1,00	30,82	2,65	0,46
20	In-depth	1	0,05	0,05	4,53	0,21
23	In-depth	1	0,60	0,19	4,09	0,34
24	In-depth	1	0,97	0,23	4,06	0,35
25	In-depth	2	0,79	1,16	3,38	0,31
26	In-depth	1	0,74	1,68	3,41	0,53
27	In-depth	2	0,99	1,74	3,50	0,48
28	In-depth	3	0,53	2,02	3,09	0,48
29	In-depth	3	0,73	2,53	3,35	0,43
30	In-depth	2	1,00	2,78	3,38	0,42
31	In-depth	1	0,97	3,37	3,09	0,49
32	Mixed	1	0,34	(0.60, **8.24**, 0.40)	3,38	0,42
33	Mixed	1	0,05	(**11.02**, 5.88, 0.79)	3,56	0,33
34	Mixed	1	0,05	(5.87, **14.2**5, 0.92)	3,68	0,37
35	Mixed	2	0,48	(2.96, **3.08**, 0.73)	3,71	0,41
36	Mixed	1	0,75	(**7.61**, 5.01, 0.67)	3,79	0,34
37	Mixed	1	0,16	(**9.61**, 4.67, 1.02)	3,79	0,37
38	Mixed	1	0,39	(0.51, 0.84, **0.67**)	4,09	0,36
39	Mixed	1	0,80	(1.05, **1.25**, 0.20)	4,35	0,26

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. The results are shown in Table 9.1.

9.3.2 Experiment 2: PC test conducted at the IVC lab

To compare the experimental results between the ACR and PC methods, a PC test was conducted with the experimental setup as close to Experiment 1 as possible. To reduce the number of comparisons, our proposed OSD method was used. 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 is 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 [58]. This setup is consistent with the experiment conducted in IVY lab besides the size of the screen. The test environment is shown in Figure 9.2.

Figure 9.2: Test environment.

Viewers

40 naive viewers participated in this test. 22 are females and 18 are males. Their ages are ranged from 19 to 65, with an average age of 30.2. All of them 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.

Test methodology

As there are in total 36 video sequences, and the MOS from Experiment 1 is available as shown in Table 9.1, the OSD (*Optimized Square Design*) method is used. Thus, the square matrix in OSD is arranged based on the ranking ordering of the MOS. The video sequences with the closest visual discomfort MOS will be 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 pair comparison binary data.

The square matrix of the OSD method is designed as follows according to the rank ordering of the MOS results in IVY, the number in the matrix represents the index of the stimulus:

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

Conforming to the OSD method, only stimuli in the same column or row are

Figure 9.3: Test interface.

compared. Thus, there are in total 180 pairs for each observer.

Procedure

The test includes a training session and a test session. Five pairs are included in the training session. After watching a pair of video sequences, the observers are asked to select the one which is more uncomfortable. A touch screen is used for the viewers to make the selection. If the observer is not very sure about the selection, he can replay the video sequences as many times as he wants. The test interface is shown in Figure 9.3.

There are 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 are as balanced as possible, which means 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-B} should appear as often as the condition {B-A} for all observers. In this way, the presentation bias effect is avoided as much as possible as we already introduced in Section 4.5.

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

9.4 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 by two main aspects.

One is focusing on the scale values after converting from the raw paired comparison data. The other is focusing on the raw paired comparison data.

The MOS and confidence intervals for all sequences from Experiment 1 are shown in Table 9.1. The Bradley-Terry model is used to generate visual discomfort scores of Experiment 2. As the question for the viewers in the experiment was "which one is more uncomfortable", the higher the BT score, the higher the degree of visual discomfort. It should be noted that in [65], the MOS represents the degree of visual comfort. The higher the MOS, the lower the degree of visual discomfort, which is opposite to the condition of BT scores.

It should be noted that as the BT scores are the converted scores from the paired comparison raw data, the confidence interval does not have the same meaning as the MOS confidence interval. It is related to the total number of comparisons and the goodness of model fit. In our study, as the chi-square test for goodness of model fit is passed, the BT scores can be used as the scale value from paired comparison data. The confidence interval can be used to explain the reliability of the estimated value.

9.4.1 Comparison between the scales values: MOS and BT scores

The scatter plot of the MOS and BT scores is shown in Figure 9.4. Both the confidence intervals of the MOS and BT scores are added in the figure. After mapping the BT scores to MOS based on the logistic fitting function recommended by VQEG's final report II [138], the CC, SROCC, RMSE between the MOS and BT scores are calculated, which are 0.53, -0.50 and 0.33, respectively.

Furthermore, it is shown that the confidence intervals of the MOS are larger than the BT scores. For better visualization, the sorted MOS and BT scores are shown in Figure 9.5. For MOS, they are ranged from 2.5 to 4.5. According to the confidence intervals, it is shown that a large amount of the scores is 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.

9.4.2 Comparison of the raw data

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

There are in total 35 adjacent pairs in MOS, for example, sequence{19,15},

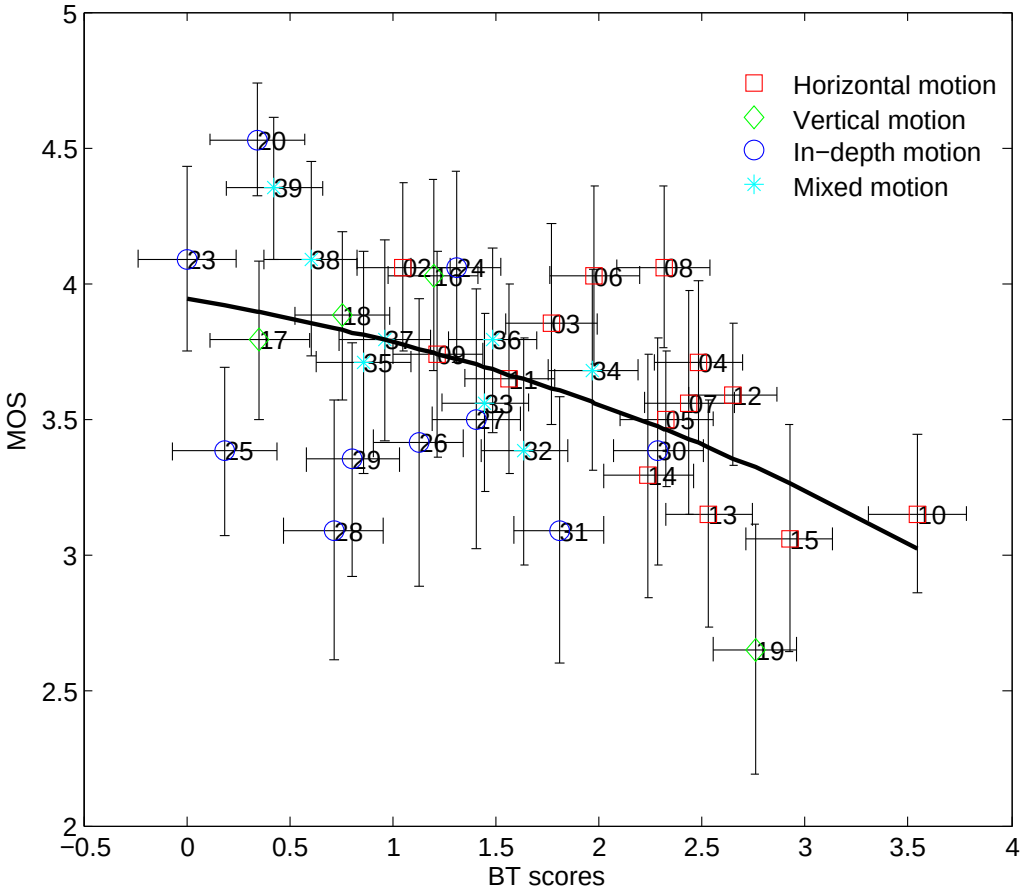


Figure 9.4: The scatter plot of the MOS results and BT scores. The error bars represent the confidence intervals of the MOS and BT scores. The labeled numbers next to the error bars are the sequence indexes. Different markers represents different types of motion in the database according to [60]. The black line is the fitting curve from BT scores to MOS.

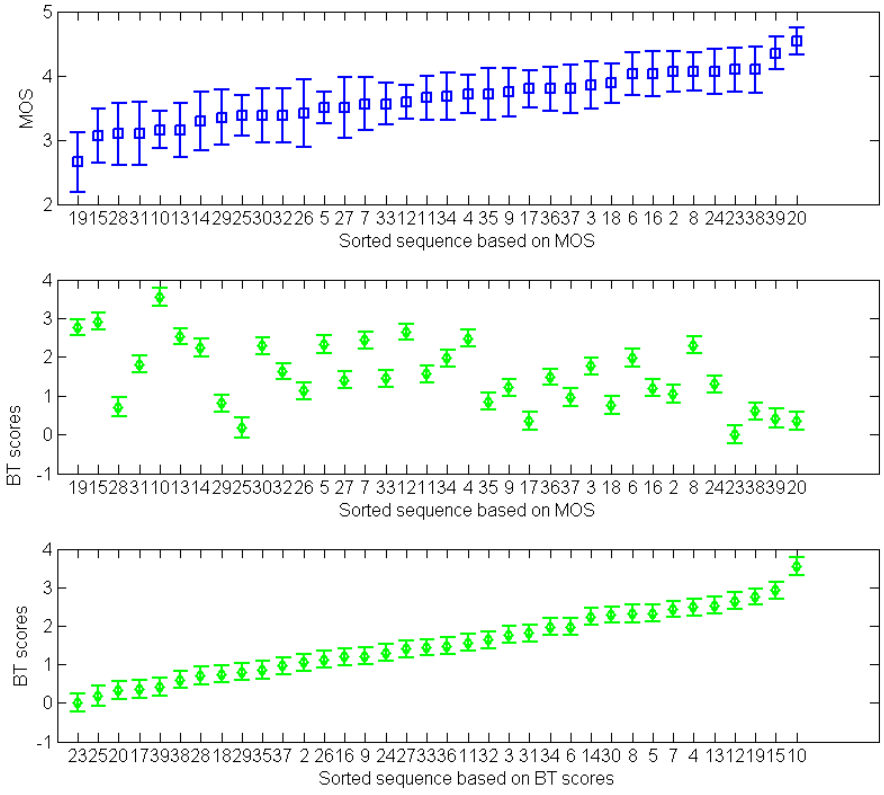


Figure 9.5: The comparison between the sorted MOS and BT scores.

Table 9.2: Barnard's test results on the adjacent pairs of the MOS. * indicates that the preference on the pair is significant at significance level of 0.05.

Sequence-A	Sequence-B	Vote on A	Vote on B	Barnard's p-value
19	15	16	24	0.16
15	28	34	6	0.00*
28	31	7	33	0.00*
31	10	9	31	0.00*
10	13	31	9	0.00*
13	14	24	16	0.16
14	29	33	7	0.00*
29	25	33	7	0.00*
25	30	5	35	0.00*
30	32	26	14	0.08
32	26	25	15	0.09
26	5	11	29	0.02*
5	27	29	11	0.02*
27	7	12	28	0.03*
7	33	26	14	0.08
33	12	8	32	0.00*
12	11	27	13	0.06
11	34	19	21	0.6
34	4	16	24	0.16
4	35	32	8	0.00*
35	9	17	23	0.26
9	17	32	8	0.00*
17	36	11	29	0.02*
36	37	24	16	0.16
37	3	11	29	0.02*
3	18	26	14	0.08
18	6	8	32	0.00*
6	16	28	12	0.03*
16	2	23	17	0.25
2	8	9	31	0.00*
8	24	29	11	0.02*
24	23	33	7	0.00*
23	38	16	24	0.16
38	39	20	20	1.00
39	20	23	17	0.26

Table 9.3: 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

sequence{15, 28}, ..., sequence{39, 20}. Based on the confidence intervals of these adjacent pairs, the MOS of the stimuli in each pair are not significantly different. The votings of the 40 viewers in our test and the corresponding Barnard's test on the preference are shown in Table 9.2. p-value < 0.05 means there is significant difference between the votings on the video sequence A and B at the significance level of 0.05. According to Table 9.2, 20 out of 35 pairs are significantly different.

If check those significantly different pairs, it would be found 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), the object is perceived as being cut off by the borders. This unnatural shooting distortion would induce visual discomfort [20].

In this 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 9.2, it might be inferred that when using the paired comparison 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 condition that one had window violation while the other did not, such as Stimuli {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 work as in paired comparison test because according to the confidence intervals of the MOS, the visual discomfort induced by these pairs are not 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, "PC1_ACR0" is used to represent the number of pairs that paired comparison 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 9.3. The results indicated that there are in total 27 pairs can be discriminated by the ACR method and 78 pairs can be discriminated by paired comparison 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 Paired comparison method has higher discriminability than the ACR method on the visual discomfort induced by different video sequences.

This study verifies the conclusions from [83] that the paired comparison method has higher discriminability on closer stimuli. In addition, the results showed that the viewer's behavior during the test might be affected by the test methodology. For example, in our paired comparison 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, such as the displays used in the test or the different cultures of the two labs (One is in Korea, the other is in France). The discussions above are only based on the results of this test. To validate the conclusions, more experiments are needed.

9.5 Discussions and Conclusion

To what extent the paired comparison methodology is different from the ACR method is the question to be resolved in this study. In this study, the visual discomfort results obtained by the ACR and PC test methodologies are compared. The results verified the conclusions that the paired comparison method has higher discriminability on the stimuli which have closer or similar test targets, e.g., quality. It is also found that the viewer's behavior during the test might be affected 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.

10

Objective visual discomfort model for stereoscopic 3D videos

In Chapter 8, the influence of 3D motion on visual discomfort in S-3DTV has been investigated. A corresponding visual discomfort model has been also proposed. In this chapter, the proposed model is evaluated on natural video sequences and verified by the subjective experimental results in Chapter 9.

10.1 Introduction

The industry would largely benefit from the availability of an objective visual discomfort model as it could be used to optimize the stereoscopic 3D images/videos production or broadcasting chain by predicting the visual discomfort. As we already discussed before, visual discomfort is the result of a combination of different factors. A summary of the features usually used to predict visual discomfort is presented in Table 10.1. The corresponding existing models are briefly introduced in Table 10.2.

Generally, objective visual discomfort models can be classified into two types, one considers only the characteristics of the stereoscopic images/videos, e.g., the disparity distribution, the object's geometry features, 3D motion, etc. The other also considers the binocular distortions, e.g., crosstalk, keystone distortion, etc.

For stereoscopic videos, one important issue for the modeling of visual discomfort is the influence of 3D motion. Some of the earlier studies on visual discomfort [150][100][79] didn't consider the differences between the planar motion and the in-depth motion, they used the combined motion as a feature to predict visual dis-

Table 10.1: Summary of the features that used in objective models

	Classification	Feature	Feature index
Statistic features	Disparity distribution	Disparity magnitude	11
		Disparity skewness	12
		Disparity distribution of the top and the bottom part of the image	13
		Disparity dispersion	14
		Disparity gradient	15
	Binocular distortion	Crosstalk	21
		Vertical disparity (keystone)	22
		Brightness	23
		Sharpness	24
	Image brightness	Image brightness	31
	Object geometry	Width	41
		Thickness	42
	Motion	Planar motion	51
		In-depth motion	52
		Camera motion	53

comfort. However, a recent study [65] shows that the influence of the planar motion and the in-depth motion on visual discomfort are significantly different. In [65], the proposed visual discomfort model is a function of the most salient object's disparity, planar horizontal motion velocity, planar vertical motion velocity and in-depth motion velocity. In our previous study, as shown in Chapter 8, we also found that the influences of planar motion and in-depth motion on visual discomfort are significantly different. Moreover, the static condition was taken into account in our study. It was shown that in the condition of large relative disparity between the foreground and the background, the static condition would induce more visual discomfort than the motion conditions.

Another important issue is the study of the combination effects of different factors on visual discomfort, for example, the disparity, and binocular distortions. In [22], the authors combined the features of stereoscopic images and motion components in stereoscopic videos to predict visual discomfort induced by S3D videos. The motion components in [22] include the objects' 3D motion, camera's movement and scene change. However, the planar motion and the in-depth motion were not considered individually. Furthermore, the interaction between the object's motion and disparity has not been taken into account, which actually is a significant influence factor on visual discomfort.

In Chapter 8, we already analyzed the visual discomfort induced by disparity

Table 10.2: Summary of the objective visual discomfort models for stereoscopic images and videos.

Model type	Model	Features	Remark
For stereoscopic image	Yano2002[150]	14	No image distortion
	Nojiri2006[100]	13,14	No image distortion
	Kim2011[71]	11, 14, 22	Range and maximum angular disparity, range and maximum vertical disparity, keystone distortion, location and spatial frequency
	Sohn2012[116]	11, 42	No image distortion
	Lee2013[85]	11, 41	No image distortion
For stereoscopic video	Yano2002[150]	13, 51+52	motion used in the model is a combination of planar motion and in-depth motion
	Nojiri2006[100]	11, 51+52	the same as above
	Choi2010[23]	11, 14, 52, 53	Depth complexity, depth position, temporal complexity, scene movement(camera motion or scene change)
	Lambooij2011[79]	11, 14, 15, 51+52	Average amount of motion, average amount of screen disparity, screen disparity range and gradient
	Jung2012[65]	11, 51, 52	Salient object's disparity, horizontal, vertical and depth motion
	Choi2012[22]	11, 14, 15, 21, 23, 24, 31, 51+52, 53	Spatial(depth) complexity, depth position, temporal complexity, scene movement, depth gradient, crosstalk, brightness, binocular differences in brightness and focus

and 3D motion on synthetic stimuli, and proposed a model which belongs to the category of without considering binocular distortion. In this chapter, the proposed visual discomfort model is evaluated by natural 3D video sequences. The remainder of this chapter is organized as follows. Section 10.2 presents the framework of the proposed visual discomfort model for which two implementations are proposed: one is based on the tracked moving objects in the video sequence; the other is based on the moving objects in each frame. Section 10.3 introduces the feature extraction methods in the models. Section 10.4 provides the 3D motion pooling strategy and the spatial and temporal pooling strategies. In Section 10.5, the performances of the proposed models are evaluated by the results of the two subjective experiments in Chapter 9, and a more comprehensive understanding of the two test methodologies can be obtained. In Section 10.6, an objective visual discomfort algorithm is introduced where an object tracking algorithm is integrated in the model and the performance is shown. Finally, Section 10.7 concludes this chapter.

10.2 Overview of the proposed model

In our previous study (Chapter 8), the relationship between the visual discomfort, disparity, and motion velocity is found as follows (the unit for disparity and velocity are all in visual angular degree):

$$D_{static} = 2.53r_o + 0.54 \tag{10.1}$$

$$D_{planar} = 1.45r_o + 0.18v_p - 0.04r_o v_p - 1.11 \tag{10.2}$$

$$D_{depth} = 0.31r_o + 1.23v_d + (0.45 - 0.21r_o)d_a v_d + 2.51 \tag{10.3}$$

where D_{static}, D_{planar} and D_{depth} represent the visual discomfort score induced by static, planar and in-depth motion, respectively. r_o is the relative disparity offset (disparity difference between the object and background), d_a is the disparity amplitude, v_p is planar motion velocity and v_d is the in-depth motion velocity.

For the static condition, there was only one attribute that has been studied, i.e., r_o (see Equation 10.1). This is because in our previous study, the disparity of the background was fixed. The disparity of the foreground is a variable. However, in natural video sequences, as the position of the background varies, there would be two variables, i.e., the disparity of the static object, as well as the relative disparity r_o. Studies already showed that the crossed disparity would generate more visual discomfort than the uncrossed disparity[54][99]. Thus, in this chapter, this equation

for static condition has been updated as follows:

$$D_{static} = a_1 r_o + a_2 d + a_3 r_o d + a_4 \tag{10.4}$$

where r_o represents the relative disparity between the static object and the background, and d represents the disparity of the static object (in our study, crossed disparity has a positive value, uncrossed disparity has a negative value). Their interaction is also considered. a_1 to a_4 are coefficients which need to be trained by the subjective data.

Based on the relationship above, two frameworks of the proposed model are proposed. One is based on the tracked object in the video sequence, which is named "Model T". The other is based on the objects in each frame, which is named "Model F". In Model T, all the features in Equation 10.2, 10.3 and 10.4 are calculated based on the tracked objects. In Model F, the features are calculated based on the moving objects in each frame, i.e., no object tracking across the frames. The overview of the frameworks for the two models are shown in Figure 10.1 and 10.2.

The proposed framework includes the following steps:

1. The estimation of the disparity;
2. The estimation of the 3D motion components. In this chapter, the term "3D motion" refers to both the planar motion and the in-depth motion;
3. The extraction of the foreground, background and the moving object(s) in each frame;
4. For Model T, there is one more step for detecting and tracking the moving objects. It should contain the following information: the number of objects that occur in the whole video sequence, the start and the end frames for each object, and the unique identification label for each object.
5. Computation of the visual discomfort induced by foreground $D_{foreground}$. According to the foreground disparity and the background disparity, the visual discomfort induced by the foreground can be calculated according to Equation 10.4.
6. Computation of the visual discomfort induced by the moving objects. For Model T, the features in Equation 10.2 and 10.3 can be calculated based on the definitions. For Model F, the visual discomfort score is calculated based on the objects in each frame. In this case, the disparity d_a is replaced by the in-depth motion velocity v_d if the object in each frame is considered as an independent object.
7. Motion, spatial and temporal pooling to generate a visual discomfort score. As there are two motion components, i.e., planar motion and in-depth motion, there are two visual discomfort scores induced by 3D motion. A pooling

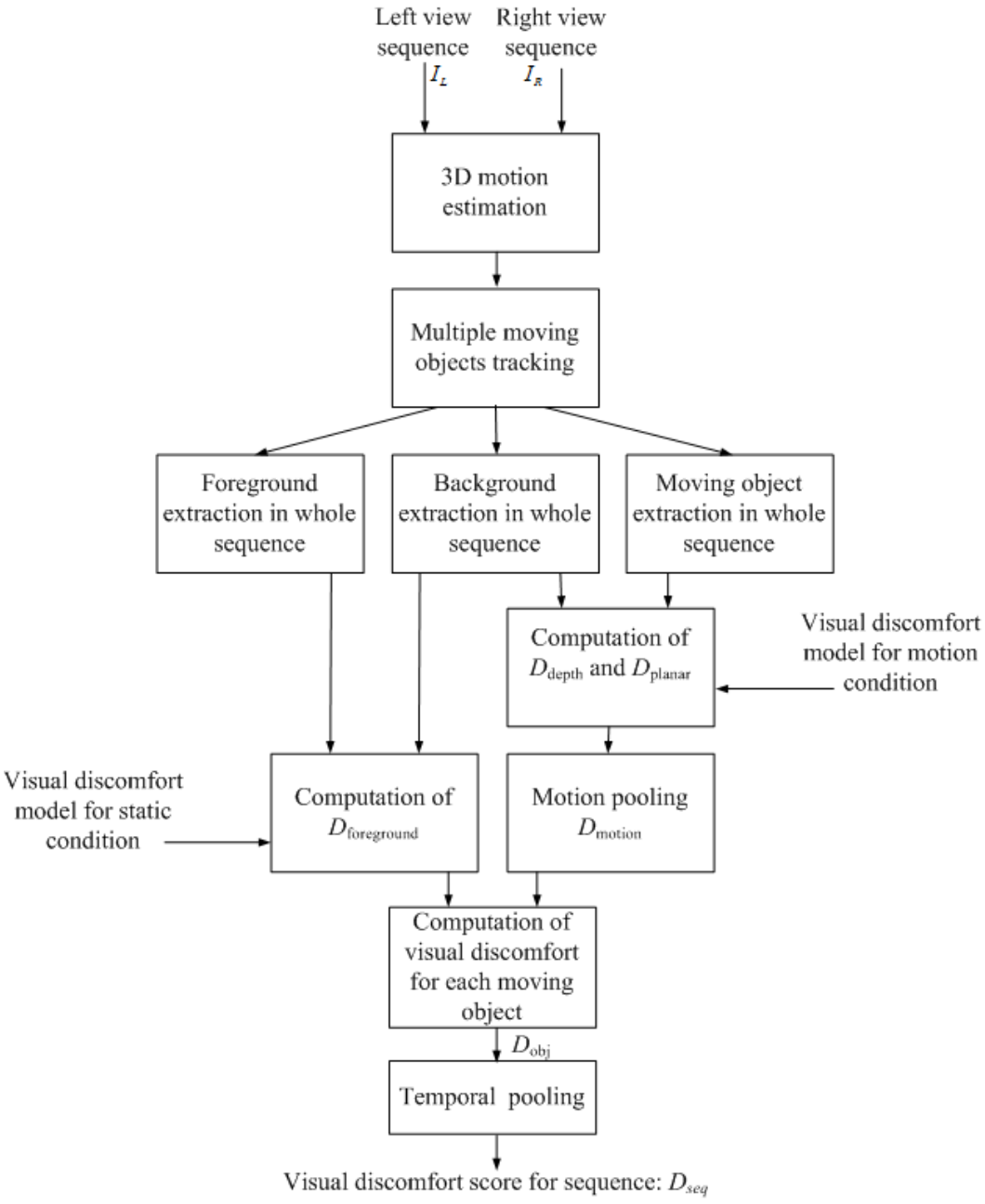


Figure 10.1: Overall framework of Model T.

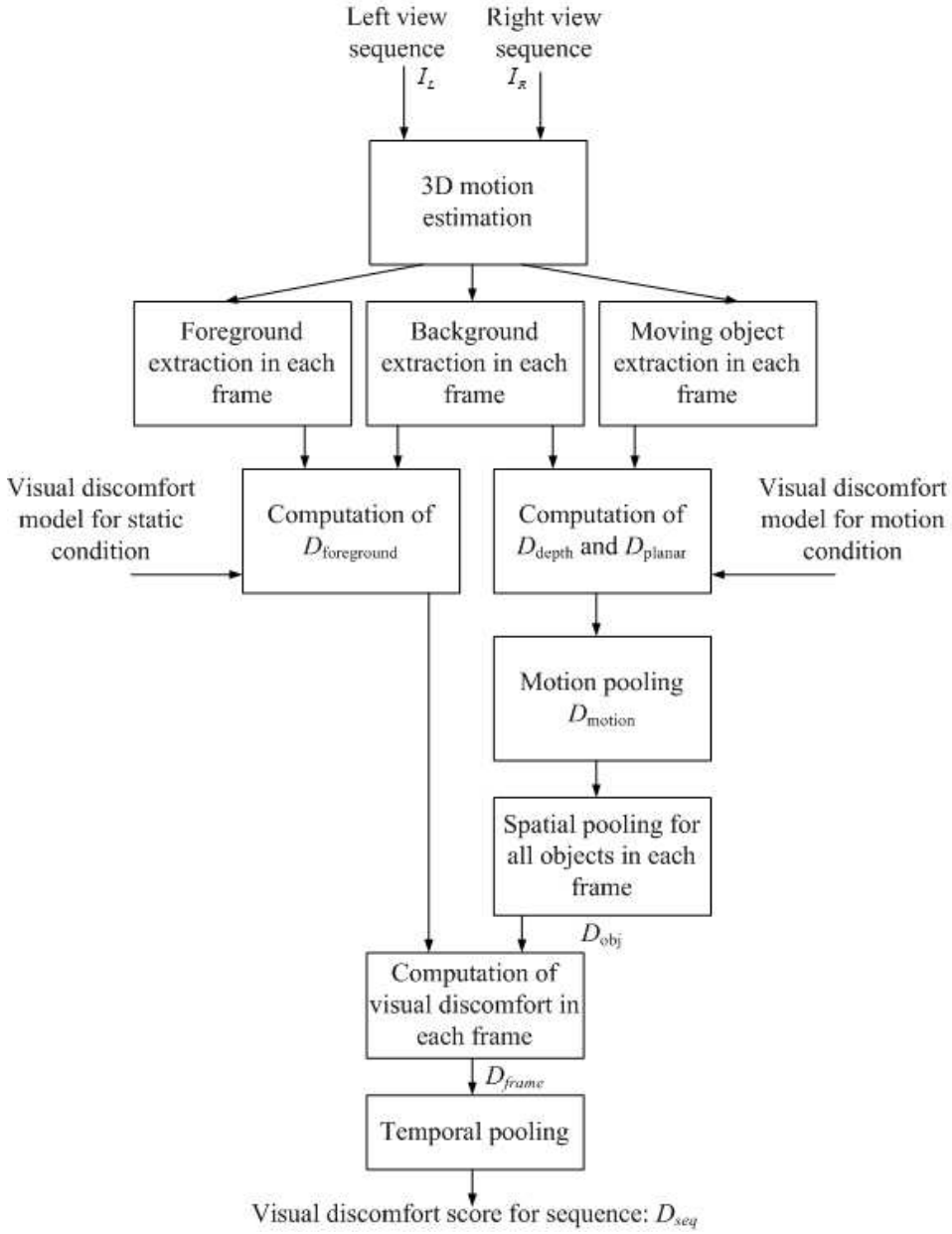


Figure 10.2: Overall framework of Model F.

strategy for the motion score thus needs to be investigated. Furthermore, the possible solutions for spatial and temporal pooling are investigated to generate a final visual discomfort score for a video sequence, D_{seq}.

10.3 Feature extraction

In this section, the methods used to extract the features of the proposed models are introduced. In particular, how to utilize the 3D disparity map and 2D motion map to generate 3D motion maps is described, as well as the requirements for multiple moving object tracking.

10.3.1 3D motion calculation

Figure 10.3 shows the procedure of the extraction of the 3D motion. Firstly, a 2D motion estimation algorithm is used to estimate the motion maps of the left views frame by frame. The obtained maps include a 2D motion map at x direction, $M_{2D_x}(i,j,t)$, and a motion map at y direction, $M_{2D_y}(i,j,t)$, where i represents the row position, j represents the column position on the screen and t represents the current frame.

The disparity map $d(i,j,t)$ can be obtained by applying a disparity estimation algorithm on the left and right views frame by frame, where i, j and t have the same meaning with the 2D motion maps.

As the in-depth motion also affects the results of 2D motion estimation on x-direction, the estimated 2D motion map $M_{2D_x}(i,j,t)$ is a combination of the in-depth motion and 2D x-direction motion. Thus, the in-depth motion related part should be removed from $M_{2D_x}(i,j,t)$ to obtain the "real" planar x-direction motion. The methods are shown in Figure 10.4.

Assuming there is an object moving from position Obj_t to position Obj_{t+1} virtually, the corresponding left view and right view at frame t are in position (x_l^t, y_l^t) and (x_r^t, y_r^t), and at frame $t+1$ they are in positions (x_l^{t+1}, y_l^{t+1}) and (x_r^{t+1}, y_r^{t+1}) on the screen. For better understanding, a compensated object Obj_t^c is created with the same depth level as the object at frame $t+1$, but in line with the previous object Obj_t, i.e., the center position of (x_l^t, y_l^t) and (x_r^t, y_r^t), the virtual position Obj_t and Obj_t^c are in the same line, as shown in Figure 10.4. The positions of the left and right views of the compensated object on the screen is (x_l', y_l') and (x_r', y_r'), respectively. Thus, if there is no vertical disparity, the distance between left and right view of object Obj_t^c is the same as the object Obj_{t+1}, i.e.,

$$x_r' - x_l' = x_r^{t+1} - x_l^{t+1} \tag{10.5}$$

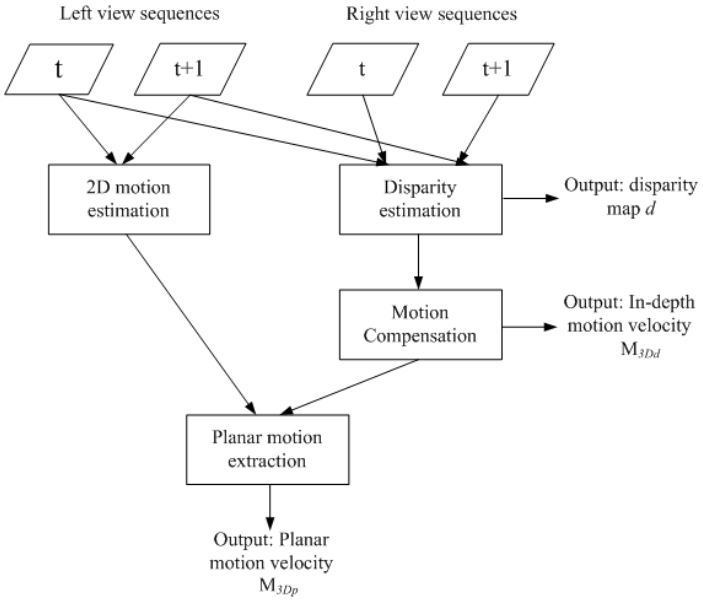


Figure 10.3: Extraction procedure of 3D motion magnitude maps (M_{3D_p}, M_{3D_d}).

Let's take the left view as an example. The 2D motion map $M_{2D_x}(i,j,t)$ in fact measured the distance between x_l^t and x_l^{t+1}. However, the distance between x_l^t and x_l' is induced by in-depth motion. The "real" 2D motion component is from x_l' to x_l^{t+1}.

Therefore, the in-depth motion velocity can be calculated by

$$M_{3D_d}(x_l^t, y_l^t, t) = d(x_l^t + M_{2D_x}(x_l^t, y_l^t, t), y_l^t + M_{2D_y}(x_l^t, y_l^t, t), t+1) - d(x_l^t, y_l^t, t) \tag{10.6}$$

The "real" planar motion maps at x and y directions are:

$$M_{3D_x} = M_{2D_x} - \frac{1}{2} M_{3D_d} \tag{10.7}$$

$$M_{3D_y} = M_{2D_y} \tag{10.8}$$

The planar motion magnitude map is:

$$M_{3D_p} = \sqrt{{M_{3D_x}}^2 + {M_{3D_y}}^2} \tag{10.9}$$

M_{3D_d} and M_{3D_p} are the 3D motion components used in this study.

10.3.2 Multiple moving objects tracking

In a video sequence of natural contents, there are usually several salient moving objects. In the proposed models, the moving objects are considered as a factor that may induce visual discomfort. Thus, it is necessary to detect the moving objects and in particular, for Model T, to track the objects and label them with unique IDs.

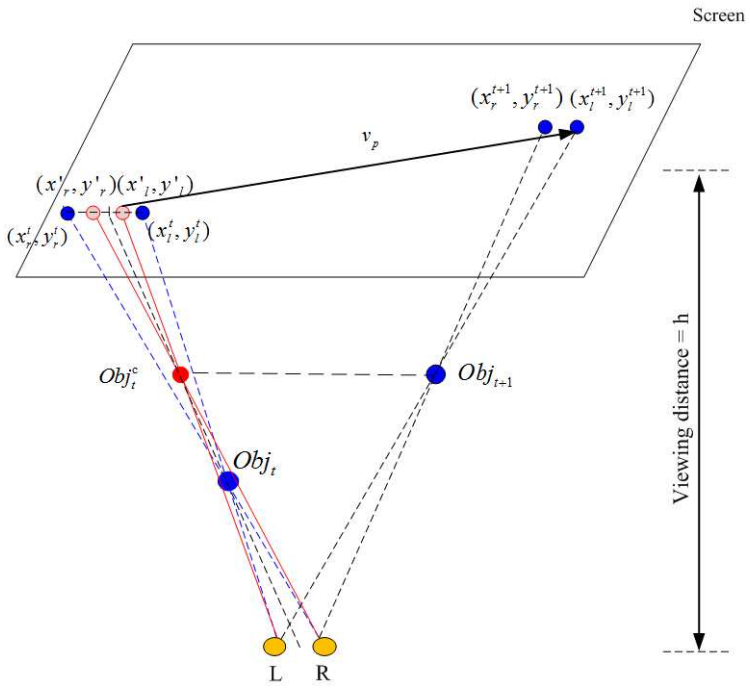


Figure 10.4: Calculation of the 3D motion magnitude maps (M_{3D_p}, M_{3D_d}).

The key points of the object tracking are [28]:

1. Detecting moving objects in each frame;
2. Associating the detections corresponding to the same object over time.

For Model T, based on the tracking results, each moving object's disparity offset, disparity amplitude and velocity are obtained and then used for the calculation of visual discomfort score.

10.4 Pooling strategies of the objective models

In this section, the pooling strategies for the two proposed objective models are introduced, including (1) the pooling strategy for the planar motion and the in-depth motion induced visual discomfort scores, and (2) the pooling strategy for the multiple moving objects as well as the pooling strategy for all frames of a video.

According to the tracked objects in Model T, and the detected objects of each frame in Model F, the disparity, 3D motion velocity are expressed as follows for better explanation:

- $d_{obj}(I,t)$: disparity value of object I at frame t.
- $v_{3D_x}(I,t)$: 3D x-direction motion velocity for object I at frame t.
- $v_{3D_y}(I,t)$: 3D y-direction motion velocity for object I at frame t.
- $v_{3D_p}(I,t)$: 3D planar motion velocity for object I at frame t.
- $v_{3D_d}(I,t)$: 3D in-depth motion velocity for object I at frame t.

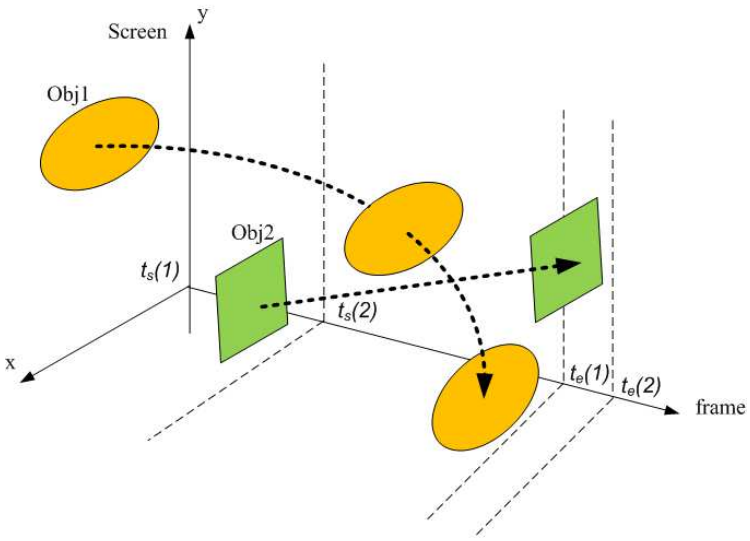


Figure 10.5: An example of the tracking of multiple objects in a video sequence.

10.4.1 Pooling strategy for Model T

Motion pooling strategy

Assuming there are in total N objects appearing in the whole video sequence, an example is shown in Figure 10.5. Each object is assigned with a unique label $I, I = 1, 2, ..., N$. The start frame and the end frame for Object I is $t_s(I)$ and $t_e(I)$.

For a moving object, pooling the visual discomfort scores induced by different types of motion is an important issue in the model. Here, three strategies are proposed, i.e., the visual discomfort induced by 3D motion D_{motion} can be obtained by the mean, the L2-Norm, or the maximum value of the two visual discomfort scores D_{planar} and D_{depth}:

1. $D_{motion}(I) = \frac{1}{2}(D_{planar}(I) + D_{depth}(I))$.
2. $D_{motion}(I) = \sqrt{D^2_{planar(I)} + D^2_{depth(I)}}$
3. $D_{motion}(I) = \max(D_{planar}(I), D_{depth}(I))$

For Model T, as we already described in Section 10.2, the $D_{planar}(I)$ and $D_{depth}(I)$ can be calculated according to the original definition of the disparity amplitude, relative disparity offset, and the planar and in-depth motion velocity of Object I. The $d_{obj}(I,t)$, $v_{3D_p}(I,t)$ and $v_{3D_d}(I,t)$ are obtained by calculating the median value of the corresponding disparity map and 3D motion map of the detected object's area at frame t. The median value is used here because it is robust to the errors along the edge of the object, where the disparity estimation and motion estimation were prone to large estimation errors.

Spatial & Temporal pooling strategy

Besides the tracked objects, the corresponding foreground and background should be extracted as well. The visual discomfort induced by the foreground at frame t is $D_{foreground}(I,t)$. The tracked object's related foreground induced visual discomfort is:

$$D_{fore}(I) = median(D_{foreground}(I,t)), t = t_s(I) : t_e(I) \tag{10.10}$$

The visual discomfort based on Object I is:

$$D_{obj}(I) = D_{fore}(I) + D_{motion}(I) \tag{10.11}$$

For the video sequence, the visual discomfort score can be calculated by the following ways:

1. $D_{seq} = \frac{1}{N}\sum_{I} D_{obj}(I)$
2. $D_{seq} = median(D_{obj}(I)), I = 1 : N$
3. $D_{seq} = max(D_{obj}(I)), I = 1 : N$

10.4.2 Pooling strategy for Model F

Motion pooling strategy

Unlike Model T, the moving objects for Model F are detected in each frame, and each object in each frame is labeled with a unique ID. Assuming there are in total N objects in one frame, and each object is assigned with a unique label $I, I = 1, 2, ..., N$, the visual discomfort induced by 3D motion D_{motion} can be obtained by the mean, L2-Norm, or maximum value of the two visual discomfort scores D_{planar} and D_{depth}:

1. $D_{motion}(I,t) = \frac{1}{2}(D_{planar}(I,t) + D_{depth}(I,t))$.
2. $D_{motion}(I,t) = \sqrt{D^2_{planar(I,t)} + D^2_{depth(I,t)}}$
3. $D_{motion}(I,t) = \max(D_{planar}(I,t), D_{depth}(I,t))$

For Model F, the D_{planar} and D_{depth} are calculated according to Equation 10.2 and 10.3. As each object in each frame is considered as an independent object, the disparity amplitude is replaced by the in-depth motion velocity, and the relative disparity offset r_o is the difference between the object's disparity and the background's disparity.

The same with Model T, the moving object's disparity, planar motion velocity and in-depth motion velocity are obtained by calculating the median value of the corresponding disparity map and 3D motion map of the detected object's area.

Spatial & Temporal pooling strategy

The foreground and background in each frame should be extracted. The visual discomfort induced by the foreground at frame t is $D_{foreground}(t)$. The visual discomfort score of frame t is:

$$D_{frame}(t) = D_{obj}(t) + D_{foreground}(t) \tag{10.12}$$

where $D_{obj}(t)$ can be obtained by the following ways:

1. $D_{obj}(t) = \frac{1}{N}\sum_{I=1}^{N} D_{motion}(I,t)$
2. $D_{obj}(t) = median(D_{motion}(I,t)), I = 1:N$
3. $D_{obj}(t) = max(D_{motion}(I,t)), I = 1:N$

For the video sequence, the visual discomfort score can be calculated by the following way:

$$D_{seq} = median(D_{frame}(t)), t = 1 : \text{total number of frames} \tag{10.13}$$

10.5 Performances of the proposed models

In this section, the proposed two models are evaluated by the two subjective experimental results (i.e. ACR and PC results) in Chapter 9.

As there is no ground truth for the disparity and 2D motion velocity, the disparity map and 2D motion map for each frame are generated by a disparity estimation algorithm based on a first order primal-dual convex optimization algorithm proposed by Chambolle *et al.* [16]. Then, the 3D motion maps are calculated according to the method in Section 10.3.1. The disparity of the background is obtained by averaging the minimum 10% disparity values. The disparity of the foreground is obtained by averaging the maximum 10% disparity values.

The moving objects tracking algorithm and detection algorithm would largely affect the results, thus, to evaluate the proposed models without being interrupted by other factors, the moving object were tracked by manually labeling the ID to the objects every 10 frames. Thus, the positions of all objects were obtained. Then, the object's planar motion and in-depth motion velocities and disparity value were obtained according to the object's position on 3D motion maps and disparity map.

10.5.1 Evaluation of the disparity and motion estimation algorithm

Before evaluating the performance of the proposed models, the accuracy of the estimation algorithm on disparity and 3D motion velocity needs to be evaluated. For

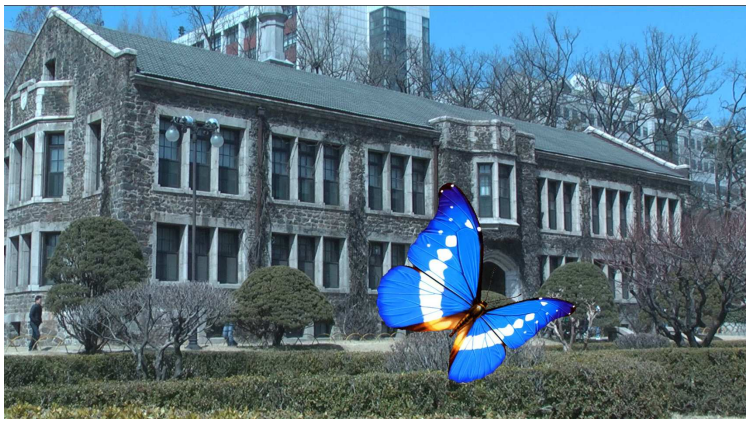


Figure 10.6: A preview of the used synthetic stimuli to evaluate the performance of the disparity and motion estimation algorithm.

Table 10.3: The estimated results on the synthetic stimuli

Sequence	Estimated value					Ground truth value				
	v_p	v_d	d_o	d_a	d_b	v_p	v_d	d_o	d_a	d_b
1	6.15	0.07	-1.29	0.00	-1.41	6	0	-1.3	0	-1.4
5	15.86	0.00	-0.64	0.01	-1.41	15	0	-0.65	0	-1.4
8	15.84	0.00	0.00	0.00	-1.41	15	0	0	0	-1.4
12	25.92	0.01	0.64	0.03	-1.42	24	0	0.65	0	-1.4
13	6.42	0.02	1.30	0.01	-1.41	6	0	1.3	0	-1.4
21	0.14	1.32	0.05	1.30	-1.41	0	1	0	1.3	-1.4
25	0.01	2.02	0.03	2.06	-1.41	0	2	0	2	-1.4
29	0.06	3.12	0.06	2.58	-1.41	0	3	0	2.6	-1.4
30	0.06	0.91	-0.59	1.29	-1.41	0	1	-0.65	1.3	-1.4
33	0.05	3.02	-0.63	0.64	-1.41	0	3	-0.65	0.65	-1.4
RMSE	0.65	0.11	0.03	0.02	0.01	-	-	-	-	-

the purpose of evaluation, the disparity and motion velocity estimation algorithm were applied on some typical synthetic stimuli. These stimuli were generated by the same way as the stimuli in Chapter 8 except for the foreground and the background images. The foreground is replaced by a butterfly and the background is replaced by a building image from IEEE P3333.1 3D Image Database, which is member-only available in http://grouper.ieee.org/groups/3dhf/. An example of the synthetic stimuli is shown in Figure 10.6. With the known disparity and 3D motion velocity values, the performance of the tested algorithm on natural content can be evaluated.

In this study, five planar motion stimuli (indexed with 1, 5, 8, 12 and 13) and five in-depth motion stimuli (indexed with 21, 25, 29, 30, 33) were chosen. The disparity and velocity values are shown in Table 8.2. These stimuli covered various disparity levels and velocity levels.

The moving object in the synthetic stimuli were manually tracked by assigning a unique ID for each object every 10 frames and the corresponding positions of all objects were recorded. The moving object's disparity, planar motion velocity and

in-depth motion velocity were obtained by extracting the values of the disparity map and 3D motion map at the recorded position. The estimated disparities, velocities and the ground truth (designed data) are shown in Table 10.3. The unit for planar motion velocity v_p and in-depth motion velocity v_d are degree/s. The disparity offset d_o, disparity amplitude d_a and background disparity d_b are in unit of degree.

As shown in Table 10.3, the RMSE between the estimated data and the ground truth is very low. Thus, the performances of the algorithm used on disparity and motion estimation are satisfactory in this study and thus can be applied on the proposed models.

10.5.2 Evaluation of the proposed model

In this section, the proposed object tracking-based model and frame-based model are evaluated by the results of two subjective experiments which have been previously introduced in Chapter 9. For better visualization and comparison, the MOS is converted to represent the degree of **visual discomfort**. Thus, the original MOS values are subtracted from 5 and get the converted score, where "0" represents "very comfortable" and "4" represents "extremely uncomfortable" (thus, the following MOS is in fact the converted MOS if there is no other specific explanation). The manually tracking results are used in this section to avoid the influence from the performance of the tracking algorithm.

10.5.3 Performance of the proposed models

According to the proposed pooling strategies in this study, there are in total nine conditions to be evaluated, i.e., for the pooling of visual discomfort induced by planar motion D_{planar} and in-depth motion D_{depth}, the "Mean","Max" and "L2-Norm" are used; for the spatial/temporal pooling of the visual discomfort induced by multiple moving objects, the "Mean", "Median" and "Max" pooling strategies are used. In both models, only four parameters need to be trained, i.e., a_1, a_2, a_3 and a_4 in Equation 10.4. They are trained by the MOS and BT scores separately. In this study, the "Particle Swarm Optimization (PSO)" algorithm is used for training the parameters, the objective function is the CC value between predicted scores and the MOS or BT scores. Thus, the obtained fitting parameters would generate the highest CC between the subjective data and the predicted scores. For more details about this optimization algorithm, the readers are referred to [24].

The performances of Model T and Model F are shown in Table 10.4 and Table 10.5, respectively. The CC, SROCC and RMSE between the predicted scores and the subjective scores are used for evaluation. Please note that the predicted scores have been fitted to the subjective scores before calculating the CC, SROCC and

Table 10.4: Performance of Model T

Motion pooling	Spa.& Tem. pooling	BT scores			MOS		
		CC	SROCC	RMSE	CC	SROCC	RMSE
Mean	Mean	**0.80**	0.61	**0.48**	0.82	0.77	0.23
Mean	Median	0.77	0.56	0.52	0.79	0.74	0.25
Mean	Max	0.70	0.50	0.58	**0.84**	0.78	**0.22**
Max	Mean	0.77	0.61	0.52	0.83	0.78	**0.22**
Max	Median	0.77	0.58	0.52	0.79	0.73	0.24
Max	Max	0.64	0.52	0.63	0.83	**0.81**	**0.22**
L2-Norm	Mean	0.79	0.56	0.50	0.82	0.77	0.23
L2-Norm	Median	**0.80**	0.60	**0.48**	0.79	0.73	0.24
L2-Norm	Max	0.66	**0.63**	0.61	**0.84**	0.79	**0.22**
Jung's model[65]		-	-	-	0.81	0.77	0.27

RMSE, which is recommended by the VQEG final report on the validation of objective models of video quality assessment [138] (The same for all the other CC, SROCC, RMSE values in this study). For better visualization but in limited space, some examples of the scatter plot of the predicted scores and the subjective data are shown in Figure 10.7 and 10.8.

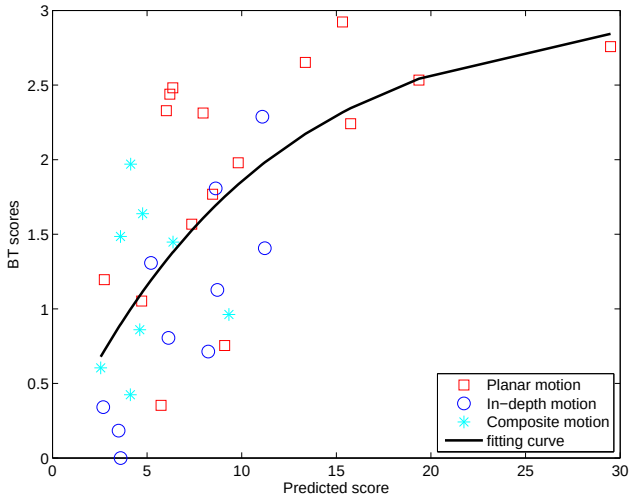
The results in Table 10.4 and Table 10.5 indicate that the performances of the two proposed models are comparable in this database. This result might be due to the simplicity of this database, i.e., no scene cut in the video sequences. Therefore, the comparability of the performances of the two models might requires more databases to be verified.

Generally, for both models, the results of the "Mean-Mean" method and the "L2-Norm-Median" method showed higher correlation with the subjective BT scores. The results of the "Mean-Max" and "L2-Norm-Max" methods showed higher correlation with the MOS while the "Max-Max" method would generate the most consistent results with the MOS on rankings. The results also indicate that there might be no universal optimal pooling strategy for different test methodologies. Large databases are required for the verification of this conclusion.

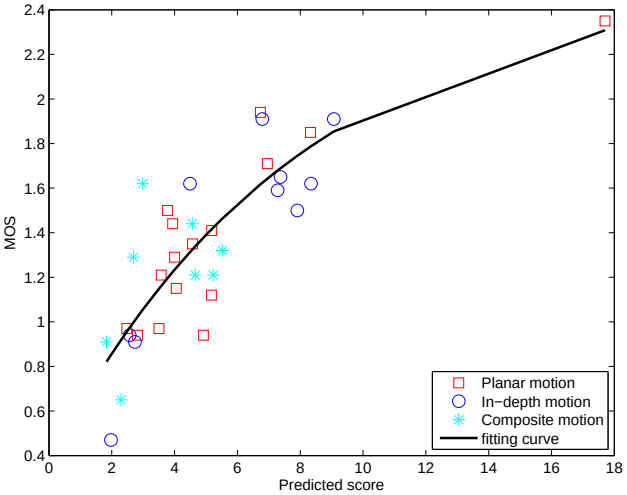
The trained a_1, a_2, a_3, and a_4 for Model T and Model F are shown in Table 10.6 and Table 10.7, respectively. Generally, the main difference between the parameters for BT scores and MOS is that a_1 is negative for BT score but positive for MOS. a_1 is the coefficient for the relative disparity of the foreground r_o. The positive value is consistent with our previous study, i.e., the relative disparity has a significant influence on visual discomfort, larger relative distance between the foreground and the background would generate more visual discomfort. However, for the BT scores, the negative value of a_1 might be explained by the effects of window violation in this database. In paired comparison test, the window violation has a dominant influence on the results, thus, the higher of the disparity of the foreground (the disparity of the foreground in front of the screen is a positive value), the more perceived visual

Table 10.5: Performance of Model F

Motion pooling	Spa.& Tem. pooling	BT scores			MOS		
		CC	SROCC	RMSE	CC	SROCC	RMSE
Mean	Mean	**0.80**	0.62	**0.49**	0.81	0.77	0.23
Mean	Median	0.78	0.53	0.50	0.79	0.74	0.25
Mean	Max	0.70	0.58	0.58	**0.85**	0.79	**0.21**
Max	Mean	0.79	0.61	0.50	0.82	0.79	0.23
Max	Median	0.73	0.53	0.56	0.79	0.73	0.24
Max	Max	0.68	**0.65**	0.59	0.84	**0.81**	0.22
L2-Norm	Mean	0.79	0.63	**0.49**	0.82	0.77	0.23
L2-Norm	Median	**0.80**	0.60	**0.49**	0.79	0.73	0.25
L2-Norm	Max	0.68	0.64	0.60	**0.85**	0.79	**0.21**
Jung's model[65]		-	-	-	0.81	0.77	0.27

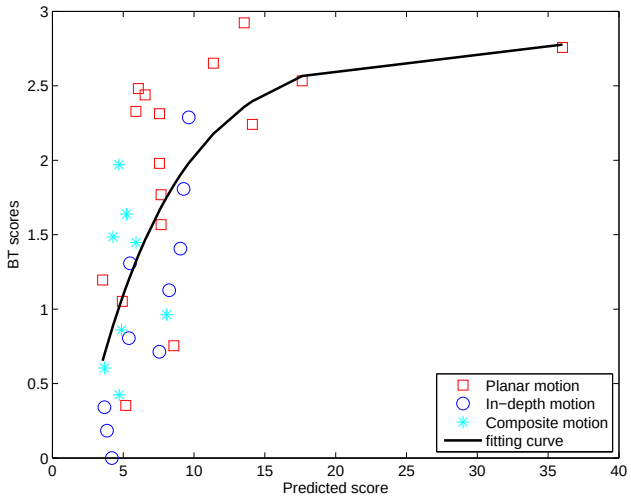


(a) Scatter plot of the BT scores and Model T using 'Mean-Mean' method.

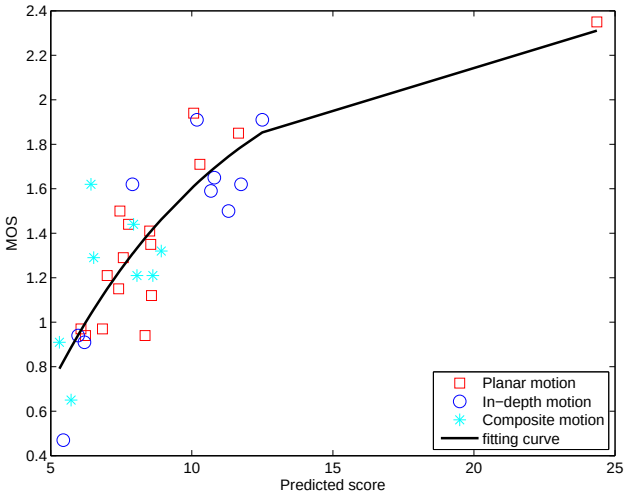


(b) Scatter plot of MOS and Model T using 'Mean-Max' method.

Figure 10.7: Scatter plot of the trained Model T results and the subjective scores.



(a) Scatter plot of the BT scores and Model F using 'Mean-Mean' method.



(b) Scatter plot of MOS and Model F using 'Mean-Max' method.

Figure 10.8: Scatter plot of the trained Model F results and the subjective scores.

Table 10.6: The trained parameters of Model T

Motion pooling	Spa.& Tem. pooling	BT scores				MOS			
		a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
Mean	Mean	-4.40	1.64	-3.33	-1.58	2.19	2.56	-1.72	1.73
Mean	Median	-5.19	1.25	-3.49	-0.74	4.34	4.18	-1.52	0.54
Mean	Max	-0.39	1.70	-7.74	3.16	0.85	0.16	-0.79	-0.11
Max	Mean	-1.42	4.02	-7.92	-0.39	3.06	2.48	-1.87	5.79
Max	Median	-2.94	3.02	-6.39	1.12	5.82	4.58	-1.98	-4.94
Max	Max	-8.43	8.62	2.59	2.37	2.45	0.44	-1.01	-1.00
L2-Norm	Mean	-5.20	0.19	0.61	-0.17	3.17	3.09	-1.99	-1.03
L2-Norm	Median	-2.30	2.12	-3.18	0.32	6.53	5.72	-2.25	-4.70
L2-Norm	Max	-1.10	6.39	-10	2.18	1.91	0.30	-1.17	0.34

Table 10.7: The trained parameters of Model F

Motion pooling	Spa.& Tem. pooling	BT scores				MOS			
		a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
Mean	Mean	-2.24	2.15	-3.03	2.42	2.28	2.61	-1.74	2.54
Mean	Median	-4.42	0.06	-0.91	-2.97	4.37	4.28	-1.54	-8.16
Mean	Max	-2.56	0.74	-2.18	-1.72	0.79	0.01	-0.59	3.33
Max	Mean	-2.87	1.65	-2.40	5.58	3.28	2.81	-2.10	-0.17
Max	Median	-1.82	0.05	-0.98	-2.90	5.87	4.70	-2.06	0.39
Max	Max	-0.54	3.93	-6.83	2.06	2.27	0.18	-0.78	3.59
L2-Norm	Mean	-3.11	3.05	-4.38	-0.48	3.49	3.51	-2.32	-0.49
L2-Norm	Median	-2.15	1.73	-2.71	0.72	6.64	5.94	-2.37	0.01
L2-Norm	Max	-0.81	6.32	-10	4.70	1.79	0.11	-1.05	2.36

discomfort. However, the sequences which contain window violation are mainly with small relative disparity between the foreground and the background, for example, sequence 6, 12 to 15 and 27, 30. Thus, for this database, the coefficient a_1 is negative when fitting to the BT scores.

10.5.4 Considering the effects of window violation

As we already discussed in Chapter 9, window violation in the sequences is a possible significant factor that induces the difference between ACR results and the PC results. Generally, windows violation can be detected by the following ways [28]:

1. Perform connected component analysis on pixels with large crossed disparity;
2. If the detected object is in contact with the boundary of the screen, there is windows violation in the sequence.

Thus, in this section, the effect of window violation is considered in the objective models: the final visual discomfort score for a video sequence D_{seq} is the sum of the original score and window violation induced score D_{wv}.

Table 10.8: Performance of Model T considering window violation

Motion pooling	Spa.& Tem. pooling	BT scores			MOS		
		CC	SROCC	RMSE	CC	SROCC	RMSE
Mean	Mean	0.78	0.58	0.51	0.82	0.78	0.23
Mean	Median	0.79	0.56	0.50	0.79	0.76	0.24
Mean	Max	0.70	0.51	0.58	**0.84**	0.79	**0.22**
Max	Mean	0.77	0.58	0.52	0.83	0.79	**0.22**
Max	Median	0.77	0.59	0.52	0.80	0.76	0.24
Max	Max	0.67	**0.64**	0.63	0.83	**0.82**	**0.22**
L2-Norm	Mean	**0.80**	0.63	0.49	0.82	0.78	0.23
L2-Norm	Median	**0.80**	0.61	**0.48**	0.79	0.76	0.24
L2-Norm	Max	0.66	0.61	0.61	**0.84**	0.79	**0.22**
Jung's model[65]		-	-	-	0.81	0.77	0.27

In this database, Sequence 6, 10, 12, 13, 14, 15, 19, 24, 27, 30, 31, 33, and 36 contain windows violation by visually check. It is observed that higher crossed disparity of window violation would lead to higher perceived visual discomfort. Thus, the effect of window violation on visual discomfort is:

$$D_{wv} = a_5 d_w \tag{10.14}$$

where d_w is the maximum crossed disparity of the detected moving objects that contain window violation, which is computed based on the rule of "winner takes all". a_5 is the coefficient. In Model T, D_{wv} is added to Equation 10.11 and in Model F, it is added to Equation 10.12.

The performance of Model T and Model F considering the window violation effect are shown in Table 10.8 and Table 10.9, respectively. Their corresponding trained coefficients are shown in Table 10.10 and Table 10.11.

The objective models which have considered the windows violation did not show significant improvement compared to the original one. This might be because the interdependency between the relative disparity and the windows violation in this database. As we already discussed in the previous section, the sequences which contain windows violation also have small relative disparity between the foreground and the background. The factors r_o and d_w cannot be separately studied in this study.

10.6 Evaluation of a proposed objective visual discomfort algorithm

In Section 10.5, to evaluate the performances of the proposed models, the moving object is tracked manually by assigning a unique ID. In this section, a multiple moving objects tracking algorithm is applied in the proposed models to replace the

Table 10.9: Performance of the Model F considering window violation

Motion pooling	Spa.& Tem. pooling	BT scores			MOS		
		CC	SROCC	RMSE	CC	SROCC	RMSE
Mean	Mean	**0.80**	0.64	0.49	0.81	0.78	0.23
Mean	Median	**0.80**	0.61	0.49	0.79	0.75	0.24
Mean	Max	0.74	0.55	0.54	**0.85**	0.79	**0.21**
Max	Mean	**0.80**	0.64	**0.48**	0.83	0.80	0.22
Max	Median	**0.80**	0.60	0.49	0.80	0.77	0.24
Max	Max	0.72	0.57	0.56	0.84	**0.82**	0.22
L2-Norm	Mean	**0.80**	**0.65**	0.49	0.82	0.78	0.23
L2-Norm	Median	**0.80**	0.60	0.49	0.79	0.77	0.24
L2-Norm	Max	0.68	0.64	0.60	**0.85**	0.80	0.22
Jung's model[65]		-	-	-	0.81	0.77	0.27

Table 10.10: The trained parameters of Model T considering the window violation

Motion pooling	Spa.& Tem. pooling	BT scores					MOS				
		a_1	a_2	a_3	a_4	a_5	a_1	a_2	a_3	a_4	a_5
Mean	Mean	-3.74	3.31	-6.11	-1.11	0.42	2.05	3.35	-2.21	-0.97	-1.06
Mean	Median	-3.05	1.02	-1.36	4.29	1.20	4.23	6.11	-2.29	4.11	-2.50
Mean	Max	-0.72	1.75	-7.53	6.21	0.16	0.97	1.06	-1.40	0.88	-0.89
Max	Mean	-4.11	2.29	-5.13	0.31	0.13	2.84	3.49	-2.15	-1.41	-1.65
Max	Median	-3.19	2.82	-5.90	5.87	0.10	5.35	6.45	-2.60	-0.30	-3.23
Max	Max	-0.42	8.43	-8.48	3.80	0.82	2.34	1.56	-1.66	2.01	-1.62
L2-Norm	Mean	-5.32	3.20	-5.22	2.96	1.62	3.15	4.40	-2.69	1.88	-1.62
L2-Norm	Median	-2.78	0.67	-2.11	-0.32	1.54	6.23	8.24	-3.26	-5.65	-3.54
L2-Norm	Max	-0.73	5.44	-10	2.33	2.02	1.89	1.32	-1.89	-4.69	-1.31

Table 10.11: The trained parameters of Model F considering window violation

Motion pooling	Spa.& Tem. pooling	BT scores					MOS				
		a_1	a_2	a_3	a_4	a_5	a_1	a_2	a_3	a_4	a_5
Mean	Mean	-2.79	3.18	-4.32	3.46	1.24	2.27	3.59	-2.22	0.47	-1.19
Mean	Median	-2.41	1.80	-2.68	-0.72	1.11	4.19	6.04	-2.18	4.65	-2.43
Mean	Max	-1.05	1.92	-7.55	-0.60	0.02	0.80	0.29	-0.79	3.55	-0.29
Max	Mean	-2.11	1.43	-2.63	2.16	0.62	3.13	4.13	-2.64	1.83	-1.91
Max	Median	-0.90	1.50	-2.29	-0.30	0.80	5.41	6.58	-2.50	1.22	-2.94
Max	Max	-5.30	4.04	-7.44	4.39	2.59	2.27	1.12	-1.34	-1.43	-1.06
L2-Norm	Mean	-2.91	2.73	-3.98	-0.24	1.57	3.38	4.77	-2.85	-1.09	-1.72
L2-Norm	Median	-1.99	2.88	-3.98	-0.29	1.74	6.40	8.62	-3.33	-0.89	-3.61
L2-Norm	Max	-0.20	4.71	-10	6.22	0.62	1.77	0.60	-1.08	2.36	-0.67

Table 10.12: Performance of an implementation of Model F

Motion pooling	Spa.& Tem. pooling	BT scores			MOS		
		CC	SROCC	RMSE	CC	SROCC	RMSE
Mean	Mean	0.78	0.63	0.51	0.77	0.74	**0.25**
Mean	Median	**0.79**	**0.76**	**0.50**	0.75	0.69	0.26
Mean	Max	0.74	0.64	0.55	**0.78**	0.76	**0.25**
Max	Mean	0.64	0.53	0.63	0.75	0.70	0.26
Max	Median	0.70	0.70	0.53	0.74	0.66	0.27
Max	Max	0.64	0.53	0.63	0.76	0.73	0.26
L2-Norm	Mean	0.76	0.70	0.53	0.76	0.71	0.26
L2-Norm	Median	**0.79**	0.74	**0.50**	0.74	0.67	0.27
L2-Norm	Max	0.66	0.65	0.61	0.77	**0.77**	**0.25**

manually tracking results. Thus, by combining together the disparity estimation algorithm, motion estimation algorithm and the multiple object tracking algorithm, the integrated algorithm can be considered as an objective visual discomfort algorithm.

The *multiObjectTracking* function in Computer Vision System Toolbox of MATLAB 2013 is used (http://www.mathworks.fr/fr/help/vision/examples/motion-based-multiple-object-tracking.html), which is a simple and fast implementation algorithm on object tracking. In the *multiObjectTracking* function, the detection of the moving objects uses a background subtraction algorithm based on Gaussian mixture models. The association of detections to the same object is based solely on motion. The motion of each track is estimated by a Kalman filter. The filter is used to predict the track's location in each frame, and determine the likelihood of each detection being assigned to each track. The results include the ID of each detected object, the positions of the object, the size of the object, and the started frame and the ended frame for each object.

The performance of the objective visual discomfort algorithm for Model F is shown in Table 10.12. As shown in the table, the performance of this algorithm is a little worse than the manually tracking based model for the MOS results. However, for paired comparison results, the "Mean-Median" pooling method, and the "L2-Norm-Median" methods could generate very consistent results with the subjective data. For the ACR test results, similar as Model F using manual tracking results, the "Mean-Max" and "L2-Norm-Max" methods would generate more reliable prediction results on visual discomfort. According to the CC, SROCC and RMSE values, the results indicate that this algorithm is applicable to obtain good prediction results on visual discomfort.

10.7 Conclusions

In this chapter, the performance of the our objective visual discomfort model was evaluated by the natural 3D video sequences. Two implementations for the objective model are proposed which lead to two models in this chapter. One is based on the tracked object called “Model T” and the other is based on the moving objects in each frame, called “Model F”. Both models generate highly correlated results with the subjective scores.

In this study, the parameters of the models are trained by the results of two different subjective experimental results, one was obtained by the ACR method, and the other was obtained by the PC method. The resulting parameters indicate and verify the differences between the two test results, i.e., window violation has different effect on different test methodologies.

An objective visual discomfort algorithm is proposed in this study, where an easy and fast multiple moving object tracking algorithm, as well as an disparity and motion estimation algorithm are integrated. The results of the proposed model shows a high correlation with the subjective scores. Thus, this algorithm is applicable in the real application of automatically assessing the visual discomfort induced by 3D videos.

Future work would be the generalization of the proposed models to distorted video sequences, particularly for the shooting errors which would induce binocular distortions.

11

Objective psychophysical prediction of visual discomfort

Visual discomfort can be predicted by physiological signals. In existing studies, the relationship between the psychophysical predictor and visual discomfort is usually studied by using video sequences of natural content, where the influence factors are combined together. It would be very interesting to find the influence of each factor on the physiological signals. Thus, in this chapter, the synthetic stimuli in Chapter 8 were employed again for the study on the relationship between 3D video characteristics (e.g., motion type, disparity, velocity, etc), visual discomfort and eye blinking rate.

11.1 Introduction

As introduced in previous chapters, visual discomfort can be predicted by subjective assessment and objective devices. Subjective assessment is based on the participant's subjective opinion, e.g., Questionnaire, Paired Comparison test, SSCQE (Single Stimulus Continuous Quality Evaluation), etc. Objective prediction is often based on physiological signals, e.g., eye pressure, blinking rate, electrical activity of the brain, etc. In this chapter, we focus on the objective psychophysical prediction.

In the study of [73], the authors used an electroencephalography (EEG) device to detect visual fatigue. The results showed that in the beta band of EEG, the power of the EEG signals in watching 3D video was significantly larger than in watching 2D conditions. In [70], the authors used the functional magnetic resonance imaging

(fMRI) to test visual fatigue in 3D condition, the results showed that there were strong activities in the frontal eye field (FEF) [111]. Studies already showed that FEF region plays a significant role in the planning and execution of saccadic eye movements and participates in the control of visual selective attention [7][14]. This result might be an indicator that the eye movement and eye blinks are possible measures for assessing visual fatigue. Nahar et.al [97] studied the electromyography (EMG) response of the orbicularis oculi muscle to different visual stress conditions, the results showed that only for the squint-beneficial test conditions (e.g., refractive error, glare), the power of the EMG response increased with the degree of eyestrain.

Eye blinking rate has been considered as an indicator for predicting visual discomfort or visual fatigue. Studies showed that when in relaxed conditions, people would blink more often than in book reading and computer reading tasks [131]. In [82][151], the results showed that blinking rate was higher in watching 3D video than in 2D. The study of [69] gives the conclusion that eye blinking rate increases with visual fatigue when watching 3D images. For the conditions employing a visual display unit (VDU, a visual display device for a computer), the blinking frequency was significantly decreased during the fatigued condition (e.g., read information from the screen for a long time) [30]. In conclusion, eye blinking performs quite differently in different conditions, e.g., in relax condition, reading, long term use of VDU, watching 2D images and 3D images.

So far, there is no distinct study on the relationship between eye blinks and watching synthetic stereoscopic stimuli with controlled disparity and velocity. Thus, the objective of this study is to find out the relationship between eye blinking rate, 3D video characters (e.g., disparity offset, disparity amplitude, velocity, motion type) and visual discomfort.

11.2 Experiment

11.2.1 Apparatus and environment

The display used in this study is the same as in Section 8.2.3, i.e., the Dell Alienware AW2310 23-inch 3-D LCD screen (1920×1080 full HD resolution, which featured 0.265-mm dot pitch, 120 Hz) with active shutter glasses (NVIDIA 3D vision kit). Viewing distance was about 90 cm (three times of the screen height). The viewing environment was adjusted according to ITU-R BT.500 [58].

The electro-physiological measurement device Porti from TMSi was used to obtain the EMG signals (EMG: the electrical activity produced by skeletal muscles) with eye-blinking data. The sample rate is 2048 Hz. Eight surface electrodes were affixed with conducting paste (Tac-Gel) at the outer canthus, inner canthus, top eyelid and bottom eyelid positions of both eyes. Besides, a reference channel is

Table 11.1: Design of the added 8 stimuli in the experiment

Index	d_a (degree)	d_o (degree)	v_p (degree/s)	v_i (degree/s)
37	0.65	0	0	1
38	0.65	0	0	2
39	0.65	0	0	3
40	1.3	-0.65	0	2
41	0.65	-0.65	0	2
42	1.3	0.65	0	2
43	0.65	0.65	0	1
44	0.65	0.65	0	2

placed on the forehead about 2 cm above the eyes.

11.2.2 Stimuli

There are 44 stimuli used in this test. 36 out of them are exactly the same as we already used in Chapter 8.2.4. More details can be found in Table 8.2. The remaining 8 stimuli are designed as shown in Table 11.1. d_o represents disparity offset, d_a represents disparity amplitude, v_p is the planar motion velocity, v_i is the in-depth motion velocity. The foreground and background are exactly the same as in our previous test.

11.2.3 Subjects and Procedure

Twenty-eight naive observers participated in this subjective test. 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. Observers were asked to watch each of the stimuli for a duration of 10 seconds. 44 stimuli were displayed. The presentation order was randomly permuted for each observer. The EMG signals were recorded from the 8 electrodes.

11.3 Relationship between eye blinking rate and visual discomfort

11.3.1 Influence factors of eye blinking

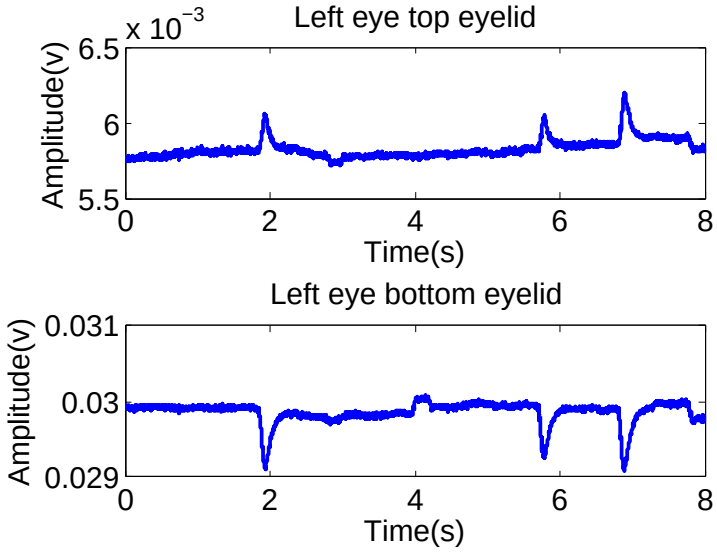
The EMG signals of the first second and the last second were removed in order to avoid transient effects. The duration of the signals in the analysis was 8 seconds.

Eye blinking is easy to detect from the raw EMG signal data according to some criteria. For example, the average length of a blink is 100-400 milliseconds. The amplitude of the blinking signal is larger than other EMG signal. The same position of the left and right eyes will generate similar responses for eye-blinking. For the position of the top and bottom eyelid, they always generate opposite responses on eye-blinking for the same eye. According to the signals from 8 positions, the number of blinks in 8 seconds for all stimuli were counted by manually inspecting the captured signal. Some examples of the EMG signals from the top eyelid and bottom eyelid of both eyes are shown in Figure 11.1.

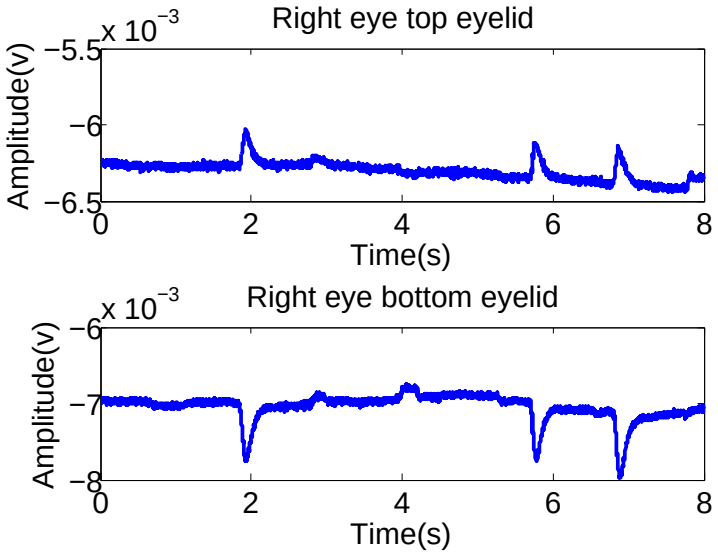
The average blinking rate for each stimulus is obtained by averaging all observers' data. It should be noted that the obtained eye blinking rate may be influenced by the electrodes around the eyes. Thus, the eye blinking rate in this chapter is not an absolute value. However, in this experiment, due to the fact that all of the data were influenced by the electrodes, these values can be used to make a comparative analysis on the relationship between the eye blinking rate, the 3D video characteristics, and visual discomfort.

The N-way ANOVA test was conducted on the mean blinking rate to test the main factors on blinks for each stimulus condition. The results showed that only velocity was the main factor in both the planar motion stimuli and the in-depth motion stimuli, with p-value of 0.005 and 0.0222. The disparity offset for static stimuli, planar motion stimuli and in-depth motion stimuli as well as the disparity amplitude for the in-depth motion stimuli did not have significant influence on eye blinks. The Multiple Comparison test was conducted based on the N-way ANOVA results. The results are shown in Figure 11.2. For planar motion and in-depth motion stimuli, only the velocity levels between slow and fast have a significant difference. The results indicated that the performance of eye blinks was affected significantly and differently by different video stimuli. Blinking rate increased with velocity when watching in-depth motion stimuli. However, it decreased with increasing velocity when watching planar motion stimuli. Though other factors were tested as not having significant influence on eye blinks, there was a trend of blinking rate with the increase of disparity offset, and this trend was different for different stimuli. For static and in-depth motion stimuli, the blinking rate increased with the disparity offset. However, for the planar motion stimuli, the blinking rate decreased with increasing disparity offset.

As shown in the Figure 11.2, the relative disparity between the foreground and the background plays a more important role in eye blinks than the absolute disparity, which shows a strong link with our previous study [87]. With the increase of the relative disparity, the eye blinking rate increases as well for the static and in-depth motion condition. But for the planar motion condition, the results are opposite.



(a)



(b)

Figure 11.1: Examples of the raw EMG signal for the left and right eye at the position of the top and bottom eyelid. Three eye blinks are detected in this example.

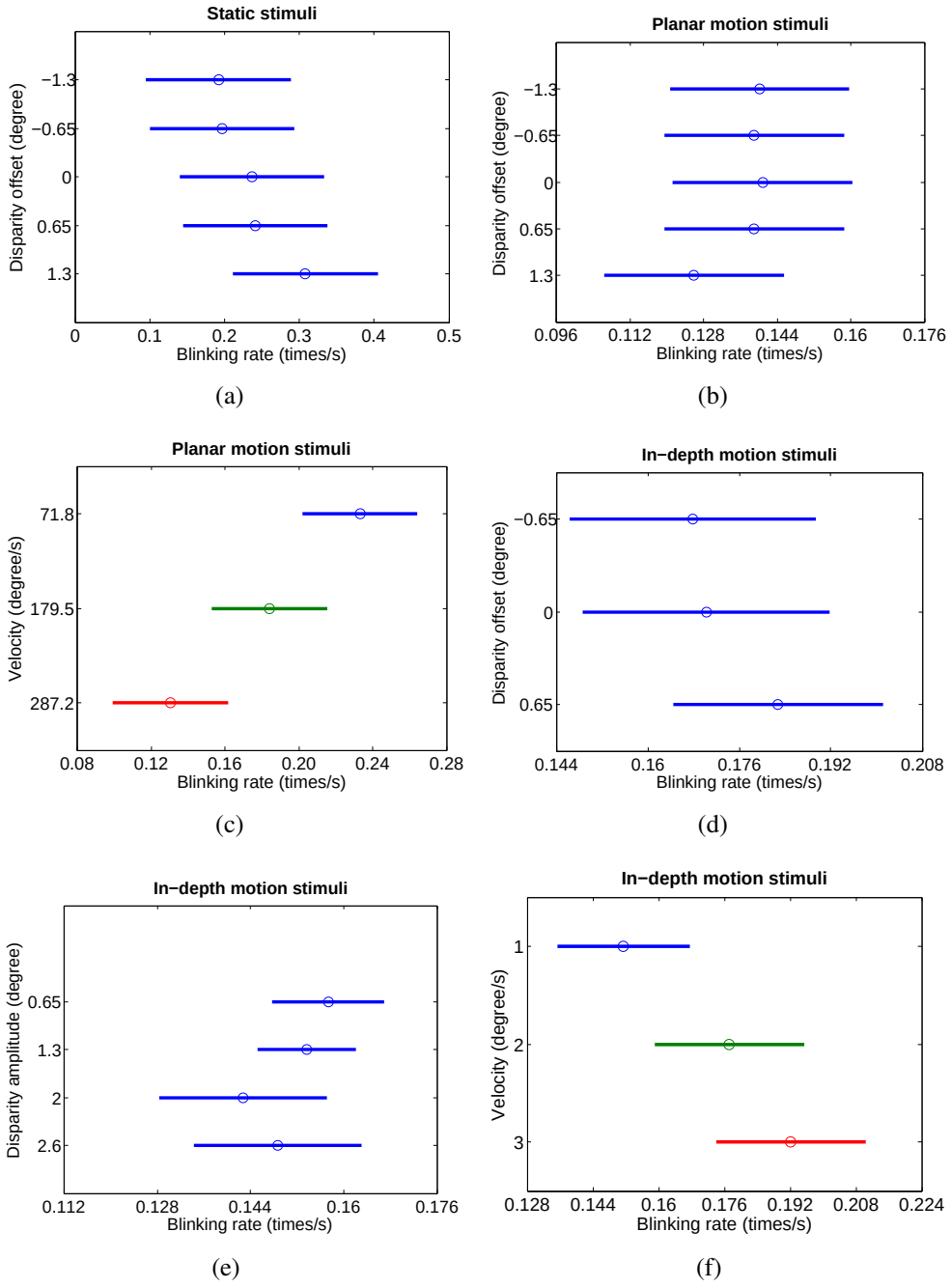


Figure 11.2: The Multiple Comparison test results for different factor levels. The mean value and the 95% confidence interval for each level of the factor are provided. (a) is the comparison of disparity offset levels for static stimuli. (b)-(c) are comparisons of disparity offset and velocity for planar motion stimuli, respectively. (d)-(f) are comparisons on disparity offset, disparity amplitude and velocity levels for in-depth motion stimuli, respectively.

Table 11.2: The linear regression results for different motion types

Type	Objective model	RMSE	R^2
Static	0.1752+0.0426r_o	0.0187	0.8792
Planar	0.2834-0.0110 r_o-0.0005v_p	0.0302	0.7177
In-depth	0.1345+0.0155r_o-0.0116d_a+0.0184v_d	0.0258	0.3751

11.3.2 Objective eye blinking models in function of 3D video characteristics

According to the results above, the relationship between eye blinking rate and relative disparity and velocity was nearly linear, thus, linear regression was used here to generate the objective models for different type of motion stimuli. The regression results are shown in Table 11.2. r_o represents the relative disparity, d_a represents the disparity amplitude, v_p and v_d are velocities for planar and in-depth motion.

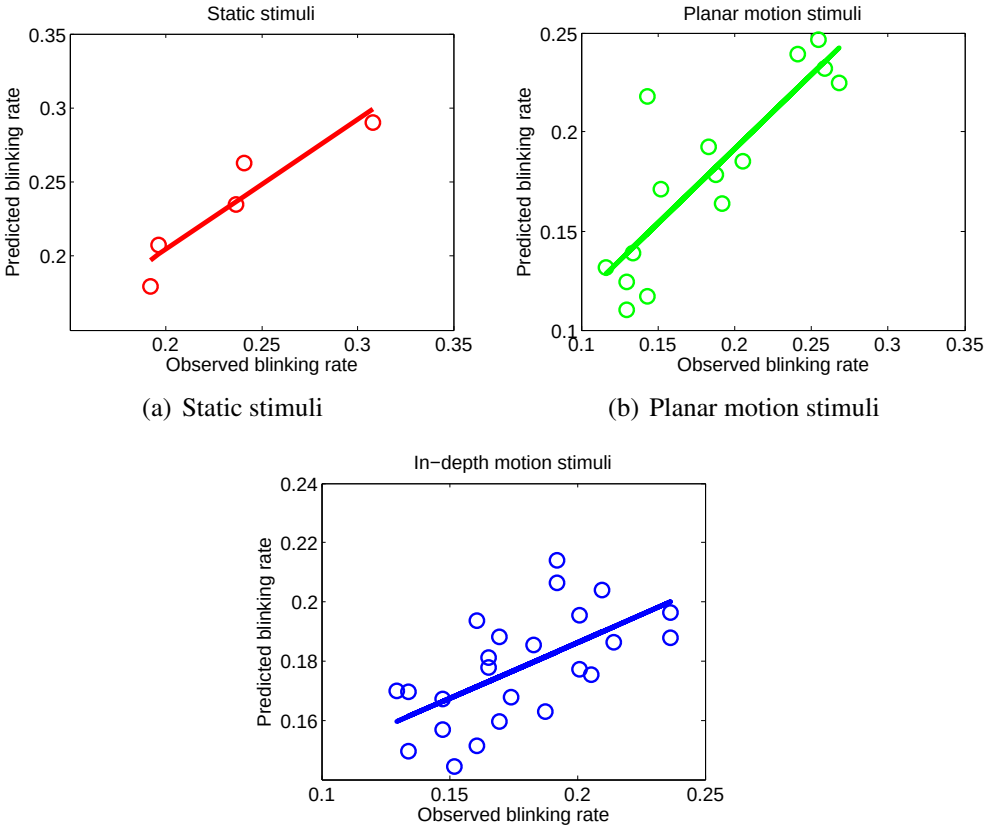
As shown in the objective models, for the static and the in-depth motion stimuli, the relative disparity offset is proportional to eye blinks, i.e., eye blinks increases with the relative disparity. For the planar motion stimuli, the relative disparity is inversely proportional to eye blinking rate. The velocity of the planar motion stimuli is inversely proportional to eye blinking rate while vice versa for the in-depth motion stimuli.

The Root Mean Square Error (RMSE) and R^2 for the observed eye blinking rate and the predicted value are shown in the table as well. The scatter plot of the observed value and the predicted value are shown in Figure 11.3. As shown in the results, generally, this model can predict the eye blinking reasonably well, especially in static and planar motion conditions.

11.3.3 The link between blinking rate and visual discomfort

The visual discomfort score for the 36 stimuli (Stimuli 1-36) had been previously obtained in Chapter 8 and shown in Table 8.2. The BT scores represent the degree of visual discomfort. The higher the value, the higher the visual discomfort degree. The BT scores are considered as the ground truth of visual discomfort in this study. Figure 11.4 shows the scatter plot of the mean eye-blinking rate and the visual discomfort score in each type of motion stimuli. The PLCC between eye blinking rate and visual discomfort are 0.9888, -0.8199 and 0.5347 for static, planar motion and in-depth motion stimuli, respectively.

As shown in Figure 11.4, the visual discomfort has a linear relationship with eye blinking rate. The linear relation to in-depth motion stimuli is less evident as in the static and the planar motion situation. The results indicated that when

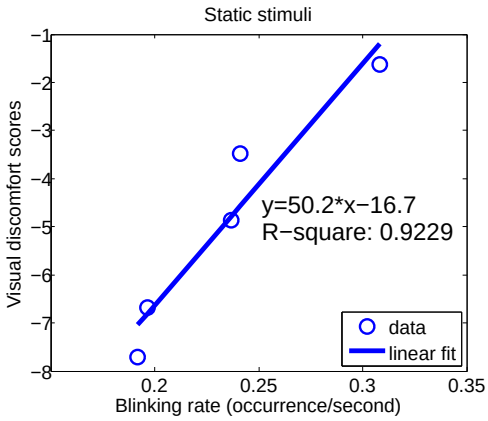


(a) Static stimuli

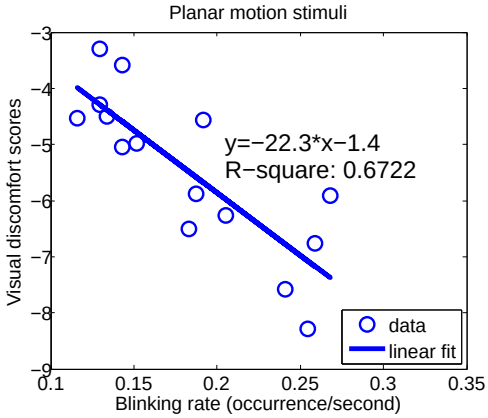
(b) Planar motion stimuli

(c) In-depth motion stimuli

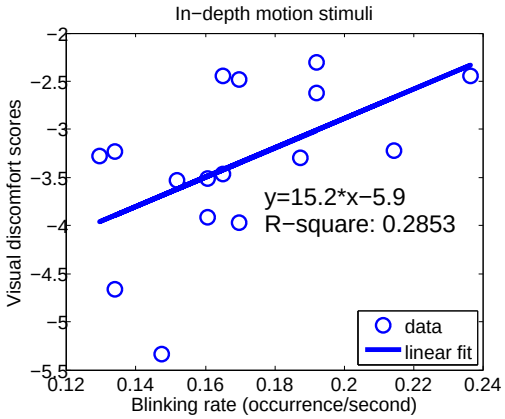
Figure 11.3: The scatter plot of the true blinking rate and the predicted blinking rate for all conditions.



(a)



(b)



(c)

Figure 11.4: The linear correlation of visual discomfort and eye-blinking rate for static stimuli, planar motion stimuli and in-depth motion stimuli. The x-axis represents the blinking rate. The y-axis represents the visual discomfort degree, higher scores represent more visual discomfort.

watching a still stereoscopic image or 3D video with in-depth motion, the blinking rate increased with the visual discomfort. However, when watching a 3D video with only planar motion, the blinking rate decreased with the visual discomfort.

11.4 Conclusions

In this study, the relationship between eye blinking rate, 3D video characteristics and visual discomfort are investigated. The eye blinking signals were extracted from the EMG signal which was obtained by an electro-physiological measurement device. The N-way ANOVA test results showed that velocity in 3D videos was a main factor for eye blinking. Its effect on eye blinks was significantly different for the planar motion stimuli and the in-depth motion stimuli. Eye blink frequency decreased with increasing motion velocity for the planar motion stimuli while it increased for the in-depth motion stimuli. An objective eye blinking prediction model for 3D stimuli was developed which showed linear relationship between 3D video characteristics and eye blinking rate.

It was also shown that the relationship between eye-blinking rate and visual discomfort was nearly linear. For the static and in-depth motion stimuli, the frequency of eye blinks increased with visual discomfort. However, for the planar motion stimuli, the blinking rate decreased with increasing visual discomfort. It seems that the blinking mechanisms for planar motion and in-depth motion stimuli are different. Further psychophysical studies are needed.

12

Conclusion and perspectives

12.1 Summary and contribution

In this thesis, we presented our work on the methods of subjective assessment and objective prediction on QoE, preference and visual discomfort in 3DTV. Two main goals were included. One was to develop a subjective methodology which could assess the multi-dimensional concept reliably and effectively. The other was to improve the viewing experience in 3DTV by means of decreasing the perceived visual discomfort.

Part I of this thesis focused on the subjective assessment methodology, including the analysis, evaluation and application of the proposed efficient paired comparison designs. The contributions are listed as follows.

Boosting paired comparison methodology

It is usually unfeasible to conduct a full paired comparison test with a typical number of test stimuli in quality assessment field. Thus, an efficient and effective design for paired comparison is necessary. In this thesis, we proposed a design to select the pairs based on the rank ordering of the stimuli and a spiral-based position arrangement in a matrix. The proposed design could avoid the bias effects from the presentation of the stimuli and meanwhile, it could largely reduce the time complexity from N^2 for the full paired comparison (FPC) to $N\sqrt{N}$. For instance, for a subjective experiment with 36 images, the number of pairs for each observer using the FPC method is 630 while using the proposed design it is 180. The test duration is shortened approximately 70%. In addition, experiments have shown that with the

same number of trials and under the condition that there were observation errors in the subjective experiments, the proposed design would generate more accurate results than the FPC method. For the proposed design (e.g., ASD), the RMSE between the estimates and the ground truth value is decreased approximately 10% compared with the FPC method, while for the existing designs, e.g., balanced sub-set design [32] and the sorting algorithm based design [113], the RMSE are increased 10% and 39% respectively. Thus, the proposed design not only boosts the paired comparison subjective experiment but also improves the accuracy of the estimates (Chapter 4, 5).

Novel practices for statistical analysis of paired comparison results

The traditional analysis tools for paired comparison data in our community are Thurstone-Mosteller (TM) model and Bradley-Terry (BT) model, which are used to convert the paired comparison binary data to scale values for all stimuli. However, when the model fails in model fit, i.e., the converted scale value could not be used to explain the raw data, it is necessary to use some other tools to analyze the data. This thesis provides the readers with some novel analysis methods which have been seldom used in this community, for example, (1) the Barnard's exact test and the Fisher's exact test which are used to test the significance of two proportional values with small sample sizes, e.g., whether 10/20 is significantly different from 8/22; (2) the Monte-Carlo significant test which is proposed to test the significant factors; and (3) the EBA model, which has the same function as TM and BT model but with more attributes being considered. These methods complement the statistical tools to analyze the paired comparison data (Chapter 4, 6).

Identification of the influence of different display technologies on viewing preference in 3DTV broadcasting contents

Different 3D display technologies would generate different viewing experience. In this thesis, the influence of the shutter glasses display and the polarized display was studied in subjective paired comparison experiments in terms of the preference of 3DTV broadcasting video contents. One important conclusion is that in the condition of high quality video sequences, the preference of viewing 2D format over 3D format in polarized displays is much higher than in shutter glasses displays. This result provides an important hint to the industry that 3D effect does not necessarily mean higher preference of the viewers when compared to 2D. The display technology used should be taken into account. Furthermore, the bandwidth of the broadcasting chain might benefit from this conclusion by providing 2D format rather than 3D format in certain cases(Chapter 6).

Side effects: hints on influence factors in 3DTV

The proposed efficient paired comparison design was employed to evaluate the possible factors that would influence the QoE. Many factors were considered in the experiments at the same time, including the system influence factor, context influence factor and human factors. As these influence factors were interacted and could not be independently analyzed in this thesis, the results from the experiments provide a hint on the influence factors rather than a conclusion. For example, the results in this thesis showed that test environment may not have significant influence on QoE; gender and viewing experience of the observers may be significant influence factors, etc. These results are important for the researchers who are interested in QoE. For instance, to design a subjective experiment, the researcher may need to pay more attention on the balance of the observer's gender and viewing experience distribution rather than the viewing environment(Chapter 6).

Part II of this thesis focused on the study on visual discomfort, including subjective evaluation, objective modeling and psychophysical prediction. The main contributions are listed as follows.

A psychophysical visual discomfort model considering 3D motion in 3D videos

3D motion (including static condition, planar motion and in-depth motion) in stereoscopic video sequences is considered as the main cause of visual discomfort. However, the three components of 3D motion were usually independently studied in literatures. There is no study to quantitatively compare the influences of different types of motion on visual discomfort. Based on this issue, we proposed a psychophysical visual discomfort model in function of 3D motion by utilizing the synthetic stimuli. The proposed model was then evaluated on natural stereoscopic video sequences. Two frameworks for the model were proposed, one is based on the tracked moving objects across the whole video sequence; the other is based on the moving objects in each frame of the video sequence. The test results showed that the performances of the two frameworks are comparable (CC=0.84 and CC=0.85 with MOS). Furthermore, both frameworks showed higher correlation with the subjective results than the existing model [65](CC=0.81 with MOS). Due to the lack of the stereoscopic video databases, the proposed model was only verified in one database. Further validation of the proposed model is necessary (Chapter 8, 10).

Comparative study on Paired Comparison (PC) and Absolute Category Rating (ACR) methodology on visual discomfort

Studies on visual discomfort in 3DTV were usually conducted based on subjective experiments with different methodologies. Thus, it would be interesting to know the influence of test methodologies on the final results. In this thesis, the PC method and the Absolute ACR method were compared in the context of visual discomfort induced by natural 3D video sequences. The results showed that the discriminability of the PC method is approximately 3 times higher than the ACR method. In addition, viewers' behavior might vary with the test methodology. For instance, in this thesis, the viewers in Paired Comparison test would pay more attention on the window violation of the video sequences, in particular, when in a pair one stimulus contains window violation but the other does not, viewers may judge the window violation as a significant factor to induce perceived differences on visual discomfort. In the ACR test, there was no such phenomenon observed. Thus, this conclusion indicates that when draw a conclusion from a subjective study, more attention should be paid on the subjective test methodology as it may have affected the results (Chapter 9).

Eye blinking rate is not always proportional to visual discomfort

Eye blinking rate is a widely used indicator for visual discomfort. In this thesis, we analyzed the relationship between 3D video characteristics (disparity, motion type, velocity, etc.), visual discomfort and eye blinking rate. It was shown that eye blinking rate is indeed an indicator for visual discomfort, however, it is not always proportional to the degree of visual discomfort. For static and in-depth motion stimuli, eye blinks increase with the degree of visual discomfort. However, for the planar motion stimuli, eye blinks decrease with increasing degree of visual discomfort. It seems that the blinking mechanisms for planar motion and in-depth motion stimuli are different. Thus, in stereoscopic 3DTV, it should be careful to directly use eye blinking rate as an indicator for visual discomfort. The characteristics of the video content should be taken into account as well (Chapter 11).

12.2 Limitation and perspectives

Considering the objectives of this thesis, there are still some studies need further investigation:

– The proposed design for selecting pairs in Paired Comparison test in this thesis is based on a spiral arrangement in a matrix. The advantage of the spiral

arrangement is that it is easy to understand and implement for the requirement that closer pairs would generate more precise results than distant pairs. Furthermore, it is replicable in other tests. However, there might be a mathematical solution for the arrangement of the matrix which could generate the results with the minimum estimation errors.

– The influence of Paired Comparison and ACR test methodologies on final results needs further study. In this thesis, the two methodologies were employed in two labs individually. There are some other possible factors that may affect the results, for example, the observer's culture (Korean and French), the observer's viewing experience, and the 3D displays used in the two labs. To draw a general conclusion on the differences of the performances between the two test methodologies, more experiments should be conducted.

– Due to the lack of stereoscopic video databases on visual discomfort, the proposed psychophysical model in this thesis was only trained and verified by one database. In the future, an effort should be made on the construction of a variety of databases. The proposed model would then be evaluated on other databases.

It should be noted that even though the proposed efficient paired comparison design, the novel statistical analysis tools, the influence of test methodologies, and the psychophysical prediction method introduced in this thesis were serving for the QoE in 3DTV, the goals of this thesis are far beyond. The new multimedia technology, e.g., Ultra High Definition television, might face similar issues as in 3DTV, for instance, multi-dimensional viewing experience or visual discomfort. Thus, the research work conducted in this thesis is not limited to 3DTV but open to any new technologies in multimedia.

List of Tables

List of Figures

Abbreviations

ACR	Absolute Category Rating
ANOVA	ANalysis Of VAriance
ARD	Adaptive Rectangular Design
ASD	Adaptive Square Design
BT	Bradley-Terry
DOF	Depth of Focus
DSCQS	Double Stimulus Continuous Quality Scale
DSI	Disparity Spatial Information
DTI	Disparity Temporal Information
DVB	Digital Video Broadcasting
EBA	Elimination by Aspects
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
FCC	Frame Compatible Compatible
FPC	Full Paired Comparison
fMRI	Functional magnetic resonance imaging
GDD	Group Divisible Design
GLM	Generalized Linear Model
HRC	Hypothetical Reference Circuit
HRRG	HodgeRank on Random Graphs
IF	Influence Factor
LOOCV	Leave-one-out Cross Validation
MOS	Mean Opinion Score

ORD	Optimized Rectangular Design
OSD	Optimized Square Design
PC	Paired Comparison
PLCC	Pearson Linear Correlation Coefficient
PoE	Preference of Experience
QoE	Quality of Experience
RD	Rectangular Design
RMSE	Root Mean Square Error
SROCC	Spearman's Rank-Order Correlation Coefficient
R/PC	Randomised Paired Comparison
SA	Sorting Algorithm based design
SBS	Side by Side
SD	Square Design
SI	Spatial Information
SRC	Source Reference Circuit
SS	Single Stimulus
SSCQE	Single Stimulus Continuous Quality Evaluation
SSQ	Simulator sickness questionnaire
STB	Set-top box
TD	Triangular Design
TI	Temporal Information
TM	Thurstone-Mosteller
VA	Vergence-Accommodation
VDU	Visual Display Unit
VQEG	Video Quality Experts Group

Publications

Journal

1. **Jing Li**, Jesús Gutiérrez, Romain Cousseau, Marcus Barkowsky, Fernando Jaureguizar, Julián Cabrera, Narciso García and Patrick Le Callet, “Evaluation Study of Preference of Experience in 3DTV: Influence of System, Context and Human Factors”, submitted to IEEE Trans. on Broadcasting.
2. **Jing Li**, Marcus Barkowsky, Patrick Le Callet, “Visual discomfort of stereoscopic videos: influence of motion”, Displays (accepted).
3. **Jing Li**, Marcus Barkowsky, Patrick Le Callet, “Recent Advances in Standardization on 3D Quality of Experience”, IEEE COMSOCMMTC E-Letter, vol.8, no. 3, May 2013.

Conference

1. **Jing Li**, Ondrej Kaller, Francesca De Simone, Jussi Hakala, Dawid Juszka, Patrick Le Callet, “Cross-lab study on Preference of Experiencein 3DTV: Influence from display technology and test environment”, QoMEX, IEEE, 2013.
2. **Jing Li**, Marcus Barkowsky, Patrick Le Callet, “Subjective assessment methodology for Preference of Experience in 3DTV”, IEEE IVMSP, 2013.
3. **Jing Li**, Marcus Barkowsky, Patrick Le Callet, “Boosting Paired Comparison methodology in measuring visual discomfort of 3DTV: performances of three di erent designs”, Proceedings of the SPIE Electronic Imaging, Stereoscopic Displays and Applications, 2013.

4. Marcus Barkowsky, **Jing Li**, Taehwan Han, Sungwook Youn, Jiheon Ok, Chulhee Lee, Christer Hedberg, Inirajith V. Ananth, Kun Wang, Kjell Brunnström, Patrick Le Callet, "Towards standardized 3DTV QoE assessment: Cross-lab study on display technology and viewingenvironment parameters", SPIE Electronic Imaging, 864809-864809-7, 2013.
5. **Jing Li**, Marcus Barkowsky, Patrick Le Callet, "Visual discomfort is not always proportional to eye blinking rate: exploring some effects of planar and in-depth motion on 3D QoE", VPQM, 2013.
6. **Jing Li**, Marcus Barkowsky, Patrick Le Callet, "Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment", ICIP, 2012, Orlando, U.S.A.
7. **Jing Li**, Marcus Barkowsky, Junle Wang and Patrick Le Callet, "Study on visual discomfort induced by stimulus movement at fixed depthon stereoscopic displays using shutter glasses", 17th International Conference on Digital Signal Processing, 2011.
8. **Jing Li**, Marcus Barkowsky, Patrick Le Callet, "The influence of relative disparity and planar motion velocity on visual discomfort ofstereoscopic videos", QoMEX, IEEE, 2011.
9. **Jing Li**, Marcus Barkowsky, Patrick Le Callet, "Visual discomfort induced by relative disparity and planar motion of stereoscopic images", the first Sino French Workshop on Information and Communication Technologies, France, 2011.

Book Chapter

1. **Jing Li**, Marcus Barkowsky, Patrick Le Callet, "Chapter 11: Assessing the Quality of Experience of 3DTV and beyond - Tackling the multidimensional sensation", in 3D Future Internet Media, Tasos Dagiuklas and Ahmet Kondoz, Springer, 2014.
2. Matthieu Urvoy, Marcus Barkowsky, **Jing Li**, patrick Le Callet, "Confort et fatigue visuels en stéréoscopie" dans *Vidéo 3D : Capture, traitement et diffusion*, Traité IC2, série Signal et image, Chapitre: 16, pp.309-327. Hermès Sciences Publications Lavoisier, Lucas, Laurent and Loscos, Céline and Rémion, Yannick. September 2013.

Project Report

1. Junle Wang, Emilie Bosc, **Jing Li**, Vincent Ricordel, "Livrable D1. 2 of the PERSEE project: Perceptual Modelling: Definition of the Models", 2011.

2. Junle Wang, Josselin Gautier, Emilie Bosc, **Jing Li**, Vincent Ricordel, “Livrable D6. 1 of the PERSEE project: Perceptual Assessment: Definition of the scenarios”, 2011.

Bibliography

[1] P3333.1 - Standard for the Quality Assessment of Three Dimensional (3D) Contents based on Psychophysical Studies. 2

[2] ATEME. Satellite bit rate recommendation. 2010. 84

[3] BT. Backus, DJ. Fleet, AJ. Parker, and DJ. Heeger. Human cortical activity correlates with stereoscopic depth perception. *Journal of Neurophysiology*, 86(4):2054–2068, 2001. 119

[4] M. Barkowsky, R. Cousseau, and P. Le Callet. Is visual fatigue changing the perceived depth accuracy on an autostereoscopic display? In *IS&T/SPIE Electronic Imaging*, pages 78631V–78631V. International Society for Optics and Photonics, 2011. 17

[5] M. Barkowsky, J. Li, T. Han, S. Youn, J. Ok, C. Lee, C. Hedberg, I.V. Ananth, K. Wang, and K. Brunnström. Towards standardized 3DTV QoE assessment: Cross-lab study on display technology and viewing environment parameters. In *IS&T/SPIE Electronic Imaging*, pages 864809–864809. International Society for Optics and Photonics, 2013. 24, 25, 26, 104, 199

[6] G. Barnard. A new test for 2×2 tables. *Nature*, 156:177, 1945. 67

[7] E. Bizzi. Discharge of frontal eye field neurons during saccadic and following eye movements in unanesthetized monkeys. *Experimental Brain Research*, 6(1):69–80, 1968. 180

[8] J-L. Blin. SAMVIQ–Subjective assessment methodology for video quality. *Rapport technique BPN*, 56, 2003. 24

[9] 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 as-

sessment. *Selected Topics in Signal Processing, IEEE Journal of*, 5(7):1332–1343, 2011. 30

[10] R.A. Bradley. 14 paired comparisons: Some basic procedures and examples. *Handbook of Statistics*, 4:299–326, 1984. 43, 47, 77, 82, 90, 130

[11] R.A. Bradley and M.E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, Dec. 1952. 43, 47, 130

[12] D.H. Brainard. The psychophysics toolbox. *Spatial vision*, 10(4):433–436, 1997. 125

[13] T. C. Brown and G. L. Peterson. *An enquiry into the method of paired comparison: reliability, scaling, and Thurstone's Law of Comparative Judgment*. US Department of Agriculture, Forest Service, Rocky Mountain Research Station, 2009. 64

[14] T. J Buschman and E. K Miller. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *science*, 315(5820):1860–1862, 2007. 180

[15] B. Cassin, S. Solomon, and M. L Rubin. *Dictionary of eye terminology*. Wiley Online Library, 1990. 12

[16] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, May 2011. 83, 167

[17] L. Chen, Y. Tu, W. Liu, Q. Li, K. Teunissen, and I. Heynderickx. 73.4: Investigation of Crosstalk in a 2-View 3D Display. In *SID Symposium Digest of Technical Papers*, volume 39, pages 1138–1141. Wiley Online Library, 2008. 116

[18] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet. New requirements of subjective video quality assessment methodologies for 3DTV. *International Workshop on Video Processing and Quality Metrics*, Jan. 2010. 20, 88, 111

[19] W. Chen, J. Fournier, 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*, pages 1–6, Jan. 2012. 16

[20] 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. 17, 152

[21] S-H. Cho and H-B. Kang. An Assessment of Visual Discomfort Caused by Motion-in-Depth in Stereoscopic 3D Video. In *Proceedings of the British*

Machine Vision Conference, pages 65.1–65.10. BMVA Press, 2012. 116, 121

[22] J. Choi, D. Kim, S. Choi, and K. Sohn. Visual fatigue modeling and analysis for stereoscopic video. *Optical Engineering*, 51(1):017206–1, 2012. 156, 157

[23] J. Choi, D. Kim, B. Ham, S. Choi, and K. Sohn. Visual fatigue evaluation and enhancement for 2d-plus-depth video. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 2981–2984. IEEE, 2010. 157

[24] M. Clerc. *Particle swarm optimization*, volume 243. Iste London, 2006. 169

[25] L.E. Coria, D. Xu, and P. Nasiopoulos. Quality of Experience of stereoscopic content on displays of different sizes: A comprehensive subjective evaluation. *IEEE International Conference on Consumer Electronics*, pages 755–756, Jan. 2011. 103, 202

[26] BG. Cumming and GC. DeAngelis. The physiology of stereopsis. *Annual Review of Neuroscience*, 24(1):203–238, Mar. 2001. 13

[27] J.E. Cutting. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. *W. Epstein*, 6:69–117, 1995. 17

[28] S. Delis, N. Nikolaidis, and Pitas I. Automatic Detection of Depth Jump Cuts and Bent Window Effects in Stereoscopic Videos. *11th IEEE IVMSP Workshop : 3D Image/Video Technologies and Applications*, 2013. 164, 173

[29] G. Devroede. Constipation–a sign of a disease to be treated surgically, or a symptom to be deciphered as nonverbal communication? *Journal of clinical gastroenterology*, 15(3):189, 1992. 109

[30] M. Divjak and H. Bischof. Eye blink based fatigue detection for prevention of computer vision syndrome. In *IAPR Conference on Machine Vision Applications, Tokyo*, 2009. 120, 180

[31] N. R. Draper, H.y Smith, and E. Pownell. *Applied regression analysis*, volume 3. Wiley New York, 2nd edition, 1981. 134

[32] O. Dykstra. Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetitions on pairs. *Biometrics*, 16(2):176–188, Jun. 1960. 34, 36, 50, 55, 190

[33] A. Eichhorn, P. Ni, and R. Eg. Randomised pair comparison: an economic and robust method for audiovisual quality assessment. In *Proceedings of the 20th international workshop on Network and operating systems support for digital audio and video*, pages 63–68. ACM, 2010. 34

[34] M. Emoto, T. Niida, and F. Okano. Repeated vergence adaptation causes the decline of visual functions in watching stereoscopic television. *Journal of Display Technology*, 1(2):328–340, 2005. 119

[35] 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*, pages 183–188, Sep. 2011. 29, 30, 200

[36] EF. Fincham and J. Walton. The reciprocal actions of accommodation and convergence. *The Journal of Physiology*, 137(3):488–508, 1957. 18

[37] RA. Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society*, 98(1):39–82, 1935. 67

[38] T. Fukushima, M. Torii, K. Ukai, J.S. Wolffsohn, and B. Gilmartin. The relationship between CA/C ratio and individual differences in dynamic accommodative responses while viewing stereoscopic images. *Journal of Vision*, 9(13), 2009. 125

[39] J.A. Gaspar. Serra da estrela (portugal). 2007. 12, 199

[40] S. Georgieva, R. Peeters, H. Kolster, J.T. Todd, and G.A. Orban. The processing of three-dimensional shape from disparity in the human brain. *The Journal of Neuroscience*, 29(3):727–742, 2009. 119

[41] M.E. Glickman and S.T. Jensen. Adaptive paired comparison design. *Journal of statistical planning and inference*, 127(1-2):279–293, 2005. 34

[42] J. Gutiérrez, P. Perez, F. Jaureguizar, J. Cabrera, and N. Garcia. Validation of a novel approach to subjective quality evaluation of conventional and 3d broadcasted video services. In *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, pages 230–235. IEEE, 2012. 24

[43] MG. Habib and DR. Thomas. Chi-square goodness-if-fit tests for randomly censored data. *The Annals of Statistics*, 14(2):759–765, 1986. 44

[44] A. Hanazato, M. Okui, and I. Yuyama. Subjective evaluation of cross talk disturbance in stereoscopic displays. In *SDI 20th Int. Display Res. Conf*, pages 288–291, 2000. 116

[45] J.C. Handley. Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment. *Proc. IS&T Image Processing, Image Quality, Image Capture, Systems Conference*, 4:108–112, Apr. 2001. 42, 90

[46] D.S. Hands, M.D. Brotherton, A. Bourret, and D. Bayart. Subjective quality assessment for objective quality model development. *Electronics Letters*, 41(7):408–409, 2005. 21

[47] J.M. Harris, S.P. McKee, and S.N.J. Watamaniuk. Visual search for motion-in-depth: Stereomotion does not pop out from disparity noise. *Nature neuroscience*, 1(2):165–168, 1998. 116

[48] L.R. Harris and M.R.M. Jenkin. *Vision in 3D environments*. Cambridge University Press, 2011. 116

[49] I.E. Heynderickx and S. Bech. Image quality assessment by expert and non-expert viewers. In *Electronic Imaging 2002*, pages 129–137. International Society for Optics and Photonics, 2002. 21

[50] D.M. Hoffman, A.R. Girshick, K. Akeley, and M.S. Banks. Vergence–Accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision*, 8(3):1–30, Mar. 2008. 110

[51] N.S. Holliman. 3D display systems. *Science*, 38(8):31–36, 2010. 14

[52] J.R. Hughes. Gamma, fast, and ultrafast waves of the brain: their relationships with epilepsy and behavior. *Epilepsy & Behavior*, 13(1):25–31, 2008. 119

[53] M. Hyder, K.R. Laghari, N. Crespi, M. Haun, and C. Hoene. Are QoE requirements for multimedia services different for men and women? Analysis of gender differences in forming QoE in virtual acoustic environments. *Emerging Trends and Applications in Information Communication Technologies*, 281:200–209, Mar. 2012. 21, 102

[54] 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*, pages 38–45. International Society for Optics and Photonics, 2002. 111, 158

[55] W.A. Ijsselsteijn, H. de Ridder, and J. Vliegen. Subjective evaluation of stereoscopic images: Effects of camera parameters and display duration. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(2):225–233, Mar. 2000. 114

[56] ITU-R BT.2021. Subjective methods for the assessment of stereoscopic 3DTV systems. *International Telecommunication Union, Geneva, Switzerland*, Aug. 2012. 2, 3, 24, 25, 31, 65, 117

[57] ITU-R BT.2160. Features of three-dimensional television video systems for broadcasting. *International Telecommunication Union, Geneva, Switzerland*, 2011. 114, 202

[58] ITU-R BT.500-13. Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union, Geneva, Switzerland*, Jan. 2012. 20, 21, 23, 72, 117, 125, 143, 145, 180

[59] ITU-T P.910. Subjective video quality assessment methods for multimedia applications. *International Telecommunication Union*, Apr. 2008. 23, 25, 31, 32, 83, 200

[60] IVY Lab stereoscopic video dataset. *Available: http://ivylab.kaist.ac.kr/demo/ivy3D-LocalMotion/index.htm*. 142, 144, 149, 196, 204

[61] T.Y. Jang. A study of visual comfort of 3d crosstalk and binocular color-rivalry thresholds for stereoscopic displays. Master's thesis, National Taiwan University of Science and Technology, 2012. 114

[62] J. A. John. Reduced group divisible paired comparison designs. *The Annals of Mathematical Statistics*, pages 1887–1893, 1967. 33

[63] S. Jumisko-Pyykkö. *User-Centered Quality of Experience and Its Evaluation Methods for Mobile Television*. PhD thesis, Tampere University of Technology, 2011. 19

[64] S. Jumisko-Pyykkö and T. Vainio. Framing the context of use for mobile HCI. *International Journal of Mobile Human Computer Interaction*, 2(4):1–28, 2010. 19

[65] Y.J. Jung, S. Lee, H. Sohn, H. W. Park, and Y.M. Ro. Visual comfort assessment metric based on salient object motion information in stereoscopic video. *Journal of Electronic Imaging*, 21(1):011008–1, 2012. 7, 141, 148, 156, 157, 170, 171, 174, 175, 191

[66] O. Kaller, L. Bolecek, and T. Kratochvil. Subjective evaluation and measurement of angular characteristics of the 3D stereoscopic displays. *International Conference Radioelektronika*, pages 1–4, Apr. 2012. 19

[67] R.G. Kaptein, A. Kuijsters, M.T.M. Lambooij, W.A. Ijsselsteijn, and I. Heynderickx. Performance evaluation of 3D-TV systems. *Image Quality and System Performance V*, 6808:1–11. 16

[68] R.S. Kennedy, N.E. Lane, K.S. Berbaum, and M.G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993. 117

[69] D. Kim, S. Choi, S. Park, and K. Sohn. Stereoscopic visual fatigue measurement based on fusional response curve and eye-blinks. In *Digital Signal Processing (DSP), 2011 17th International Conference on*, pages 1–6. IEEE, 2011. 120, 180

[70] D. Kim, Y.J. Jung, E. Kim, Y.M. Ro, and H.W. Park. Human brain response to visual fatigue caused by stereoscopic depth perception. In *Digital Signal*

Processing (DSP), 2011 17th International Conference on, pages 1–5. IEEE, 2011. 119, 179

[71] D. Kim and K. Sohn. Visual fatigue prediction for stereoscopic image. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(2):231–236, 2011. 114, 157

[72] 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):324–324, 2011. 110

[73] Y.J. Kim and E.C. Lee. EEG Based Comparative Measurement of Visual Fatigue Caused by 2D and 3D Displays. *HCI International 2011–Postersï¿½ Extended Abstracts*, pages 289–292, 2011. 119, 179

[74] L.S. King. *Medical thinking: A historical preface*. Princeton University Press Princeton, NJ, 1982. 110

[75] F.L. Kooi and A. Toet. Visual comfort of binocular and 3D displays. *Displays*, 25(2-3):99–108, 2004. 111, 113, 116

[76] J. Kuze and K. Ukai. Subjective evaluation of visual fatigue caused by motion images. *Displays*, 29(2):159–166, 2008. 111, 118

[77] M. Lambooij, M.F. Fortuin, W.A. Ijsselsteijn, and I. Heynderickx. Visual discomfort associated with 3D displays. In *5th Int. Workshop Video Process. Quality Metrics for Consum. Electron*, 2010. 71

[78] M. Lambooij, W. IJsselsteijn, D.G. Bouwhuis, and I. Heynderickx. Evaluation of stereoscopic images: beyond 2d quality. *Broadcasting, IEEE Transactions on*, 57(2):432–444, 2011. 16

[79] M. Lambooij, WA. IJsselsteijn, and I. Heynderickx. Visual discomfort of 3D TV: Assessment methods and modeling. *Displays*, 32(4):209–218, 2011. 155, 157

[80] P. Le Callet, S. Möller, and A. Perkis. Qualinet white paper on definitions of quality of experience v.1.1. *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*, Jun. 2012. 15, 19

[81] P. Lebreton, A. Raake, M. Barkowsky, and P. Le Callet. A subjective evaluation of 3D IPTV broadcasting implementations considering coding and transmission degradation. *IEEE International Symposium on Multimedia*, pages 506–511, Dec. 2011. 19

[82] E.C. Lee, H. Heo, and K.R. Park. The comparative measurements of eyestrain caused by 2d and 3d displays. *Consumer Electronics, IEEE Transactions on*, 56(3):1677–1683, 2010. 120, 180

[83] J-S. Lee, L. Goldmann, and T. Ebrahimi. Paired comparison-based subjective quality assessment of stereoscopic images. *Multimedia Tools and Applications*, pages 1–18, Feb. 2012. 30, 31, 152, 200

[84] S. Lee, Y.J. Jung, H. Sohn, Y.M. Ro, and H.W. Park. Visual discomfort induced by fast salient object motion in stereoscopic video. In *Proceedings of SPIE*, volume 7863, page 786305, 2011. 71, 116, 121, 122

[85] Seong-il Lee, Yong Ju Jung, Hosik Sohn, and Yong Man Ro. Subjective assessment of visual discomfort induced by binocular disparity and stimulus width in stereoscopic image. In *IS&T/SPIE Electronic Imaging*, pages 86481T–86481T. International Society for Optics and Photonics, 2013. 157

[86] H-C. Li, J. Seo, K. Kham, and S. Lee. Measurement of 3d visual fatigue using event-related potential (erp): 3d oddball paradigm. In *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 2008*, pages 213–216. IEEE, 2008. 119

[87] 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*, pages 155–160, Sep. 2011. 71, 122, 182

[88] 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)*, pages 1–8. IEEE, 2011. 71, 122

[89] J. Li, O. Kaller, F. De Simone, J. Hakala, D. Juszka, and P. Le Callet. Cross-lab study on Preference of Experience in 3DTV: influence from display technology and test environment. *International Workshop on Quality of Multimedia Experience*, pages 1–2, 2013. 104

[90] A. Martín Andrés, M.J. Sánchez Quevedo, and A. Silva Mato. Asymptotical tests in 2 $\times$ 2 comparative trials (unconditional approach). *Computational Statistics & Data Analysis*, 40(2):339–354, Aug. 2002. 67

[91] G. Mather. *Foundations of sensation and perception*, volume 2. Psychology Press, 2009. 10

[92] LN. McLin and CM. Schor. Voluntary effort as a stimulus to accommodation and vergence. *Investigative ophthalmology & visual science*, 29(11):1739, 1988. 125

[93] D.V. Mehrotra, I.SF Chan, and R.L. Berger. A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics*, 59(2):441–450, Jun. 2003. 67

[94] C. R Mehta and P. Senchaudhuri. Conditional versus unconditional exact tests for comparing two binomials. 2003. http://www.cytel.com/Papers/twobinomials.pdf. 67

[95] B. Mendiburu. *3D movie making: stereoscopic digital cinema from script to screen*. Focal Press, 2012. 111

[96] F. Mosteller. Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16(1):3–9, 1951. 42

[97] N.K. Nahar, J.E. Sheedy, J. Hayes, and Y.C. Tai. Objective measurements of lower-level visual stress. *Optometry & Vision Science*, 84(7):620, 2007. 120, 180

[98] B.H Natelson. *Facing and fighting fatigue: a practical approach*. Yale University Press, 1998. 110

[99] Y. Nojiri, H. Yamanoue, A. Hanazato, and F. Okano. Measurement of parallax distribution, and its application to the analysis of visual comfort for stereoscopic hdtv. In *Proc. SPIE*, volume 5006, pages 195–205, 2003. 111, 158

[100] Y. Nojiri, H. Yamanoue, S. Ide, S. Yano, and F. Okana. Parallax distribution and visual comfort on stereoscopic hdtv. In *Proc. IBC*, pages 373–380, 2006. 111, 155, 157

[101] L. OHare and P.B. Hibbard. Spatial frequency and visual discomfort. *Vision research*, 51(15):1767–1777, 2011. 125

[102] S. Ohno and K. Ukai. Subjective evaluation of motion sickness following game play with head mounted display. *The journal of the institute of image information and television engineers*, 54:887–891, 2000. 118

[103] S. Pastoor. Human factors of 3d imaging: results of recent research at heinrich-hertz-institut berlin. In *Proc. IDW*, volume 95, pages 69–72, 1995. 113, 116

[104] S. Pastoor and M. Wöpking. 3-D displays: A review of current technologies. *Displays*, 17(2):100–110, 1997. 14

[105] D.G. Pelli. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, 10(4):437–442, 1997. 72, 125

[106] M.H. Pinson, J. Lucjan, P. Romuald, H-T. Quan, S. Christian, C. Phillip, Y. Audrey, P. Le Callet, B. Marcus, and I. William. The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):640–651, October 2012. 20, 199

[107] M. Rangaswamy, B. Porjesz, D.B. Chorlian, K. Wang, K.A. Jones, L.O. Bauer, J. Rohrbaugh, S.J. OĆonnor, S. Kuperman, and T. Reich. Beta power in the EEG of alcoholics. *Biological psychiatry*, 52(8):831–842, 2002. 119

[108] ITUT Rec. P.10/G.100 Amendment 2: New definitions for inclusion in Recommendation ITU-T P.10/G.100. *International Telecommunication Union, Geneva, Switzerland*, 2008. 15

[109] F. Rumsey and S. Bech. On some biases encountered in modern audio quality listening tests-a review. *Journal of the Audio Engineering Society*, 56(6):427–451, 2008. 24, 25, 199

[110] M. Santoro, G. AlRegib, and Y. Altunbasak. Misalignment correction for depth estimation using stereoscopic 3-d cameras. In *Multimedia Signal Processing (MMSP), 2012 IEEE 14th International Workshop on*, pages 19–24. IEEE, 2012. 113, 202

[111] J.D. Schall. On the role of frontal eye field in guiding attention and saccades. *Vision research*, 44(12):1453–1467, 2004. 180

[112] PJH Seuntiëns, LMJ Meesters, and WA Ijsselsteijn. Perceptual attributes of crosstalk in 3D images. *Displays*, 26(4-5):177–183, 2005. 16, 116

[113] D. A. Silverstein and Farrell J. E. Quantifying perceptual image quality. *Proc. IS&T Image Processing, Image Quality, Image Capture, Systems Conference*, 1:242–246, May 1998. 34, 35, 55, 190

[114] M. Slanina, T. Kratochvil, V. Ricny, L. Bolecek, O. Kaller, and L. Polak. Testing QoE in different 3D HDTV technologies. *Radioengineering*, 21(1):445–454, Apr. 2012. 19

[115] H. Sohn, Y.J. Jung, S. Lee, H.W. Park, and Y.M. Ro. Attention model-based visual comfort assessment for stereoscopic depth perception. In *Digital Signal Processing (DSP), 2011 17th International Conference on*, pages 1–6. IEEE, 2011. 117

[116] H. Sohn, Y.J. Jung, S. Lee, H.W. Park, and Y.M. Ro. Investigation of object thickness for visual discomfort prediction in stereoscopic images. In *IS&T/SPIE Electronic Imaging*, pages 82880Q–82880Q. International Society for Optics and Photonics, 2012. 117, 157

[117] F. Speranza, F. Poulin, R. Renaud, M. Caron, and J. Dupras. Objective and subjective quality assessment with expert and non-expert viewers. In *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, pages 46–51. IEEE, 2010. 21, 102

[118] F. Speranza, W.J. 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, 6055:94–103, Jan. 2006. 71, 111, 116, 117, 121, 122, 135, 140, 141

[119] D. Strohmeier, S. Jumisko-Pyykkö, and U. Reiter. Profiling experienced quality factors of audiovisual 3D perception. *International Workshop on Quality of Multimedia Experience*, pages 70–75, Jun. 2010. 19

[120] A. Suzumura. Visual fatigue. *Ganka*, 23(8):799–804, 1981. 118

[121] W.J. Tam, F. Speranza, C. Vazquez, R. Renaud, and N. Hur. Visual comfort: stereoscopic objects moving in the horizontal and mid-sagittal planes. *Proc. SPIE 8288, Stereoscopic Displays and Applications XXIII*, 828813, 2012. 111, 116, 117, 121, 122

[122] W.J. Tam, F. Speranza, S. Yano, K. Shimono, and H. Ono. Stereoscopic 3d-tv: visual comfort. *Broadcasting, IEEE Transactions on*, 57(2):335–346, 2011. 111, 112, 202

[123] W.J. Tam and L.B. Stelmach. Display duration and stereoscopic depth discrimination. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 52(1):56, 1998. 24

[124] 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. 142

[125] M. Teplan. Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11, 2002. 119

[126] L.L. Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273–286, 1927. 39, 40

[127] L.L. Thurstone. Psychophysical analysis. *The American journal of psychology*, 38(3):368–389, 1927. 39

[128] R.F. Tiltman. How stereoscopic television is shown. *Radio News*, 10(418-419), Nov. 1928. 13

[129] R. BH Tootell, J.D Mendola, N.K. Hadjikhani, P.J. Ledden, A.K. Liu, J.B. Reppas, M.I. Sereno, and A. M. Dale. Functional analysis of v3a and related areas in human visual cortex. *The journal of Neuroscience*, 17(18):7060–7078, 1997. 119

[130] D.Y. Tsao, W. Vanduffel, Y. Sasaki, D. Fize, T. A. Knutsen, J. B. Mandeville, L.L. Wald, A.M. Dale, B.R. Rosen, D.C. Van Essen, and others. Stereopsis activates v3a and caudal intraparietal areas in macaques and humans. *Neuron*, 39(3):555–568, 2003. 119

[131] K. Tsubota and K. Nakamori. Dry eyes and video display terminals. *New England Journal of Medicine*, 328(8):584–584, 1993. 120, 180

[132] A. Tversky. Elimination by aspects: A theory of choice. *Psychological review*, 79(4):281–299, Jul. 1972. 99

[133] A. Tversky and S. Sattath. Preference trees. *Psychological Review*, 86(6):542–573, Nov. 1979. 99

[134] C. W. Tyler, L. T. Likova, K. Atanassov, V. Ramachandra, and S. Goma. 3d discomfort from vertical and torsional disparities in natural images. *Proc. 17th SPIE Human Vision Electron. Imaging*, 8291:82910Q–1, 2012. 113, 115, 203

[135] H. Urey, K.V. Chellappan, E. Erden, and P. Surman. State of the art in stereoscopic and autostereoscopic displays. *Proceedings of the IEEE*, 99(4):540–555, Apr. 2011. 14

[136] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricordel, P. Le Callet, J. Gutierrez, and N. Garcia. NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences. *International Workshop on Quality of Multimedia Experience*, pages 109–114, Jul. 2012. 83

[137] Video Quality Experts Group. Report on the validation of video quality models for High Definition Video Content. 2010. 55

[138] Video Quality Experts Group. Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, phase II. *VQEG*, Aug. 2003. 148, 170

[139] VQEG. Video Quality Expert Group. *http://www.its.bldrdoc.gov/vqeg*. 2

[140] K. Wang, M. Barkowsky, R. Cousseau, K. Brunnström, R. Olsson, P. Le Callet, and M. Sjöström. Subjective evaluation of HDTV stereoscopic videos in IPTV scenarios using absolute category rating. *IS&T/SPIE Electronic Imaging*, pages 78631T–78631T, Jan. 2011. 19, 21, 102

[141] L. Wang, Y. Tu, L. Chen, P. Zhang, K. Teunissen, and I. Heynderickx. Crosstalk acceptability in natural still images for different (auto) stereoscopic display technologies. *Journal of the Society for Information Display*, 18(6):405–414, 2010. 116

[142] C. Wheatstone. Contributions to the physiology of vision. Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London*, 128:371ï¿½–394, Jan. 1838. 13

[143] F. Wickelmaier and C. Schmid. A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior Research Methods, Instruments, and Computers*, 36(1):29–40, Feb. 2004. 75, 99

[144] J.W. Wilkinson. An analysis of paired comparison designs with incomplete repetitions. *Biometrika*, 44(1/2):97–113, 1957. 34

[145] A.J. Woods, T. Docherty, and R. Koch. Image distortions in stereoscopic video systems. In *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*, pages 36–48. International Society for Optics and Photonics, 1993. 114

[146] Q. Xu, Q. Huang, T. Jiang, B. Yan, W. Lin, and Y. Yao. Hodgerank on random graphs for subjective video quality assessment. *Multimedia, IEEE Transactions on*, 14(3):844–857, 2012. 35

[147] K. Yamagishi, L. Karam, J. Okamoto, and T. Hayashi. Subjective characteristics for stereoscopic high definition video. *International Workshop on Quality of Multimedia Experience*, pages 37–42, Sep. 2011. 19

[148] H. Yamanoue, M. Nagayama, M. Bitou, J. Tanada, T. Motoki, T. Mituhashi, and M. Hatori. Tolerance for geometrical distortions between l/r images in 3d-hdtv. *Systems and Computers in Japan*, 29(5):37–48, 1998. 114, 196

[149] S. Yano, M. Emoto, and T. Mitsuhashi. Two factors in visual fatigue caused by stereoscopic HDTV images. *Displays*, 25(4):141–150, 2004. 71, 111, 116, 121, 141

[150] 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):191–201, 2002. 71, 111, 117, 155, 157

[151] J.H. Yu, B.H. Lee, and D.H. Kim. Eog based eye movement measure of visual fatigue caused by 2d and 3d displays. In *Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on*, pages 305–308. IEEE, 2012. 120, 180

l'université nantes angers le mans
PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Jing LI

Methods for assessment and prediction of QoE, preference and visual discomfort in multimedia application with focus on S-3DTV

Méthodes pour l'évaluation et la prédiction de la Qualité d'expérience, la préférence et l'inconfort visuel dans les applications multimédia. Focus sur la TV 3D stéréoscopique

Résumé

La technologie multimédia vise à améliorer l'expérience visuelle des spectateurs, notamment sur le plan de l'immersion. Les développements récents de la TV HD, TV 3D, et TV Ultra HD s'inscrivent dans cette logique. La qualité d'expérience (QoE) multimédia implique plusieurs dimensions perceptuelles. Dans le cas particulier de la TV 3D stéréoscopique, trois dimensions primaires ont été identifiées dans la littérature: qualité d'image, qualité de la profondeur et confort visuel. Dans cette thèse, deux questions fondamentales sur la QoE sont étudiés. L'une a pour objet "comment évaluer subjectivement le caractère multidimensionnel de la QoE". L'autre s'intéresse à une dimension particuliére de QoE, "la mesure de l'inconfort et sa prédiction?". Dans la première partie, les difficultés de l'évaluation subjective de la QoE sont introduites, les mérites de méthodes de type "Comparaison par paire" (Paired Comparison en anglais) sont analysés. Compte tenu des inconvénients de la méthode de Comparaison par paires, un nouveau formalisme basé sur un ensemble de comparaisons par paires optimisées, est proposé. Celui-ci est évalué au travers de différentes expériences subjectives. Les résultats des tests confirment l'efficacité et la robustesse de ce formalisme. Un exemple d'application dans le cas de l'étude de l'évaluation des facteurs influençant la QoE est ensuite présenté. Dans la seconde partie, l'influence du mouvement tri-dimensionnel (3D) sur l'inconfort visuel est étudié. Un modèle objectif de l'inconfort visuel est proposé. Pour évaluer ce modèle, une expérience subjective de comparaison par paires a été conduite. Ce modèle de prédiction conduit à des corrélations élevées avec les données subjectives. Enfin, une étude sur des mesures physiologiques tentant de relier inconfort visuel et fréquence de clignements des yeux présentée.

Mots clés

qualité d'expérience, méthodes subjectives, comparaison par paires, l'expérience par préférence, inconfort visuel, modèle objectif, prédiction psychophysique, TV 3D stéréoscopique.

Abstract

Multimedia technology is aiming to improve people's viewing experience, seeking for better immersiveness and naturalness. The development of HDTV, 3DTV, and Ultra HDTV are recent illustrative examples of this trend. The Quality of Experience (QoE) in multimedia encompass multiple perceptual dimensions. For instance, in 3DTV, three primary dimensions have been identified in literature: image quality, depth quality and visual comfort. In this thesis, focusing on the 3DTV, two basic questions about QoE are studied. One is "how to subjectively assess QoE taking care of its multidimensional aspect?". The other is dedicated to one particular dimension, i.e., "what would induce visual discomfort and how to predict it?". In the first part, the challenges of the subjective assessment on QoE are introduced, and a possible solution called "Paired Comparison" is analyzed. To overcome drawbacks of Paired Comparison method, a new formalism based on a set of optimized paired comparison designs is proposed and evaluated by different subjective experiments. The test results verified efficiency and robustness of this new formalism. An application is the described focusing on the evaluation of the influence factor on 3D QoE. In the second part, the influence of 3D motion on visual discomfort is studied. An objective visual discomfort model is proposed. The model showed high correlation with the subjective data obtained through various experimental conditions. Finally, a physiological study on the relationship between visual discomfort and eye blinking rate is presented.

Key Words

QoE, subjective methodology, Paired Comparison, Preference of Experience, visual discomfort, objective model, psychophysical prediction, S-3DTV.