Liyuan Xing

Towards Reliable Stereoscopic 3D Quality Evaluation

Subjective Assessment and Objective Metrics

Thesis for the degree of Philosophiae Doctor

Trondheim, August 2013

Norwegian University of Science and Technology
Faculty of Information Technology,
Mathematics and Electrical Engineering
Department of Electronics and Telecommunications

NTNU – Trondheim
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Information Technology, Mathematics and Electrical Engineering
Department of Electronics and Telecommunications

ISBN 978-82-471-4533-3 (printed ver.)
ISBN 978-82-471-4534-0 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2013:209

Printed by NTNU-trykk

Abstract

Stereoscopic three-dimensional (3D) services have become more popular recently amid promise of providing immersive quality of experience (QoE) to the end-users with the help of binocular depth. However, various arisen artifacts in the stereoscopic 3D processing chain might cause discomfort and severely degrade the QoE. Unfortunately, although the causes and nature of artifacts have already been clearly understood, it is impossible to eliminate them under the limitation of current stereoscopic 3D techniques. Moreover, their influence on the perceived quality is not well understood. Therefore, quality evaluation, both subjective assessment and objective metrics are necessarily required to understand, measure and eventually, model and predict stereoscopic 3D quality. The thesis, composed of six collected papers, contributes to the field of quality evaluation of stereoscopic 3D media in three aspects.

First, quality evaluation of crosstalk perception was carried out on polarized stereoscopic display, since crosstalk is one of the most annoying artifacts in the visualization stage of stereoscopic 3D and can not be completely eliminated with current technologies. The subjective tests were customized for crosstalk perception with varying independent parameters of scene content, camera baseline and crosstalk level. An objective metric for crosstalk perception was proposed based on our findings of perceptual attributes of crosstalk perception, namely shadow degree, separation distance and spatial position of crosstalk. Furthermore, subjective crosstalk assessment methodologies for auto-stereoscopic displays at arbitrary viewing positions in a specified area were suggested, supported by a head and score tracking system.

Second, an extension from crosstalk perception to QoE in the simplest stereoscopic system was studied, since QoE is often referred as a criterion for the acceptance of any commercial system and determines the success or not. In addition to crosstalk level, other requisite factors of the simplest stereoscopic system, including scene content, camera baseline, screen size and viewing position have also been investigated on their relationships to perceptual attributes of QoE. Specifically, those perceptual attributes are crosstalk perception and depth enabled visual comfort, which cover the main negative and positive aspects of stereoscopic QoE. By modeling these perceptual attributes separately and combining them thereafter, an objective QoE metric was proposed.

Third, further work on the complex stereoscopic system was carried out by investigating the influence of coding artifacts on QoE. This work was done under the context of assessing 3D video compression technologies within MPEG's effort for standardizing 3D video coding techniques. However, the subjective scores can be used as ground truth dataset for proposing QoE model which will incorporate

both the coding artifacts and configurations of the complex stereoscopic system. Meanwhile, since the subjective tests were conducted at 13 laboratories around the world with large amount of test sessions, it can be used for defining a process for certification of subjective test campaigns.

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of *doctor of philosophy* (PhD) at the Norwegian University of Science and Technology (NTNU). The work was performed at the Centre for Quantifiable Quality of Service in Communication Systems (Q2S), which is a Norwegian Center of Excellence, appointed by the Research Council of Norway and funded by the Research Council, NTNU and UNINETT. Professor Andrew Perkis has been the main supervisor of this work, and professor Touradj Ebrahimi has been the co-supervisor.

Acknowledgements

I would like to thank several people who gave me a lot of support throughout my PhD study at Q2S. First and foremost I offer my sincerest gratitude to my main supervisor Andrew Perkis, who guided me to the interesting research topic when I started my PhD study and allowed me the freedom to pursue my research during the study. When I encountered difficulties in my study, either technical or non-technical, he was always there to give me guidance and confidence to stride forward. I would also like to thank my co-supervisor Touradj Ebrahimi, his outstanding research skills and elaborate research attitude influence me a lot. He always provided me with constructive advice and discussion whenever he came to Q2S for a visit. Without their patience and knowledge, this work would not be finished.

I am deeply grateful to my colleague and main co-author Dr. Junyong You for his valuable suggestion and the happy collaboration. He always first time reviewed and polished the co-author papers, which were much improved after his work. I am also very grateful to my colleague and husband Dr. Jie Xu, who has provided me unlimited support, encouragement and insightful discussion both on life and work since we first met, especially during my PhD study.

All past and present colleagues at the Q2S center are thanked for creating a warm, joyful and stimulating environment. I also thank my friends for the their encouragement and accompany over the years. My special thanks go to Anniken Skotvoll and Jo Haldor Kvello for their kindness and many helps, which make me feel warm in the heart. I wish you all happiness and a bright future.

Last while not least, I would like to thank my family, especially my parents, uncle, brother, husband and son for all your understanding, support and love.

List of Papers

The thesis is based on the following papers:

Paper A. L. Xing, J. You, T. Ebrahimi, and A. Perkis, **Assessment of Stereoscopic Crosstalk Perception**, *IEEE Transactions on Multimedia (TMM)*, vol. 14, no. 2, pp. 326-337, 2012.

Paper B. L. Xing, J. Xu, K. Skildheim, T. Ebrahimi, and A. Perkis, **Subjective Crosstalk Assessment Methodology for Auto-stereoscopic Displays**, *2012 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 515-520, 2012.

Paper C. L. Xing, J. You, T. Ebrahimi, and A. Perkis, **Stereoscopic Quality of Experience: Subjective Assessment and Objective Modeling**, *Submitted to IEEE Transactions on Multimedia (TMM)*, 2013.

Paper D. L. Xing, J. You, T. Ebrahimi, and A. Perkis, **Factors Impacting Quality of Experience in Stereoscopic Images**, *Stereoscopic Displays and Applications XXII (SDA)*, vol. 786304, pp. 786304-786304-8, 2011.

Paper E. L. Xing, J. You, T. Ebrahimi, and A. Perkis, **Objective Metrics for Quality of Experience in Stereoscopic Images**, *18th IEEE International Conference on Image Processing (ICIP)*, pp. 3105-3108, 2011.

Paper F. A. Perkis, J. You, L. Xing, T. Ebrahimi, F. De Simone, M. Rerabek, P. Nasiopoulos, Z. Mai, M. Pourazad, K. Brunnstrom, K. Wang, and B. Andren **Towards Certification of 3D Video Quality Assessment**, *6th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, pp. 1-6, 2012.

In addition we have published the following papers:

Paper i. L. Xing, T. Ebrahimi, and A. Perkis, **Subjective Evaluation of Stereoscopic Crosstalk Perception**, *SPIE Visual Communications and Image Processing (VCIP)*, pp. 77441V-77441V-9, 2010.

Paper ii. L. Xing, J. You, T. Ebrahimi, and A. Perkis, **A Perceptual Quality Metric for Stereoscopic Crosstalk Perception**, *17th IEEE International Conference on Image Processing (ICIP)*, pp. 4033-4036, 2010.

Paper iii. L. Xing, J. You, T. Ebrahimi, and A. Perkis, **An Objective Metric for Assessing Quality of Experience on Stereoscopic Images**, *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pp. 373-378, 2010.

Paper iv. L. Xing, J. You, T. Ebrahimi, and A. Perkis, **Estimating Quality of Experience on Stereoscopic Images**, *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 1-4, 2010.

Paper v. L. Xing, J. You, T. Ebrahimi, and A. Perkis, **Stereoscopic Quality Datasets under Various Test Conditions**, *5th International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 136-141, 2013.

Paper vi. J. You, L. Xing, T. Ebrahimi, and A. Perkis, **Visual Contrast Sensitivity Guided Video Quality Assessment**, *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 824-829, 2012.

Paper vii. M. Barkowsky, K. Brunnström, T. Ebrahimi, L. Karam, P. Lebreton, P. Le Callet, A. Perkis, A. Raake, M. Subedar, K. Wang, L. Xing, and J. You, **Subjective and Objective Visual Quality Assessment in the Context of Stereoscopic 3D-TV**, *3D-TV System with Depth-Image-Based Rendering, Architectures, Techniques and Challenges*, pp. 413-437, 2013.

Paper viii. J. Xu, L. Xing, A. Perkis, and Y. Jiang, **On the Properties of Mean Opinion Scores for Quality of Experience Management**, *IEEE International Symposium on Multimedia (ISM)*, pp. 500-505, 2011.

Paper ix. J. You, L. Xing, T. Ebrahimi, and A. Perkis, **Perceptual Audio-Visual Quality Metrics: Methodologies, Evolution, and Future**, *IEEE COMSOC MCTC E-letter*, pp. 5-10, 2011.

Paper x. M. Lervold, L. Xing, and A. Perkis, **Quality of Experience in Internet Television**, *5th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, pp. 1-6, 2010.

Paper xi. J. You, L. Xing, A. Perkis, and X. Wang, **Perceptual Quality Assessment for Stereoscopic Images Based on 2D Image Quality Metrics and Disparity Analysis**, *5th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, pp. 1-6, 2010.

Paper xii. J. You, G. Jiang, L. Xing, and A. Perkis and X. Wang, **Quality of Visual Experience for 3D Presentation - Stereoscopic Image**, *High-Quality Visual Experience*, pp. 51-77, 2010.

Paper xiii. C. Liu, Q. Huang, S. Jiang, L. Xing, Q. Ye, and W. Gao, **A framework for Flexible Summarization of Racquet Sports Video using Multiple Modalities**, *Computer Vision and Image Understanding (CVIU)*, vol. 113, no. 3, pp. 415-424, 2009.

Contents

Part I

Thesis Introduction

1 Introduction

Stereoscopic three-dimensional (3D) technique is based on simultaneously capturing a pair of two-dimensional (2D) images and then separately delivering them to respective eyes. Consequently, 3D perception is generated in the human visual system (HVS). Thus, it can present observers with true 3D views of a real world scene. By having the ability to add a new (third) dimension by means of stereopsis, stereoscopic 3D further blurs the lines between the real and virtual worlds. Thereby, stereoscopic 3D is believed to be a logical next step forward in the evolution of multimedia communication system towards greater realism and richer quality of experience (QoE), after introducing audio, black-and-white visual content, color visual content and high resolution. It can be expected that future stereoscopic 3D media services will transform the way people live, play and work.

However, stereoscopic 3D is not a new coming technique, it has been around over a century but not yet accepted widely by the masses. Although a successful market introduction of stereoscopic 3D is believed to be just a matter of time and has been compared to the transition from black-and-white to color TV, it has not happened because of the safety and health issues related to stereoscopic 3D displays [54]. Some common complications, such as eye strain, visual discomfort and even headache may occur when watching stereoscopic 3D media. They primarily origin from the artifacts in various stages in the stereoscopic 3D processing chain [8]. Among those, the essential ones are those unnatural for the human vision [102], such as crosstalk caused by imperfect image separation, limited comfort zone of display caused by excessive screen disparity and mismatch between convergence and accommodation. All the artifacts might modify the observer's perception of the depicted scene, result in the degradation of the perceived quality, or inducing the complications. Unfortunately, although the causes and nature of artifacts have already been clearly studied, it is impossible to eliminate them due to the limitation of current stereoscopic 3D techniques. Moreover, their influence on the perceived quality is not well understood.

Thus, quality evaluation is required during design and development of the end-to-end stereoscopic 3D media services, since the success of any commercial services depends heavily on the quality offered to users. Besides, quality evaluation is one of the most important aspects which has a significant impact on all the other aspects of the stereoscopic 3D processing chain. Although many quality evaluation methods have been proposed to assess the perceptual quality of traditional 2D images, no similar effort has been widely devoted to evaluating the perceptual quality of stereoscopic 3D media. Therefore, quality evaluation on stereoscopic 3D is required to benchmark the performance of production techniques as well as to provide a means to optimize the parameters of different algorithms. In particular, both subjective

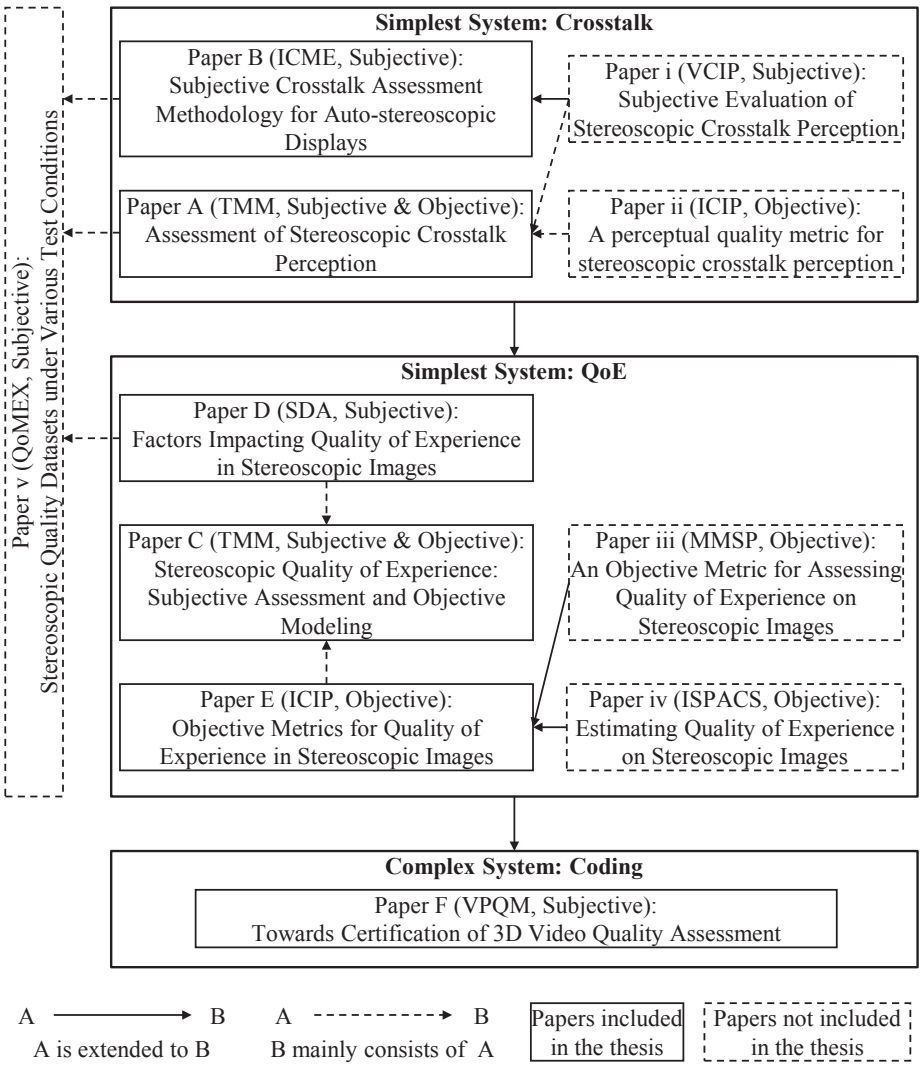


Figure 1.1: Paper overview to quality assessment on stereoscopic 3D.

assessment and objective metrics need to be adopted for stereoscopic 3D quality evaluation considering the special characteristics of stereoscopic 3D perception.

The focus of this thesis is towards reliable stereoscopic 3D quality evaluation as depicted in Figure 1.1. In particular, the simplest system consists of acquisition and visualization stages supporting two/multi videos format only. While the complex system contains representation, coding, delivery, decoding, conversion stages

as well, in addition, the acquisition and visualization stages support other formats, such as video plus depth. We emphasized on those artifacts in the simplest system which can not be eliminated completely, such as crosstalk (Paper A, Paper B, Paper i and Paper ii) and configurations of the simplest stereoscopic system which are critical to QoE (Paper C, Paper D, Paper E, Paper iii and Paper iv). Quality evaluation, both subjective tests were conducted and objective metrics were proposed for crosstalk perception and QoE in the simplest system, with more details in Figure 1.1. Particularly, the ground truth datasets were released to public through Paper v. Coding artifacts in the complex stereoscopic systems were investigated by subjective assessment (Paper F), which provides the possibility to extend the objective metrics for the simplest system to the complex system.

The introduction is organized as follows. First, a general stereoscopic 3D background is presented in Chapter 2. Then the main quality issues arising from distortions introduced in the processing chain of the simplest and complex stereoscopic system are overviewed in Chapter 3. Chapter 4 outlines state-of-the-art survey on subjective assessment and objective metrics of stereoscopic 3D quality evaluation, starting from the counterparts of 2D. Chapter 5 provides the outline and short summaries of the published papers and identifies their specific contributions. Finally, concluding remarks and a list of future work directions are provided in Chapter 6.

2 Background on Stereoscopic 3D

This chapter gives a short background for the work presented in this thesis. First, stereoscopic 3D history and applications are presented in Section 2.1. Then, Section 2.2 presents the mechanisms behind the human vision depth perception. The simplest stereoscopic system from binocular cue is introduced in Section 2.3, while the complex stereoscopic system with various stages of typical stereoscopic 3D processing chain are described in Section 2.4.

2.1 History and Applications

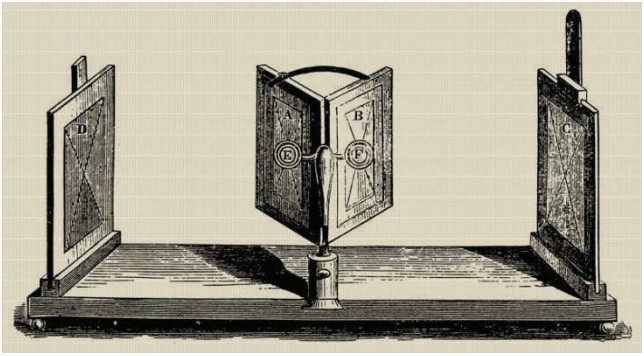


Figure 2.1: The mirror stereoscope invented by C. Wheatstone. (From Optics in [101]. This image is public domain.)

As early as in 1838, C. Wheatstone first invented stereoscopy and stereoscope [101]. He originally used his stereoscope, a rather bulky device with a pair of mirrors at 45 degree angles to the user's eyes, each reflecting a drawing located off to the side, as shown in Figure 2.1. Later in 1890, 3D motion pictures using stereoscope were first patented. In 1915, first 3D footage in cinema using anaglyph glasses was made. Soon, shutter based technique and polarization based projection were introduced in 1922 and 1936 respectively. Due to the invention of television, came the golden era of 3D movies in 1952-1955. Later in the 1980s and 1990s, 3D movies experienced a worldwide resurgence driven by IMAX high-end theaters and Disney themed-venues. Throughout the 2000s, 3D entertainments became more and more successful. In 2003, J. Cameron shoot the first full length 3D feature film for IMAX screens, and later in 2009 he made a unprecedented success

of 3D presentations of Avatar. It is also worthy to mention that animation "Polar Express" made 14 times more revenue in 3D than 2D in 2004. Moreover, music (e.g. U2 3D 2008 and In Concert 3D 2009), documentary (e.g. Biodiversity 2009), sports (e.g. NBA ALL Star Game 2009, Six Nations Cup 2009, FIFA World Cup 2020), games (e.g. 19 PS3 titles in 2010) as well as 3D blu-ray achieve great successes. More recently, stereoscopic displays (3DTV) and cameras have been introduced to the public consumer market and come into daily life of those people having fancy in or willingness to try new technologies.

In addition to the multimedia entertainments, stereoscopic 3D techniques have been quite successful for particular niche applications, such as aerial photography, minimally invasive telesurgery, undersea operations and so on. Specifically, stereoscopic 3D has potential benefits in performing tasks a) through visual clutter or noise, b) under conditions of complex and non-rectangular shapes and unfamiliar orientations, c) having floating objects that do not touch ground plane, d) requiring ballistic motions, e) with high penalty for errors and irreversible actions, f) in poor visibility or image quality.

2.2 Depth Perception

The HVS consists of eyes, especially the retina, visual pathways and visual cortex. Eyes are the optical part of the visual perception, it controls the amount of light by iris and brings the image into focus by lens, followed by information captured by retinal cells. The captured information about the image is transmitted to the brain along the visual pathways. Finally, visual cortex is responsible for processing the visual image. Among which, primary visual cortex responds to low level visual information such as frequencies, color and direction, while dorsal and ventral streams are dealing with motions and objects. Thus, our vision is active, not passive.

One of the major functions of HVS is to construct a 3D representation of the world surrounding us. Based on the two 2D retinal images at the back of your eyes, which do not preserve 3D information in the scene, it is the brain that reconstructs this 3D information by inferring the size, shape, and spatial positions of objects out there in the world. In order to provide an accurate, consistent, and useful percept of the physical environment, HVS relies on and reduces ambiguity by combination of different depth cues. The depth cues used in different layers in human vision are static pictorial cues (occlusion, linear perspective, size gradient, texture gradient, depth of focus, cast shadows), dynamic cue (motion parallax, dynamic occlusions, motion perspective), oculomotor cues (accommodation, convergence, myosis) and binocular cue (retinal disparity) [70]. Their relationships are summarized in Figure 2.2.

2.2.1 Monocular Cues

These kinds of cues can be extracted directly from monocular 2D pictures, either static or dynamic. That is why we can still perceive and judge depth in the real

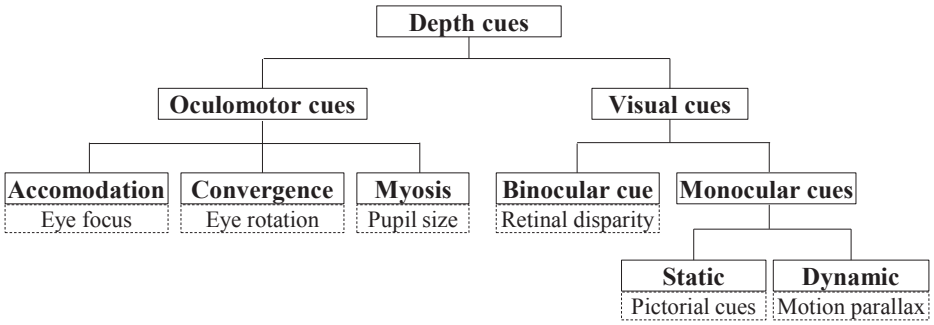


Figure 2.2: Relationship of depth cues.

world even if we close one of our eyes. This kind of characteristic enables people to perceive good depth when viewing 2D images or videos in traditional 2D display. In particular, static pictorial cues include occlusion, linear perspective, size gradient, texture gradient, depth of focus, cast shadows, and so on, while motion based cues include motion parallax, dynamic occlusions, motion perspective and so on. These monocular cues are effective for longer distances than binocular cue, especially motion based cues are very important for a wide range of scene depths. Although the monocular cues are possible to appreciate the relative location of objects, it is impossible to discriminate acute depth as binocular cue does.

2.2.2 Oculomotor Cues

These are achieved by muscles attached to the eye itself and the lens. In particular, accommodation is the ability of the eye to change the focal length of the lens to focus on objects at various distances (typically less than 2 meters). Convergence is the ability of the eye to rotate towards each other for closer objects, typically effective for distances less than 10 meters. And myosis is the ability of the eye to determine both amount of light and depth of field (DOF) by controlling its pupil size. All of them are weak depth cues for short distances and their importance quickly decreases when the distance increases. Moreover, oculomotor cues are dependent on and interactive with each other. They are affected unconsciously by monocular and binocular visual depth cues.

Accommodation and myosis: Adjusting accommodation to perceive a sharp image is not always necessary, since our eyes can tolerate small amount of retinal defocus, which is described by DOF of myosis. Specifically, the range of DOF depends on many factors [54], including target attributes (e.g. contrast, luminance, and spatial frequency) and eye/brain attributes (e.g. pupil size and age). Although the typical DOF is approximately 0.2 to 0.5 diopter, it can range from 0.04 to 3.50 diopter in the maximum.

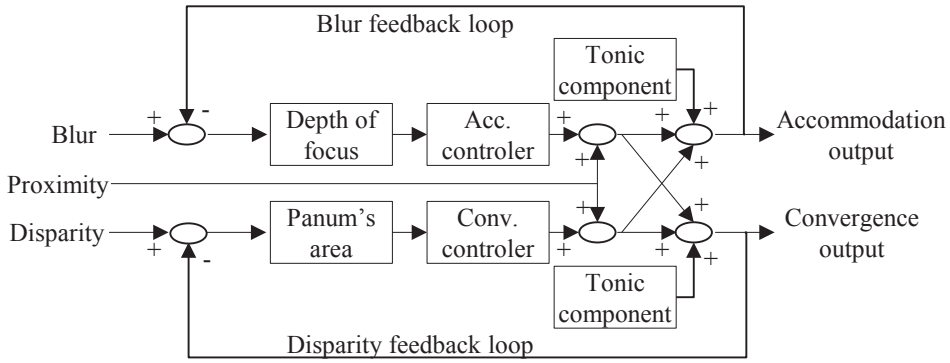


Figure 2.3: Accommodation and convergence are modeled as two dual parallel feedback control systems that interact via crosslinks. (Reproduced from Figure 1 in [54])

Accommodation and convergence: Accommodation and convergence are generally modeled as two dual parallel feedback control systems that interact via crosslinks, as depicted in Figure 2.3. Accommodation and convergence are primarily driven by retinal blur and disparity, respectively, while both of them respond to proximity information (e.g. monocular and binocular depth cues) as well. Specifically, small degree of retinal defocus within the DOF does not drive the accommodation controller, only when it exceeds DOF. Likewise, small retinal disparity within fusional area does not drive the convergence controller, only when it is outside fusional area. Each system also includes a tonic component which accounts for slower adaptations to altered viewing situations. Furthermore, both systems interact via reflexive crosslink interactions. The gains of the crosslink interactions are described by the AC/A ratio (i.e., the change in vergence due to accommodation per change in accommodation in the absence of retinal disparity) and the CA/C ratio (i.e., the change in accommodation due to vergence per change in vergence in the absence of blur). Finally, the summed output of the controller, the proximal component, the tonic component, and the crosslink describes the overall system's response and provides negative feedback to the input stimuli to obtain a stable state.

2.2.3 Binocular Cue

This is a consequence of both eyes observing scene from two slightly different angles, since human eyes are separated horizontally in a range of 50 to 70mm, by 63mm on average [16]. The field of view (FOV) of each eye is approximately 140 degrees, which results in a total FOV of 160 to 208 degrees and stereoscopic field of 120 to 180 degrees [66], as illustrated in Figure 2.4.(a). For those objects in stereoscopic field, two slightly different retinal images are formed as shown in Figure 2.4.(b). Specifically, the mechanism of binocular depth estimation consists of convergence

and stereopsis [8]. Convergence is the process that both eyes take a position which minimizes the difference of the visual information projected in both retinae. Existence of the difference in retinal images leads to binocular disparity, which is used in the process of stereopsis to estimate the relative depth between the convergence point and its surrounding area. This binocular cue is the most important depth cue for medium viewing distances.

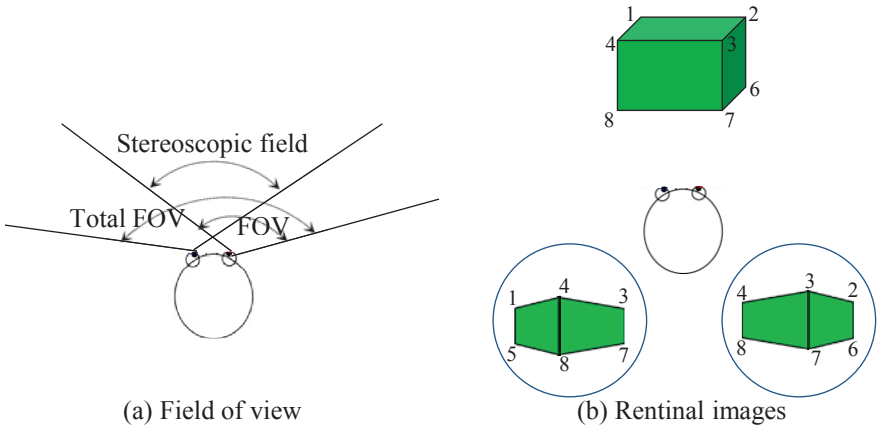


Figure 2.4: Stereoscopic vision.

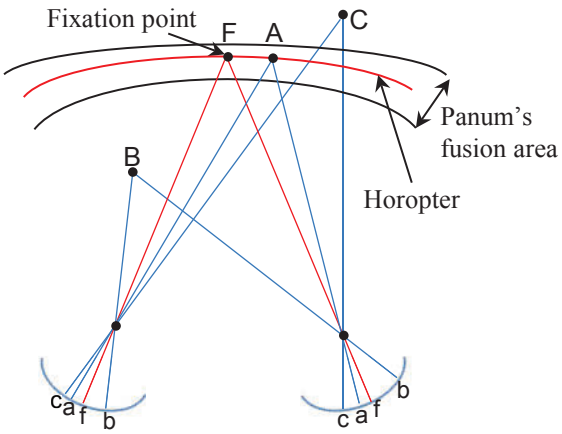


Figure 2.5: Horopter and panum's fusional area.

Panum's fusional area: A small volume of visual space around where both eyes are fixating is projected onto the retina. Those points stimulate the retinal points lying in a surface which is called horopter, they have zero retinal disparity (e.g. point A and F in Figure 2.5). While others have retinal disparity, either

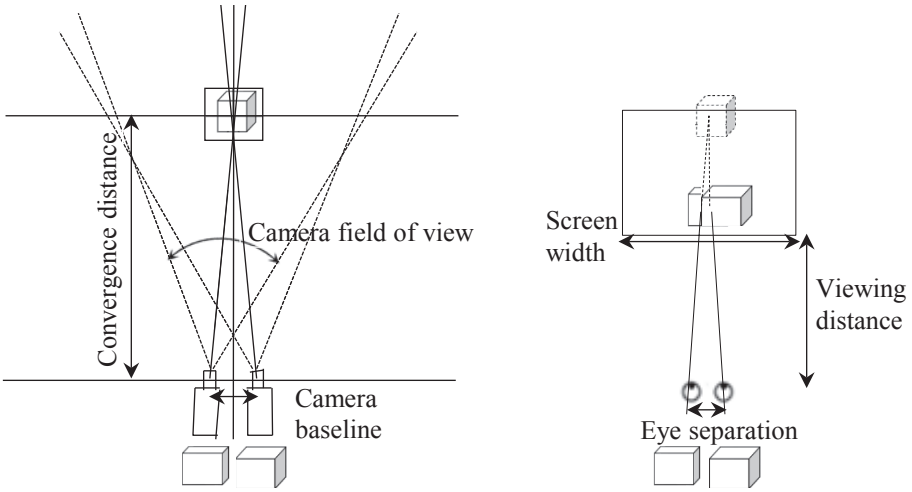
negative retinal disparity (e.g. point B in Figure 2.5) if the points located in front of the horopter or positive retinal disparity (e.g. point C in Figure 2.5) if behind. Panum's fusional area is the region around the horopter of binocular single vision. Outside the panum's fusional area, double vision occurs. In particular, the retinal disparity limits of panum's fusional area are not constant but increase along the fovea to periphery, as illustrated in Figure 2.5. Typically, the retinal disparity limit at the fovea is 0.1 degree, while it can be 0.33 or 0.66 degrees [76] at an eccentricity of 6 and 12 degrees, respectively. Moreover, convergence movements [126], stimulus durations, stimulus properties [74], temporal modulation of retinal disparity information [85], exposure duration, amount of illuminance [86], and individual differences [82] are found to have effect on the fusion limit. Usually, long stimulus durations and convergence eye movements, large moving objects and the addition of peripheral objects to the fixation object can increase the fusion limits. For example, if the retinal disparity of the reconstituted object surpasses panum's fusional area, convergence movements relocate the retinal disparity within panum's fusional area and as such, increase fusion limits (i.e., motoric fusion).

Individual differences: A person's stereo acuity determines the minimum image disparity they can perceive as depth. Stereo acuity depends on the individual differences. It is believed that approximately 5% of people are stereoscopically latent [39, 10], and up to 30% of people have very weak stereoscopic vision [108], which prevents them from depth perception based on binocular cue. However, it is possible to improve the stereo acuity by orthoptics treatment.

2.3 From Binocular Cue to the Simplest Stereoscopic System

Binocular cue is especially activated in stereoscope [101], invented by Sir Wheatstone in 1838. This device uses additional instruments (e.g. mirrors) to separate the left and right viewing channels, thus each eye receives the corresponding view respectively. This basic principle of stereoscope is the ancestor of modern stereoscopic device. The original images for the "stereoscope", which differed only in horizontal disparity, were drawings, since photography was not yet available. Nowadays, instead of drawings, stereoscopic images are captured by two cameras placed in either toed-in or parallel configuration.

Figure 2.6 depicts the principle of the simplest stereoscopic system. First, two cameras are used to record the views from left and right positions of eyes in the acquisition stage. Then, these recorded images are delivered to the left and right eyes respectively in the visualization stage. As illustrated in Figure 2.6, the requisite factors of the simplest stereoscopic imaging system consists of a) camera baseline: the distance between the cameras, b) convergence distance: the distance away from the cameras at which the optical axes of the cameras intersect, c) camera field of view: determined by the camera imaging sensors' format size and the lens focal length, d) viewing distance: the distance between the viewer and the display, e)



(a) Acquisition (b) Visualization

Figure 2.6: Principle and requisite factors of the simplest stereoscopic imaging system. (Adapted from Figure 1 in [118])

screen size: mainly measured by screen width and f) eye separation: the distance between the viewer's eyes.

Through this simplest system, binocular disparity is first represented by image disparity in a pair of left and right images, and later represented by screen disparity when the image pairs are shown in the stereoscopic display, finally recovered to be binocular disparity when the image pairs are viewed by left and right eyes. In this way, a stereoscopic imaging system is able to create an illusion of depth sensation by adding binocular disparity information. However, with different settings of requisite factors in the simplest stereoscopic system, the binocular disparity reconstructed in the virtual world may be different from the binocular disparity directly viewed in the real world. The imperfect duplication of HVS also includes a) accommodation is not implemented, b) image separation is not perfect and so on, which is due to the limitation of acquisition and visualization techniques. Therefore, the main techniques used to implement the simplest stereoscopic system will be introduced in the following subsections, followed by its quality issues in next chapter.

2.3.1 Acquisition

The most direct way to create pairs of views for the left and right eyes is to set up two/multi cameras for synchronized capturing. No representation or conversion is required for final visualization.

Two-camera system: The configurations of two cameras can be either toe-in or parallel. In particular, in toe-in camera configurations, the cameras are each

rotated inwards from parallel, thus the lens optical axes are converged. While, in parallel camera configurations, the center of each imaging sensor of cameras is moved away (outwards) from the lens optical axis, but the lens optical axes are still kept parallel. It is important to have matching cameras, in terms of photography(white balance, sensitivity, shutter speed, aperture), optics(focal length), geometry(distance, angle) and synchronization in capturing. In practice, it is impossible to make them exactly match, and 3D post-processing are usually needed, including geometric alignment, color adjustment and so on. Moreover, it is better to record all the intrinsic (focal length, image format, and principal point) and extrinsic (position of the camera center and the camera's heading in world coordinates) camera parameters together with the captured video sequences in case of further use. Practically, these parameters are not easy to measure. While it is not the case in computer-generated imagery (CGI) approach, where the parameters of virtual cameras can be set easily. With 3D models in hand, it is easy to create the two/multi video sequences by setting up two/multi virtual cameras. However, it is usually time-consuming to create the 3D models.

Multi-camera system: The two-camera system can be easily extended to multi-camera system by adding more cameras. These cameras can be organized in line, circle and array. A multi-camera system consists of 100 cameras was introduced in a real-time free viewpoint television (FTV) system [103]. In this way, a more precise 3D representation is provided.

2.3.2 Visualization

After acquiring stereoscopic sequences, stereoscopic 3D displays are required to transfer those stereoscopic sequences separately to left and right eyes of the viewers. There are all kinds of stereoscopic display techniques to display two images exclusively. Usually they are classified into two categories, stereoscopic and autostereoscopic display, depending on needing an eye wear or not.

Stereoscopic displays: Special glasses such as anaglyph glasses, polarized glasses and shutter glasses for multiplexing two images are needed in stereoscopic displays. The anaglyph glassess consist of different color filters on each eye which separate the displayed images. Those color filters should have different colors whose color spectra do not overlap and match the spectrum of the displayed images, typically red and cyan, so that each eye receives on its single 2D intensity image. The anaglyph 3D system is easy to set up due to cheap glasses, no need for special monitor and easy creation of displayed images. However, it has poor color quality and accompanies obvious ghosts because full-color anaglyph based stereograms are impossible. Therefore, either polarization or shutter based system is preferred in current stereoscopic 3D cinema because of better color quality and stereo effect.

Specifically, polarization based system is based on the fact that light is an electromagnetic wave and its direction can be controlled by polarizing filters. First, the images for the left and the right eyes are differently polarized by the polarizing filters and simultaneously displayed on a sliver screen in superposed form. Then, polarized glasses with matching polarization properties are required to filter out

each image and deliver it to the corresponding eye. The polarization forms can be either linear or circular and both can provide full color stereoscopy.

However, shutter based systems are based on time-multiplexed display of right and left eye images. Those two images are displayed alternately usually in a frequency of 140-200 images per second. Meanwhile, shutter glasses are synchronized to the displayed image by turning opaque and transparent correspondingly with the help of an electronically remote control system.

Auto-stereoscopic displays: No special eye-wear is required in auto-stereoscopic displays. There are two common technologies for auto-stereoscopic viewing: lenticular and barrier. Both technologies first display a single image by fusing the two captured stereo images on a conventional 2D display, and then extract the single image back to be two images by mounting an additional sheet on the 2D display. In particular, a barrier is like a fence consisting of opaque and transparent stripes while a lenticular consists of tiny cylindrical lenses in either vertical or slanted stripe. Importantly, a close match of the geometry between the underlying 2D display and the barrier or lenticular is needed. This ensures that the eyes of the observer at a particular location in front of the display receives their own different left and right images. Therefore, there is a specific range of positions, namely "sweet spot", where observer's eyes receive the intended separate right and left images correctly. If the observer is not in this "sweet spot", there is no 3D perception.

Extension of auto-stereoscopic display to support more than two views by using the aforementioned technologies is straightforward, which leads to multiview auto-stereoscopic display [17]. Typically there are five, eight or nine views having slight difference from a narrow horizontal viewing angle, usually 20-30 degrees. An observer, located at a particular viewing angle, receives one pair of two images corresponding to that viewing angle. If the observer moves right or left, another pair related to the new viewing angle is received. Therefore, the horizontal parallax is visible in the multiview auto-stereoscopic display, which creates more natural viewing experience. However, multiview auto-stereoscopic suffers from the "sweet spot" as well.

2.4 The Complex Stereoscopic System with Processing Chain

As shown in Figure 2.7, in addition to two/multi videos format, the acquisition and visualization stages of the complex stereoscopic system support other formats, such as video plus depth as well. Moreover, the complex stereoscopic system has various stages (e.g. representation, coding, delivery, decoding, conversion) which usually are needed in practical stereoscopic 3D communication applications (e.g. 3DTV, FTV, telesurgery). Therefore, the complex stereoscopic systems have more complexities in different scenarios, and various techniques are available in different stages of the content processing chain.

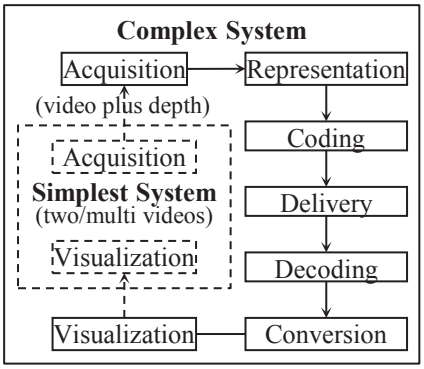


Figure 2.7: 3D processing chain: the simplest system versus the complex system.

2.4.1 Acquisition/Visualization

Instead of capturing two/multi video sequences directly, the video plus depth sensor approach captures one video sequence by a normal RGB camera and its depth map by a depth sensor. Specifically, the depth map is a gray image that contains information related to the distance of the surfaces of scene objects from the depth sensor, usually with 0 denoting the farthest depth value and 255 the nearest. By making use of the depth map, another video sequence from a slightly different viewpoint can thereby be generated. Therefore, the video plus depth format needs less data amount than two/multi videos format and is more suitable when virtual views need to be synthesized.

In 2D to 3D conversion approach, usually a depth map is first inferred based on exploiting the monocular cues in a 2D video sequence, then the same method in the video plus depth sensor approach is used to generate the other video sequence. Since there is so much available content in 2D, this is the fastest way to generate stereoscopic 3D without re-capture. However, it is an ill-problem as automatic conversion is difficult and may cause artifacts. Semiautomatic conversion based on computer vision technologies is often used for best tradeoff between efficiency and quality.

Commercial stereoscopic displays usually support several aforementioned data formats. Whatever the data format is, final two/multi views need to be generated just before displaying. The generation is prerequisite to both the video plus depth format and 2D alone format, and normally displays apply generation algorithms automatically. It may be also optional to two/multi videos, in cases when certain number of views are needed for a particular display or characteristics of two/multi videos needed to be changed to fit the viewing conditions.

2.4.2 Representation/Conversion

Both representation and conversion stages are responsible for formats changing. Specifically, representation stage mainly aims for better coding efficiency and less delivery data, while conversion stage is for format match between the delivered data and display. Furthermore, these two stages should be considered in a systematic way depending on the various available formats of acquisition and visualization stages. For example, if the acquisition and visualization format matches, we may think of avoiding format change to reduce additional artifacts, even though more data maybe required for delivery.

2.4.3 Coding/Decoding

The coding schemes for stereoscopic 3D typically utilize inter-channel similarities of two/multi videos in addition to the temporal and spatial similarities in each channel to further reduce redundancy [98]. Nowadays, there are several available H.264 coding standards for two/multi videos, such as simulcast, stereo supplemental enhancement information (SEI) and multiview video coding (MVC), where the inter-channel similarities are explored to some extents. In simulcast approach, several video sequences are individually and independently encoded, delivered and decoded, which means any 2D video coding standard can be used and inter-channel similarities are not considered at all. However, H.264/AVC SEI message exploits inter-view dependencies of stereo video by inter-field prediction and H.264/AVC MVC exploits temporal as well as inter-view reference pictures for motion- and disparity-compensated prediction, respectively.

In addition, MPEG-C part 3 and H.264/AVC auxiliary picture syntax support for video plus depth format. The former standard specifies a container format for simulcast coding with video plus depth by defining a representation format for depth maps and additional parameters for interpreting the decoded depth values at the receiver side. This allows encoding depth maps as conventional 2D sequences which is independent of the video sequences. The latter standard specifies extra monochrome pictures sent along with the main video stream, but it does not support different coding settings for video and depth. Since the depth map has different characteristics compared with the color video sequences and it is usually more important than the color video sequences, special care should be taken when it is encoded.

Recently, MPEG is working on defining a 3D Video Coding (3DVC) standard [40] for multiview plus depth data format by conducting subjective tests to assess various 3D video compression technologies. With this effort, it is possible to realize the targets defined in [40]: a) enabling stereo devices to cope with varying display types and sizes, and different viewing preferences, including the ability to vary the baseline distance for stereo video, b) facilitating support for high-quality auto-stereoscopic displays by generating many high-quality views from a limited amount of input data, e.g. the video data of 2-3 cameras and additional depth maps.

2.4.4 Delivery

The evolution of 3D video transport technology follows the same path as its 2D counterpart, namely, analog broadcast, digital video broadcasting (e.g. DVB-S, DVB-C, DVB-T, DVB-H) and streaming over the internet protocol (e.g. IPTV) [56]. There is no doubt that streaming over IP provides a more flexible means of delivery where optimization for specific needs of stereoscopic signals is feasible. Some feasible near future scenarios include unicast and multicast transmission, where peer-to-peer, as well as, server-to-client delivery can be considered. However, in the case of IPTV a common problem is burst packet losses. Therefore, congestion control mechanisms should be included in the streaming protocols to reduce packet losses, and resilience and error concealment algorithms should further be used to mitigate the impact on the video if packet losses are inevitable.

3 Stereoscopic 3D Quality Issues

Although the aforementioned principle of stereoscopic 3D sounds straightforward, it is rather difficult to implement in practice. A review of human factors issues that arise when designing stereoscopic displays is provided in [75]. The defects of technology and various signal processing in 3D processing chain [12] lead to an unnatural stereo-pair with artifacts presented to the eyes. Moreover, HVS is quite vulnerable to and not prepared to handle these binocular artifacts (only be perceived when a stereo-pair is displayed but not a single image). This can easily lead to perceptual issues and may further cause nausea and simulator sickness or cyber sickness. However, optometric test variables of viewers seem no clinically significant change after they read words on 3D displays [21]. Most artifacts are reviewed in [8, 62, 52, 118]. In cases that the positive part (naturalness, sense of presence) does not surpasses the negative part (eye strain, visual discomfort, headache) when viewing stereoscopic 3D media, its overall QoE is often not comparable to conventional 2D media.

3.1 Artifacts in the Simplest Stereoscopic System

Viewing stereoscopic 3D media on the simplest stereoscopic system may not be exactly the same as viewing a natural scene. These discrepancies are recognized as artifacts, which may result in degradation of the perceived quality. Those artifacts maybe geometrical artifacts in depth perception (depth-plane curvature caused by keystone, puppet-theater effect and cardboard effect), optic asymmetries (shift, rotation, magnification, blur), filter asymmetries (luminance, color, contrast), temporal asymmetries, static display artifacts (crosstalk, limited comfort zone), dynamic display artifacts (shear distortion, picket fence effect, image flipping). Among which, geometrical artifacts can be introduced in acquisition stage, static and dynamic display artifacts can be introduced in visualization stage, while optic, filter and temporal asymmetries can be introduced both in acquisition and visualization stages. The causes and nature of artifacts will be described in the following.

3.1.1 Geometrical Artifacts

Geometrical artifacts are related to the camera configurations, such as depth plane curvature by keystone and puppet-theater effect may exist only in the toed-in camera configuration, while not in the parallel camera configuration, as concluded in [118, 121]. However, both camera configurations may produce cardboard effect as stated in [121]. The main benefit of the toed-in camera configuration is that

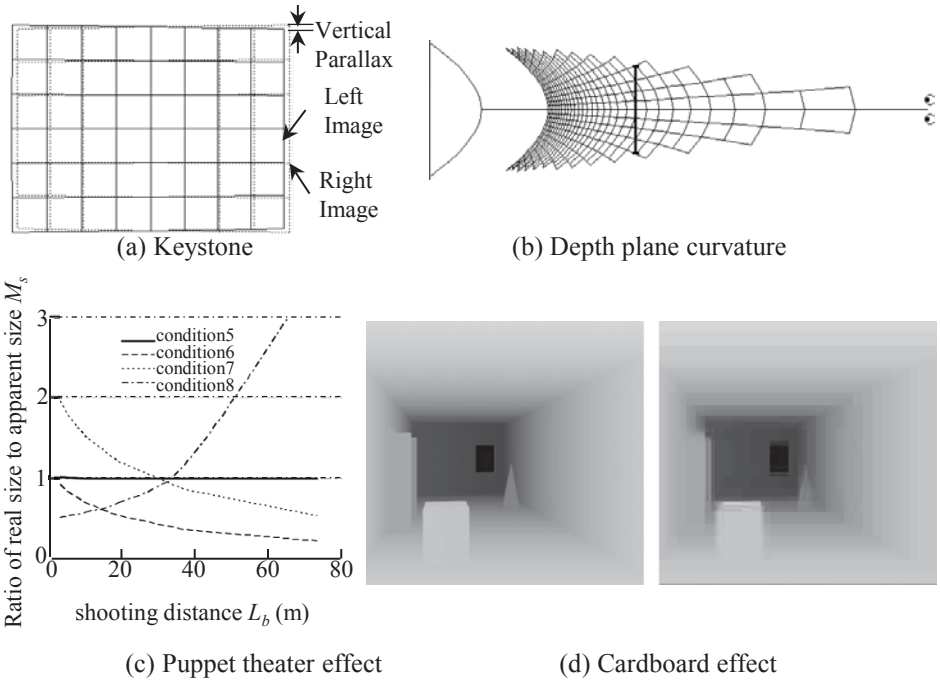


Figure 3.1: Geometrical artifacts. ((a) and (b) are generated by 3D-MAP developed by Andrew Woods, the link is http://www.andrewwoods3d.com/spie93pa.html, (c) is reproduced from Figure 9 in [121], and (d) is generated by the rendering framework developed by Danilo Hollosi, the link is http://sp.cs.tut.fi/mobile3dtv/download/)

it does not require post-production shift or charge-coupled device (CCD) shift to achieve image convergence. However, practically, in case of the toed-in camera configuration, post-productions for correcting the geometrical distortion and vertical disparity are still required to achieve the same quality as the parallel camera configuration. Thus, the parallel camera configuration is used in preference to the toed-in camera configuration nowadays.

Depth plane curvature caused by keystone: Keystone distortion causes vertical parallax and horizontal parallax in the stereoscopic image due to the imaging sensors of the two cameras being located in different planes. As shown in Figure 3.1.(a), in one of the cameras, the image of the grid appears larger on one side than the other, and in the other camera, this effect is reversed. Thus, the amount of vertical parallax is greatest in the corners of the image. The vertical parallax increases with increased camera baseline, decreased convergence distance and decreased focal length [62]. Keystone is the source of depth plane curvature [62, 118], which is a curvature of the depth planes, as shown in Figure 3.1.(b). The depth plane curvature could lead to wrongly perceived relative object distances on the display

that objects at the corners of the image appearing further away from the viewer than objects at the center of the image [118]. It also disturbs image motions during panning of the camera system [118].

Puppet-theater effect: People looks like either larger or smaller animated puppets with puppet-theater effect. This effect is caused by relative size differences that occur among shooting objects inside the image space and arise from the fact that the reproduction magnification of the objects differs between the foreground and background. Some examples of ratio of real size to apparent size of 3D objects shot under different toed in camera configurations are illustrated in Figure 3.1.(c). The authors of [121] showed that orthostereoscopic parallel shooting and display conditions (i.e., simulating human viewing angles, magnification, and convergence in the most natural way possible) do not cause the puppet-theater effect. The author of [31] describes a display technique which reduces the puppet-theater effect. An auto-stereoscopic display with collimation optics enables a large volume of depth and thus allows a larger image to be presented at a greater distance behind the screen so that the puppet-theater effect is not perceptible. Novel display techniques seem to be promising to avoid or reduce the puppet-theater effect which contributes to an unnatural appearance of a 3D image.

Cardboard effect: It is the phenomenon in which the observers of stereoscopic images get the impression that individual objects in the images are flattened like a cardboard, as shown in Figure 3.1.(d). A cardboard effect can be caused by image acquisition parameters (e.g., lens focal length, camera baseline, and convergence distance) or compression parameters resulting in a coarse quantization of disparity or depth values [122, 84]. To avoid or reduce the cardboard effect, camera parameters need to be tuned such that the thickness of objects can be perceived [122], or compression ratio should be adequately low to maintain relative accurate disparity and depth values [84].

3.1.2 Optic, Filter and Temporal Asymmetries

Stereoscopic 3D system requires the acquisition and visualization parameters for left and right images to be same. However, it is impossible to make them exactly match in practice, thus asymmetries arise [52].

Optic asymmetries: The left and right images may differ by misalignment of optics, such as shift, rotation, magnification, and blur. In particular, the misalignment can be left and right cameras in acquisition stage or projectors in visualization stage.

Filter asymmetries: Imperfection of filters and lens in acquisition system or specific visualization system may make the left and right images differ in their luminance, color, sharpness and contrast. For instance, color asymmetries can be introduced by imperfect filter in the camera (e.g. semi-transparent mirrors) or specific 3D visualization technique (e.g. anaglyph glasses).

Temporal asymmetries: The desynchronization in acquisition and visualization stages can induce temporal asymmetries, such as temporal mismatch in recoding or time-sequential stereoscopic display.

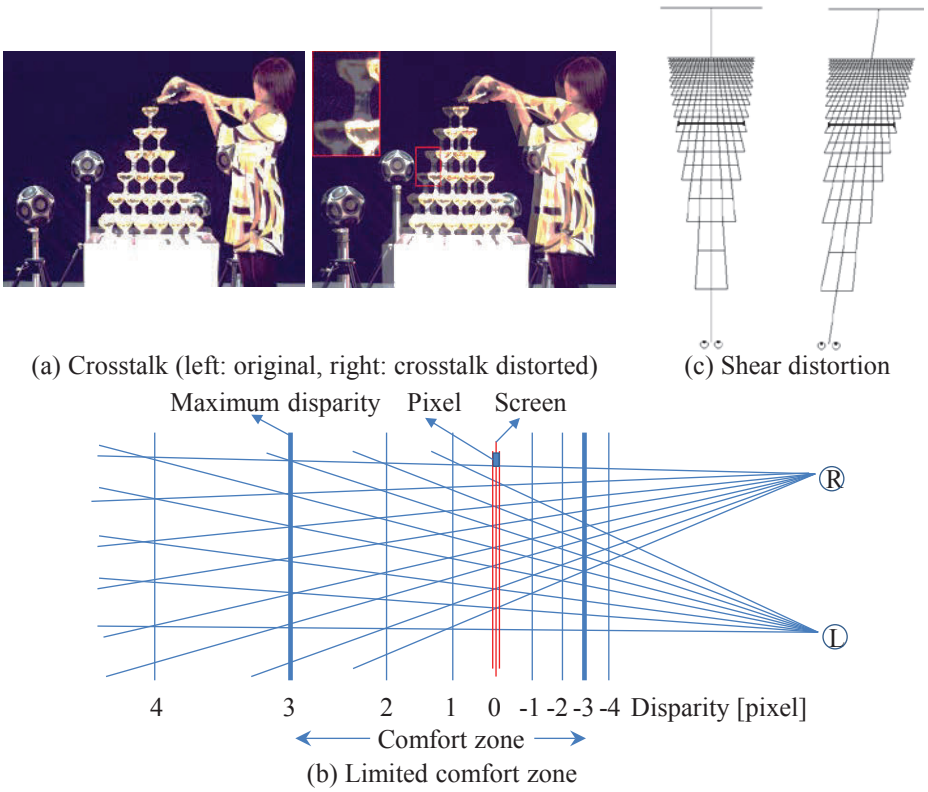


Figure 3.2: Visualization-related artifacts. ((a) is from Paper A, (b) is from Paper C, and (c) is generated by 3D-MAP developed by Andrew Woods, the link is http://www.andrewwoods3d.com/spie93pa.html)

3.1.3 Display Artifacts

The display artifacts are caused by the defects of planar stereoscopic displays, which are not able to simulate HVS completely. The defects include imperfect separation of left and right views, limited comfort zone because of excessive screen disparity and accommodation-convergence mismatch, and particular viewpoint for sweet viewing. Unfortunately, the aforementioned defects cannot be avoided because of the limitations of current planar stereoscopic displays.

Crosstalk: Crosstalk is produced by imperfect view separation that causes a small proportion of one eye's image to be seen by the other eye as well. It is perceived as ghost, shadow, double contours (Figure 3.2.(a)) and even a relative small amount of crosstalk can lead to headaches [72]. Therefore, crosstalk is probably one of the main perceptual factors which degrade image quality and visual comfort [91, 118]. However, the descriptive and mathematical definitions of crosstalk and related terms remain ambiguous in the stereoscopic literature as reviewed in [119].

Crosstalk occurs in various stereoscopic displays, while the mechanisms behind can be significantly different. The author of [117] reviewed the mechanisms behind time-sequential 3D, anaglyph 3D, polarized 3D projection, micro-polarized 3D LCDs, and auto-stereoscopic displays, in order to characterize and measure the components contributing to crosstalk. Take polarized 3D projection for example, the components which affect crosstalk include the optical quality of the polarizers, the projection screen and incorrect orientation of the coding or decoding polarizers. Auto-stereoscopic displays suffer from crosstalk mainly due to the latency in directional lenses to support motion parallax.

Owing to the different mechanisms of various 3D stereoscopic technologies, crosstalk measurement should be designed differently depending on specific 3D mechanism. For example, crosstalk occurs differently in time-sequential 3D LCDs than it does in other displays. Therefore, grey-to-grey crosstalk measurement was recently proposed in [94]. Other crosstalk measurement methods were proposed in [71, 4] for shutter type stereoscopic 3D display. While traditionally by displaying full-black and full-white in opposing eye-channels of the display, black-and-white crosstalk measurement uses an optical sensor to measure the amount of leakage between channels.

Crosstalk reduction can be achieved by reducing the effect of one or more of the above components. Since it is not possible to completely eliminate crosstalk of displays with current technologies, researchers attempted to conceal crosstalk using image processing methods before display [53, 57, 51, 43]. Such methods are usually categorized into crosstalk cancelation. Crosstalk cancelation does not always perform efficiently in all situations. In fact, none of the aforementioned methods can completely eliminate crosstalk artifacts.

In some cases, crosstalk may also have beneficial effect on perceived quality and visual comfort. Some auto-stereoscopic multiview displays intentionally introduce a certain amount of crosstalk to avoid picket fence effect and to minimize image flipping [54]. Small screen disparities limited to the foreground and background regions combined with crosstalk are perceived as blur instead of ghost [97] while perception of depth is still preserved.

Limited comfort zone: The comfort zone of stereoscopic display is limited in a depth range where objects can be reconstructed on a planar screen without inducing visual discomfort, as shown in Figure 3.2.(b). The image located inside the comfort zone remains sharp and can be fused without decoupling of accommodation and convergence. Thus, comfort zone is mainly constrained by screen disparity and accommodation-convergence mismatch.

Screen disparity is a representation of image disparity captured by the left and right cameras, leading to the binocular disparity information in the final visualization. However, screen disparity may not be identical to the original binocular disparity viewed by eyes directly since it depends on acquisition and visualization configurations. Panum's fusional area defines the limits of screen disparities in stereoscopic displays for binocular single view. Specifically, it describes a disparity offset of the whole retinal image of one eye relative to the other, which is absolute screen disparity. Lambooij et al. [54] pointed out that the fusion ability is mainly

determined by the disparity of the fixation objects within the retinal images, which is called relative screen disparity. In other words, the absolute screen disparity can be large, as long as the relative screen disparity are under the fusion limits.

In natural vision, accommodation and convergence are always reflexively coupled, as shown in Figure 3.3.(a). However, for nowadays planar stereoscopic displays, all points in the image focus at the same plane regardless of convergence point, as shown in Figure 3.3.(b). Thus, accommodation-convergence mismatch occurs. In particular, by the convergence-driven accommodation, accommodation may shift away from the display towards the reconstituted object. However, as long as the accommodation shift remains within the DOF, accommodation is able to focus the reconstituted object sharply on the retina [29], thereby still in the zone of clear single binocular vision. Otherwise, negative accommodation feedback stimulates accommodation-driven convergence which results in the convergence away from the reconstituted object, thereby it may cause either loss of fusion resulting in double vision, loss of accommodation resulting in a blurred image or both [54]. As stated in [54], although the accommodation-convergence system can handle the conflict within the DOF, viewers are still under stress and may experience eye strain, visual discomfort or even headache. Moreover, the negative effects may increase with a prolonged viewing. Therefore, the accommodation-convergence rivalry has often been theorized as a significant factor underlying the occurrence of visual discomfort and is widely thought as a major limiting factor for stereoscopic displays.

However the authors of [20] argued that the convergence-accommodation conflict present in current stereoscopic systems. They made a simultaneous measurement of the vergence and accommodation with observers viewing a real scene and its stereoscopic reproduction. They concluded that the accommodation is not fixed at the position of the stereoscopic screen but follows the position of the reconstructed 3D object in the stereoscopic scene. The conflict between accommodation and vergence is related to the DOF of the eye. Some other researchers have the

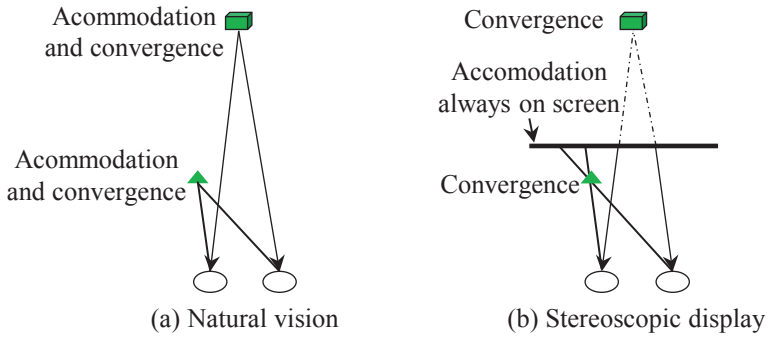


Figure 3.3: Accommodation and convergence in human vision and stereoscopic display.

same arguments. In [32], it was shown that the fixation of accommodation and convergence were almost equal when viewing 2D and 3D images respectively.

Several methods from different view points are used to define the comfort zone by providing thresholds, but lacking of consensus yet. A traditional rule-of-thumb threshold for disparity is a maximum of 70 minutes of arc, which was computed from the human eye's aperture and DOF. This threshold was confirmed in [120]. Threshold from DOF alone was recommended to be ± 0.3 diopters in ITU-R BT.1438 Recommendation [41] and ± 0.2 diopters in [125] in a more conservative way. In [54], with respect to accommodation-convergence thresholds, a DOF of 0.3 diopters and a clear and single zone of 1 degree were defined, respectively. As these two thresholds were reported to resemble each other, 1 degree for disparity was proposed as a general threshold.

Shear distortion, picket fence effect and image flipping: These three artifacts are perceived when observers move their head laterally in front of the display. In particular, shear distortion is typically experienced with stereoscopic displays, while picket fence effect and image flipping are typical multiview auto-stereoscopic display artifacts. However, they can be avoided if head tracking methods are used to update the displaying images in real time.

Shear distortion is aroused because only one correct viewing position is allowed in stereoscopic display [118, 72]. Stereoscopic images appear to follow the observer in a way that images out of the monitor will appear to shear in the direction of the observer, while images behind the surface of the monitor shear in the opposite direction, as shown in Figure 3.2.(c). Shear distortion can result in wrongly perceived relative object distances such that the images on the left view would falsely appear closer than images on the right. This will cause false perception of motion in the image [118].

The picket fence effect is the appearance of vertical banding in an image due to the black mask between columns of pixels in LCD. Image flipping indicates the noticeable transition between viewing zones which leads to discrete views and is experienced as unnatural compared to the continuous parallax experienced in the real world [92]. Display techniques can be improved such that both artifacts are less visible. For instance, in the work [107], a tilted lenticular sheet was put in front of the LCD such that a constant amount of the black mask is always visible. Owing to habituation, an observer actually does not perceive the picket fence effect anymore and image flipping is softened.

3.2 Artifacts in the Complex Stereoscopic System

Due to various data formats, display types, storage and transmission requirements in practice, additional processing stages are needed in the complex stereoscopic system, as shown in Figure 2.7. Artifacts might be introduced because of technical issues in various stages and further result in the degradation of the perceived quality when compared with the simplest stereoscopic system.

3.2.1 Representation/Conversion-related Artifacts

These artifacts are related to data format changes. The depth map based formats (e.g. 2D to 3D conversion, video plus depth) should be converted to two/multi videos format before displaying. Sometimes, two/multi videos format also need to be converted to adapt viewing condition, where a depth map is usually first inferred and then used to generate the synthesis view. Through this process, intrinsic lack of occlusion layer information and the precision of depth map may affect the quality of final synthesis view. The synthesis artifacts are the difference between the original view and synthesis view. In particular, for the occluded areas, inpainting is required, otherwise, disocclusion artifacts can be introduced [8]. These artifacts can be eliminated partially by using layered depth images (LDI) or multiview video plus depth encoding [2, 48]. View synthesis is in progress for FTV in MPEG [25].

3.2.2 Coding/Decoding-related Artifacts

While exploiting spatial, temporal and inter-view redundancies in 3D coding schemes, various 2D and 3D artifacts may be introduced. Typical 2D coding artifacts include blocking, mosaic patterns, staircase effect, ringing, color bleeding, mosquito, etc., which might destroy depth cues and thus impact 3D vision. Specific 3D artifacts include depth bleeding in depth map coding and cross-distortion in asymmetrical two/multi videos coding [8]. In particular, depth bleeding can be mitigated by using structural information of the 2D scene. Cross-distortion can be easily avoided without spatial or temporal sub-sampling of one channel in two/multi video channels, while higher bitrate is usually required. There are different view points regarding asymmetrical coding. On the one hand, the authors of [52] think that cross-distortion induces view asymmetries, thus have potential impact on visual discomfort and visual fatigue. On the other hand, the authors of [60] found that a greater weight is assigned to an un-degraded channel than a degraded one and have no degradation of perceived quality. Therefore, the effect of cross-distortion needs to be thoroughly studied. Special characteristics of depth map and inter-view similarity in two/multi videos, and binocular vision attributes should be explored when designing 3D coding algorithms.

3.2.3 Delivery-related Artifacts

Digital wireless transmissions are subject to packet losses. In DVB-H transmission, burst errors always occur [79], which results in packet losses distributed in tight groups. The presence of artifacts generated in the transmission stage also heavily depends on the employed coding algorithms and how the decoder copes with the transmission errors. With MPEG-4 encoders packet losses might result in propagating or non-propagating errors, depending on both where the errors occurred in respect to previous I-frames and the ratio between I- and P-frames. Resilience and error concealment algorithms may introduce artifacts as well. Artifacts in the tem-

poral domain (e.g. motion blur, display persistence) will affect the motion parallax depth cues.

4 Stereoscopic 3D Quality Evaluation

As introduced in previous chapters, the causes and nature of artifacts and their phenomena have been adequately studied. However, it is still unclear how the artifacts quantitatively affect user perception and QoE under various settings of the simplest stereoscopic system. Thus, quality evaluation in stereoscopic system is an urgent and important issue. The evaluation of perceived quality can be classified into two categories: subjective assessment and objective metrics. Subjective assessment is the most direct and fundamental way to evaluate the quality, in which a number of subjects are asked to watch the test images or videos and to rate their quality. Subjective tests must be carefully designed in order to create significant and reliable results. In addition, the tests should be carried out with certain number of participants. This makes subjective assessment usually time-consuming and unsuitable for real-time applications. To overcome these drawbacks, objective metrics that can predict the human subjects' judgment with high correlation are desired. To develop good objective metrics, the perception mechanisms need to be well studied and taken into account. However, this is usually fairly difficult. This chapter starts with the descriptions of common methodologies for designing subjective assessment and objective metrics in 2D, followed by state-of-the-art overview of subjective assessment and objective metrics in stereoscopic 3D.

4.1 Subjective Assessment

The quality assessment study on 2D image and video is relatively mature, while on emerging 3D image and video is still in its early stages. New characteristics of stereoscopic 3D need to be measured in addition to the picture quality of each view, including both negative aspects (e.g. binocular artifacts, fatigue, eye strain, headache) and positive aspects (e.g. binocular depth, naturalness, sense of presence). In this section, we first review standardized ITU recommendations for evaluating picture quality in 2D, then both ongoing activities towards recommendations and explorative studies for quality assessment on stereoscopic media are presented and discussed.

4.1.1 Recommendations for 2D Quality Assessment

A number of standards or recommendations have been made by the ITU for subjective quality assessment on audio, visual and audio-visual. Among which, ITU-R BT.500-"Methodology for the subjective assessment of the quality of television pictures" is one of the most widely used recommendation. It was first published in

1974, revised several times later and the latest version is ITU-R BT.500-11 published in 2002 [42]. Recommendation ITU-R BT.500 describes in extensive details the methodologies for evaluation of television picture quality, including test environment, test material, test method, and processing of subjective data. With the aforementioned full description of the experiment, a subjective test should be reproducible and experimental results should be more reliable.

Test environment: Different environments with different viewing conditions can affect the experimental results. ITU-R BT.500 distinguishes laboratory and home environment respectively. In particular, the environment luminance (room lighting and chromaticity of background), screen luminance, display brightness and contrast calibration, display resolution, viewing observation angle and viewing distance in both environments are specified.

Test material: The preparation includes selection of source signals, test materials, range of conditions and anchoring.

- Source signals: The source signals provide the reference picture directly and the input for the system under test. It should be of optimum quality for the television standard used. The absence of defects in the reference part of the presentation pair is crucial to obtain stable results.

- Selection of test materials: The number and type of test scenes should be confirmed to address particular assessment problems. New systems frequently have an impact that depends heavily on the scene or sequence content. Thus, the test materials should be selected so as to provide a reasonable generalization to normal programming. Measurement of spatial and temporal perceptual characteristics of a scene can be used to indicate the complexity of the scene.

- Range of conditions and anchoring: Because most assessment methods are sensitive to the variation in the range and distribution of conditions, judgment sessions should include a full range of varying factors or extreme examples as anchors to cover a large quality range.

Test method: Information regarding test observers, instruction for the assessment, grading scale, training session, test session and presentation of test material should be provided. Importantly, several test methods are offered in ITU-R BT.500 for different assessment problems.

- Observers: At least 15 non-expert observers should participate. Prior to a session, they should be screened for visual acuity, color vision and other visual anomalies.

- Instruction of the assessment: Assessors should be carefully introduced with the method of assessment, the types of impairment or quality factors likely to occur, the grading scale and timing. Training sequences demonstrating the range and the type of the impairment to be assessed should be used with scenes other than those used in the test, but of comparable sensitivity.

- Test session: A test session should last up to half an hour. “Dummy presentations” should be introduced to stabilize the observer’s evaluation. If several sessions are necessary, a random order should be used for the presentations. However the test condition order should be arranged so that any effects on the grading of tiredness or adaption are balanced out from session to session.

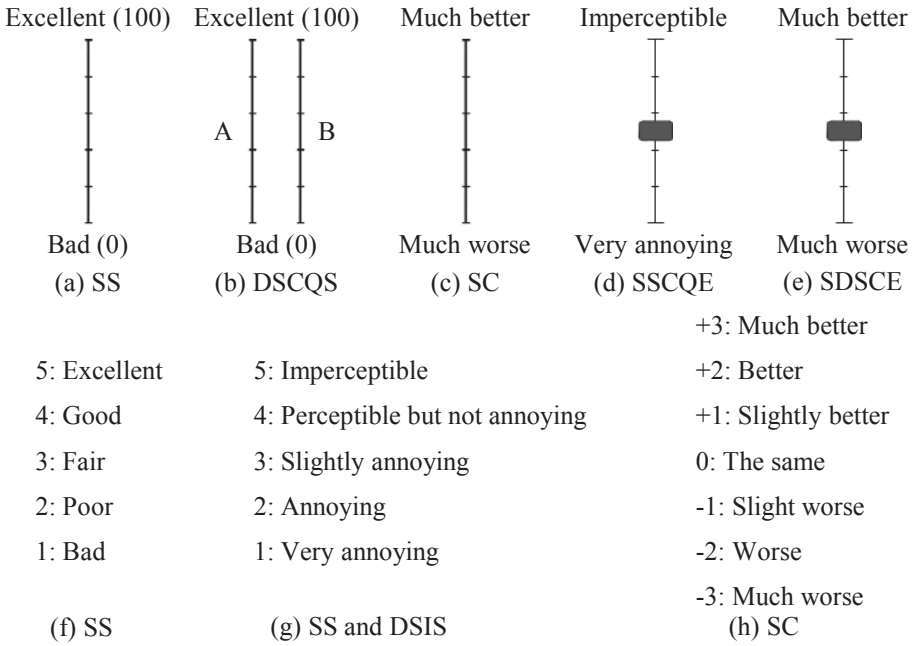
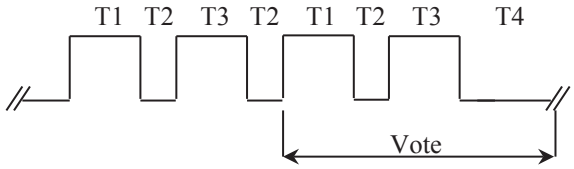


Figure 4.1: Recommended rating scales. Top: non-categorical and numerical. Bottom: categorical.

In general, four different methods are proposed to assess the overall images quality of still images or short video sequences: the double stimulus continuous quality scale method (DSCQS), double stimulus impairment scales (DSIS), single stimulus methods (SS) and stimulus comparison methods (SC). Another two test methods are proposed to assess longer video sequences with time duration from 60 seconds to 20 minutes: single stimulus continuous quality evaluation (SSCQE) and simultaneous double stimulus for continuous evaluation (SDSCE). The recommended rating scales, both non-categorical and categorical, for the above six methods are shown in Figure 4.1.

- DSCQS: Observers assess the overall image quality for a series of image pairs. Each pair consists of an unimpaired (reference) and an impaired image (test) with a length of 10 seconds per image. These two images are presented one by one twice. In the second time of image presentation, observers are asked to rate the overall quality of each image. The presentation structure is illustrated in Figure 4.2 and corresponding non-categorical grating scale is shown in Figure 4.1.(b).

- DSIS: There are two variants to the structure of presentations. The Variant II is the same as the DSCQS that the reference and test pairs are presented twice as is shown in Figure 4.1.(b). While Variant I only presents the reference and test pair once. Thus, Variant II is more time-consuming than variant I, but it is needed when the discrimination of very small impairments is required or moving sequences are



Phases of presentation:
T1=10s Test sequence A
T2=3s Mid-grey produced by a video level of around 200 mV
T3=10s Test sequence B
T4=5-11s Mid-grey

Figure 4.2: Presentation structure of DSCQS and DSIS Variant II according to ITU-R BT.500-11. (Reproduced from Figure 5 in [42])

under test. Both DSIS and DSCQS are double stimulus and similar in presentation structure, but DSIS uses categorical impairment scales as shown in Figure 4.1.(g).

- SS: A single image without a reference is presented and observers assess the overall image quality or image impairment. The rating scales can be either non-categorical (Figure 4.1.(a)) or categorical (Figure 4.1.(f) for image quality and Figure 4.1.(g) for image impairment).

- SC: Similar to SS, SC is single stimulus without a reference. While in SC, the presented stimuli are a series of image pairs, including all possible combination of two images in the stimulus set or just a selected sample of all possible image pairs. Observers are asked to compare the two images for each image pair and assign their relationship by comparison scale as shown in Figure 4.1.(c) or Figure 4.1.(h).

- SSCQE: Observers continuously assess the picture quality of a long video sequence by moving a handset slider as shown in Figure 4.1.(d). SSCQE is used to assess video that contains scene-dependent and time-varying impairments.

- SDSCE: Similar to SSCQE, but SDSCE presents two stimuli at the same time. It is used to compare the quality between the reference and test video sequence. The rating scale is similar as that of SC, but with slider that can be adjusted in real time, as shown in Figure 4.1.(e).

Results presentation: Presentation of results must cover details of the test configuration, details of the test materials, types of picture source and display monitors, number and types of the assessors, reference system used, the grand scores for the experiment, original and adjusted mean opinion scores (MOS) and 95% confidence interval (CI).

- MOS and CI: All mean scores must be associated with a CI which is derived from the standard deviation and the size of each sample [59, 99]. With a probability of 95%, the absolute value of the difference between the experimental MOS and the true MOS (for a very high number of observers) is smaller than the 95% CI, in the condition that the distribution of the individual scores meets certain requirements. CIs for the MOSs are usually calculated using student's t-distribution. The t-

distribution is appropriate when only a small number of samples are available [99]. As the number of observations increases, CI decreases.

- Outlier removal: Before calculation of MOS and CI, it is necessary to screen the observers, as described in Section 2.3 in [42]. In the screening procedure, an expected range of values is calculated for each model. A subject is not rejected for always being above the expected range or always being below the expected range but for being erratic on both sides of the expected range. This procedure is appropriate to reduce the variability of the data in small sample sets. The MOSs and their relative CIs should be calculated from the data excluding the rejected test subjects.

4.1.2 Stereoscopic 3D Quality Assessment

Recommendations: The specification of ITU-R BT.500 does not cover the features of assessing stereoscopic media. For assessing stereoscopic television pictures, ITU-R BT.1438-"Subjective assessment of stereoscopic television pictures" was published in 2000 by ITU [41]. In particular, these test methods in ITU-R BT.1438 are adapted from the ITU-R BT.500 recommendation for conventional 2DTV, but some special stereoscopic 3D characteristics are taken into consideration, such as assessment factors, viewing conditions, observers and test materials.

- Assessment factors: Besides the general factors applied to monoscopic television pictures (e.g. resolution, color rendition, motion portrayal, overall quality, sharpness), new factors peculiar to stereoscopic television system should be added, such as depth resolution, depth motion, puppet-theater effect, cardboard effect.

- Viewing conditions: The display frame effect (e.g. windows violation), inconsistency between accommodation and convergence (maximum value of depth of focus as ± 0.3 diopters) and camera parameters (camera baseline, camera convergence angle, focal length of lens) should be taken into account in determining viewing conditions.

- Observers: Besides vision tests mentioned in ITU-R BT.500, stereopsis test should be used to screen the observers. The test materials for screening observers are recommended as well.

However, ITU-R BT.1438 still lacks of specifications of many new characteristics of stereoscopic 3D and how to access them. The authors of [11] have summarized the lacks in the form of additional requirements. Moreover, there are ongoing activities for stereoscopic 3D quality assessment at ITU-R WP6, ITU-T SG9 and Video Quality Expert Group (VQEG).

Explorative study: Besides the international standardization activities, in the last decade, many explorative studies towards better understanding and assessment of the stereoscopic 3D quality have been done. These studies have been mainly focused on how various acquisition and visualization configurations and artifacts affect the perceptual attributes, such as specific artifact perception, spatial perception, image quality (texture quality and sharpness), perceived depth (amount and quality of depth), visual strain (visual discomfort, eye strain, visual annoyance), naturalness, presence and enjoyment, overall QoE (viewing experience, over-

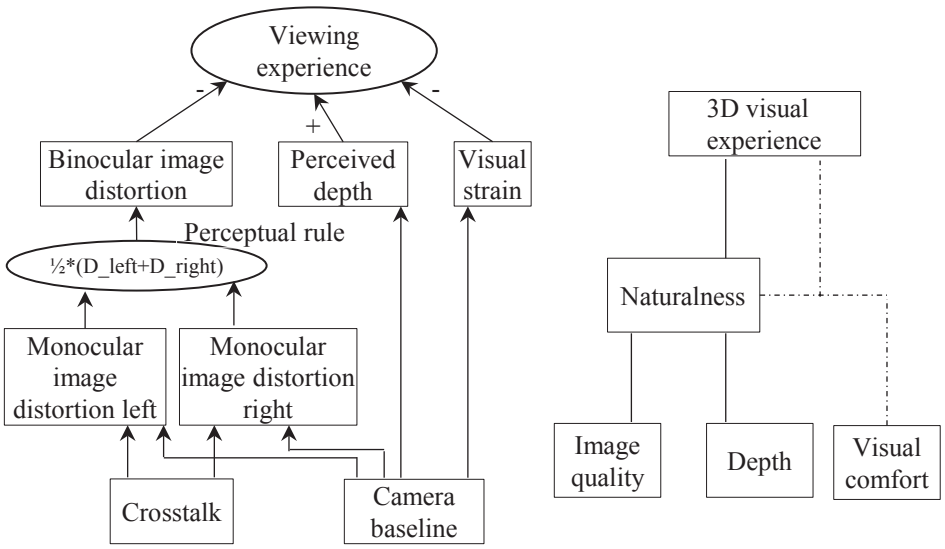


Figure 4.3: 3D viewing experience model. ((a) is reproduced from Figure 7 in [91] and (b) is reproduced from Figure 6.1 in [90])

all image quality, visual experience). Some of the relationship among perceptual attributes are described in the QoE model in Figure 4.3.(a) and Figure 4.3.(b) by Seuntiens [91, 90].

The acquisition and visualization configurations affect the final depth perception and viewing experience. In [91], two natural scenes varying in camera baseline and crosstalk level were investigated to know how they influence perceived image distortion, perceived depth, and visual strain. The same author [89] also investigated the influence of camera baseline and JPEG compression ratio on overall image quality, perceived depth, perceived sharpness and perceived eye-strain. In [23], a 3D video database with varying scene contents and camera baselines and corresponding MOS regarding overall image quality was built. In [37], the effects of stereoscopic filming parameters (camera baseline, convergence distance, and focal length) and display duration on observers' judgements of naturalness and quality of stereoscopic images were investigated. In [38], the investigated factors were image motion, stereoscopic presentation and screen size, and the measured perceptual attributes were presence, vection, involvement, and sickness, as well as observers' lateral postural responses. In [46], a series of pairs of stimuli with varying distance were shown, the subjects were asked to decide which one is closer in each pair of stimuli, and the results were analyzed to understand the the perceptual distance and the constancy of size and stereoscopic depth. The authors of [73] tried to determine the relationship between (supra)threshold perception for position offset and stereoscopic depth perception under conditions (increasing the interline gap and dioptric blur) that elevate their

Table 4.1: Explorative studies on acquisition and visualization configurations

Investigated factor	**Perceptual attribute**	**Test method**	**Ref.**
scene content, camera baseline, crosstalk level	perceived image distortion, perceived depth, visual strain	SS	[91]
scene content, camera baseline, compression ratio	overall image quality, perceived depth, perceived sharpness, perceived eyestrain	SS	[89]
scene content, camera baseline	overall image quality	SS	[23]
filming parameters, display duration	naturalness, quality of depth	SS	[37]
image motion, screen size, stereoscopic presentation	presence, vection, involvement, sickness, lateral postural responses	SS	[38]
objects distance	perceptual distance, the constancy of size, stereoscopic depth	SC	[46]
interline gap, dioptric blur	(supra)threshold perception for position offset and stereoscopic depth perception	SC	[73]
camera setting	puppet-theater and cardboard effects	SS	[121]
scene content, camera baseline, screen size, **viewing position**	QoE	SS	Paper C, Paper D

respective thresholds. The authors of [121] conducted a subjective experiment to compare the impact of camera settings to the puppet-theater and cardboard effects.

In Table 4.1 a summary of the studies presented regarding acquisition and visualization configurations is presented. While these studies strengthen our knowledge about stereoscopic perception mechanism, a comprehensive understanding of how the influence factors in the simplest stereoscopic system, namely, requisite factors, affect stereoscopic QoE is still missing. Although in [118] the grid patterns distorted by the stereoscopic display system were visualized for a rectilinear grid in front of the camera system under various settings of most requisite factors, no subjective tests were conducted. Therefore, influence of requisite factors (scene content, camera baseline, screen size and viewing position) on the perceived quality for human subjects was evaluated quantitatively in our Paper C and Paper D.

Table 4.2: Explorative studies on specific artifact: Crosstalk

Investigated factor	**Perceptual attribute**	**Test method**	**Ref.**
contrast, disparity	crosstalk visibility	SS	[72]
contrast, disparity	crosstalk perception	SS	[110]
scene content, camera baseline, crosstalk level	perceived image distortion, perceived depth, visual strain	SS	[91]
monocular cues, contrast ratio, disparity	crosstalk perception	SS	[35]
crosstalk level	image quality indicator, ghost image	SS	[34]
edges, contrast	crosstalk perception	SS	[58]
blur, vertical disparity	crosstalk perception	SS	[52]
scene content, camera baseline, crosstalk level	crosstalk perception	SS	Paper A, Paper i
scene content, crosstalk level, viewing position	crosstalk perception	SS	Paper B

In the aforementioned studies, some of them focus on the technique parameters regarding to specific artifact, such as crosstalk level [91] or compression ratio [89]. In addition, many studies have been done on perception of specific artifact and its impact on final viewing experience.

Crosstalk perception is an active research area, since crosstalk is probably one of the most annoying distortions in 3D displays. Some of the research are listed in Table 4.2. The author of [72] demonstrated that the annoyance of crosstalk increases with increasing contrast and disparity values in gray scale patches. The author suggested that the crosstalk of a display should not cross a threshold of 0.3%. In [110], images consisting of a single character and varying in contrast and disparity were computer-generated to measure crosstalk perception and an analytical formula was further proposed to predict crosstalk perception. As mentioned earlier, the authors of [91] investigated more realistic scenarios where natural scenes varying in crosstalk levels and camera baselines affect the perceptual attributes of crosstalk. However, only two, rather similar, natural scenes were used in their experiments. Moreover, the authors of [35] found out that monocular cues of images also play an important role in the crosstalk perception, in addition to contrast ratio and disparity. Later, they [34] studied the factors of stereo-images with different crosstalk levels that may affect stereopsis. In [58], it was shown that edges and high contrast of computer-generated wire-frames make crosstalk more visible when compared to natural images. This means that crosstalk can be more efficiently concealed on images with more texture or details.

These observations partially suggest a hypothesis that scene content is an important factor impacting user perception of crosstalk. Therefore, scene content,

together with other three requisite factors (camera baseline, crosstalk level, viewing position) were investigated for their impact on crosstalk perception in Paper A, Paper B and Paper i. Although other artifacts, e.g. blur and vertical disparity as investigated in [52], may also have impact on the crosstalk perception, they can be often corrected by post-processing techniques. Moreover, crosstalk itself may further impact other perceptual attributes, such as perceived image quality [45], viewing experience [91], task performance and workload [69].

In addition to crosstalk, stereoscopic 3D displays also suffer from their limited comfort zones, which are closely related to the visual discomfort and visual fatigue. Table 4.3 lists some of the work related to limited comfort zone. It was confirmed in [105] that at increasing screen disparities beyond 1 degree, the oculomotor system operates under increasing stress to preserve fusion and provide sharply focused images. In [30], stereoscopic stimuli were presented with various convergence and accommodation distances, from which two-thirds of the distances were conflicting (ranging from 0.33 diopter to 1.33 diopters). A questionnaire following an orientation detection task significantly indicated more visual discomfort for conflicting stimuli than for the nonconflicting ones. The study in [67] verified that stereoscopic stills with large parts of the images perceived beyond the DOF, received much lower scores in terms of visual comfort in contrast to stereoscopic stills perceived within the DOF. In [125], changes of accommodation and convergence were performed to evaluate subjective fatigue level after 1 hours of stereoscopic content viewing. The authors of [19] proposed that the change on fusional amplitude and accommodation response is a valid indicator of visual fatigue. It was found in [96] that negative

Table 4.3: Explorative studies on specific artifact: Limited comfort zone

Investigated factor	**Perceptual attribute**	**Test method**	**Ref.**
screen disparity	oculomotor system operates under stress	SS	[105]
various convergence and accommodation distances	visual discomfort	questions	[30]
parallax distribution	visual comfort, sense of presence	SC	[67]
changes of accommodation and convergence	visual fatigue	SS	[125]
fusional amplitude, accommodation response	visual fatigue	SS	[19]
sign of accommodation and convergence, viewing distance	discomfort, fatigue	SS	[96]
stimuli in various range of comfort zone	QoE (visual discomfort indirectly)	SS	Paper C, Paper D

accommodation-convergence conflicts (stereo content behind the screen) are less comfortable at far distances and that positive conflicts (content in front of screen) are less comfortable at near distances.

Visual discomfort and visual fatigue were indirectly studied in Paper C and Paper D, since different settings of requisite factors in stereoscopic system inevitably make test stimuli in various range of comfort zone and further impact the QoE which is a concept containing all the aspects of stereoscopic viewing, including visual strain. Specifically, the most important factors of QoE should include the most negative factors (crosstalk, visual strain) and most positive factor (binocular depth), which can be summarized from Figure 4.3.(a) and Figure 4.3.(b).

Stereoscopic coding artifacts and their impact on quality have been allocated much attention to in standardization organizations and researchers, summarized in Table 4.4. As mentioned earlier, in [89], by applying four compression ratios of JPEG coding to both the left and right stereo image separately, possible symmetric

Table 4.4: Explorative studies on specific artifact: Coding artifact

Investigated factor	**Perceptual attribute**	**Test method**	**Ref.**
compression ratio, scene content, camera baseline	overall image quality, perceived depth, sharpness and eye-strain	SS	[89]
eye dominance, asymmetric view coding	3D viewing experience	SS	[44]
asymmetric compression ratio	overall image quality, depth perception	SS	[28]
blurring, JPEG compression, JPEG2000 compression, JPEG & JPEG2000 compression	overall image quality	SAMVIQ	[9]
JPEG, JPEG2000 compression, blur filtering	overall image quality	SAMVIQ	[6]
gaussian blurring, JPEG compression, JPEG2000 compression, white noise	QoE	DSCQS	[127]
depth image compression	3D experience	DSCQS	[55]
color and depth image compression	3D experience	SC	[104]
texture and depth image compression	QoE	SC	[3]
scene content, AVC and HEVC in different coding bit rates	viewing experience	SS	Paper F

and asymmetric coding combinations were formed. Their effects on image quality, sharpness, depth and eye strain were investigated. In [44], bounds of an asymmetric stereo view compression scheme by H.264/AVC and their relationship with eye-dominance were examined based on user study. In [28], subjective tests were performed to determine the overall image quality and depth perception of a range of asymmetrically coded video sequences (two/multi videos and video plus depth) using different quantization parameters (QP) in joint scalable video model (JSVM). In [9], distortions caused by blurring, JPEG compression, JPEG2000 compression, JPEG & JPEG2000 compression were added to both images of the stereo pair, and assessed by the subjects. In [6], subjective tests on JPEG, JPEG2000 compression, and blur filtering added stereo pair were conducted. In [127], four types of distortions (gaussian blurring, JPEG compression, JPEG2000 compression and white noise) were applied to right eye images, while the left eye images were kept undistorted, followed by a subjective quality experiment.

In addition to the two/multi videos format, there are some quality evaluations on video plus depth format as well. In [55], the authors investigated the impact of depth image quality and compression on the perceived 3D experience. They found that motion and complexity of the depth image have strong influence on the acceptable depth quality in 3D videos. The results indicated that depth image can be compressed significantly, up to 0.0007 bits per pixel, without affecting 3D perception significantly. In [104], different bit budgets for the color video and the depth were tested, followed by a small scale subjective tests supplement the objective measurements on virtual view. The results showed that for similar overall quality numbers, observers favorably trade off lower depth quality for higher color quality and that depth distortions are perceived but considered less significant than the color distortions. In [3], four subjective experiments were designed with coding artifacts by using different QPs.

In Paper F, subjective tests were launched among 13 international test laboratories for evaluating the merits of proposed 3D data formats and associated compression technology and view synthesis algorithm. The test stimuli were two classes (1920×1088p 25fps and 1024×768p 30fps, 4 scene contents for each), two scenarios (2-view input configuration and 3-view input configuration), two coding categories (AVC and HEVC, in 4 different bit rates). Synthesized views were produced using a view synthesis algorithm for all sequences in all classes and all test scenarios, based on the decoded output and displayed in stereoscopic display (Hyundai S465D 46") and auto-stereoscopic display (Dimenco BDL5231V3D 52") respectively for final quality evaluation by subjects. The results were presented in two ways: MOS and CI values for each proponent and MOS versus the target bit rate.

Furthermore, with the increasing demand on sharing datasets (both test stimuli and MOS scores), several stereoscopic 3D quality datasets are freely available. For instance, LIVE released a 3D image quality database phase I [63] containing symmetrically distorted stimuli by compression using the JPEG and JPEG2000 compression standards, additive white gaussian noise, gaussian blur and a fast-fading model based on the Rayleigh fading channel. They plan a phase II release with both symmetrically and asymmetrically distorted stimuli with the same distortion

types. In the dataset track at QoMEX 2012, the authors of [106] published datasets with coding and spatial degradations, such as block-based coding, wavelet coding, resolution reduction, and edge enhancement algorithms. Besides, they reviewed the availability of 3D sequences and assessed the variety of the selected sequences in terms of spatial, temporal and depth. However, all the aforementioned datasets are for coding degraded stereoscopic 3D stimuli. To best of our knowledge, [23] is the only free available dataset for acquisition and visualization degraded stereoscopic 3D stimuli. Specifically, the influence of camera baseline on the perceived 3D image and video quality was studied. Furthermore, in addition to camera baseline, other requisite factors (e.g. scene content, crosstalk level, screen size, viewing position) in the simplest stereoscopic system were investigated for their relationships to crosstalk perception and QoE in our work and released to public through Paper v.

4.2 Objective Metrics

After obtaining the ground truth MOS scores in subjective assessment, objective metrics are usually designed to automatically predict the perceived image and video quality for three purposes [112]: a) monitoring image quality for quality control systems, b) benchmarking image and video processing systems and algorithms, c) optimizing the algorithms and the parameter settings of an image and video processing system. In conventional 2D imaging systems, image quality models have been proposed to predict 2D image quality. However, the principles of modeling 2D image quality can be used to obtain insight into modeling stereoscopic 3D quality. In this section, we will first review 2D metrics for evaluating the picture quality, and then introduce pilot work on 3D quality metrics.

4.2.1 2D Metrics

Depending on the availability of required information about the original image and video signals, objective 2D metrics can be classified into three categories [112]: full-reference (FR), reduced-reference (RR) and no-reference (NR). Most of the proposed objective quality metrics in the literature are FR, which assume that the undistorted reference signal is fully available. However, the reference images or video sequences are often not accessible in many practical video service applications. Thus, NR which evaluates image and video quality blindly is highly desirable. However, proposing NR metric turns out to be a very difficult task, although human observers can effectively and reliably assess the quality of distorted image or video without using any reference. There exists a third type of image quality assessment method RR, in which the original image or video signal is not fully available. Instead, certain features are extracted from the original signal and transmitted to the quality assessment system as side information to help evaluating the quality of the distorted image or video.

Currently, there are no widely-recognized reliable objective metrics for image and video quality assessment because of both the complexity of image and video systems and HVS, and the lack of standardization. Pixel-based metrics such as mean-squared error (MSE) or peak signal-to-noise ratio (PSNR) are the most widely used FR metrics. They are simple to understand, easy to compute, but not correlative with perceived quality very well.

Pixel-based metrics: These metrics are based on error sensitivity, which assume the loss of quality is directly related to the strength of the error signal. The MSE is the mean of the squared differences between the gray-level values of pixels in two pictures, which is defined as follows

$$MSE = \frac{1}{M * N} \sum_{y=1}^{M} \sum_{x=1}^{N} [IM_r(x, y) - IM_d(x, y)]^2 \tag{4.1}$$

where IM_r and IM_d are the pictures to compare, reference image and distorted image, respectively, M and N are the dimensions of the image in the horizontal and vertical directions, x and y are the pixel index.

PSNR is defined as:

$$PSNR = 10 \log_{10} \frac{L^2}{MSE} \tag{4.2}$$

where L is the dynamic range of the pixel values. For an 8 bits/pixel monotonic signal, L is often set to 255.

Technically, MSE measures image difference, whereas PSNR measures image fidelity, i.e. how closely an image resembles a reference image, usually the uncorrupted original one. The interpretation is that the larger PSNR the better the quality of the distorted image IM_d; that is, the closer the distorted image IM_d is to the original reference image IM_r [95].

More complicated and reliable objective metrics have been designed according to the following two approaches [115]: psychophysical approach and engineering approach. In particular, psychophysical approach basically models various factors of the HVS which are essential for visual perception, such as frequency selectivity, contrast and orientation sensitivity, spatial and temporal masking effects, color perception and so on. Since HVS is complex, those metrics are often very complex and computationally expensive, but usually correlate very well with human perception and are usable in wide range of applications. However, engineering approach is primarily based on extraction and analysis of certain features or artifacts in the image/video, such as structural elements (e.g. contours) or specific artifacts that are introduced by a particular compression technology or transmission link. It does not mean that such metrics disregard human vision, as they often consider psychophysical effects as well. However, image analysis rather than fundamental vision modeling is the conceptual basis for their design. Engineering-based metrics usually involve a lower computational complexity than psychophysical models.

Psychophysical approach: Properties of HVS are simulated in psychophysical approach. A generic block diagram of HVS based metrics [115] is illustrated in Figure 4.4.

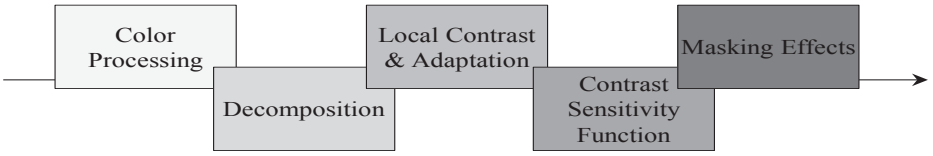


Figure 4.4: Block-diagram of a typical HVS-model. (Reproduced from Figure 5.2 in [115])

- Color processing: The first stage in the processing chain of HVS-models concerns the transformation of image into an adequate perceptual color space, usually based on opponent colors. After this step the image is represented by one achromatic and two chromatic channels carrying color difference information. This stage can also take care of the so-called luminance masking or lightness nonlinearity [87], the non-linear perception of luminance by the HVS. Such nonlinearity is inherent to more sophisticated color spaces such as the Lab color space, but needs to be added to simple linear color spaces.

- Multi-channel decomposition: It is widely accepted that the HVS bases its perception on multiple channels which are tuned to different ranges of spatial frequencies and orientations. Measurements of the receptive fields of simple cells in the primary visual cortex revealed that these channels exhibit approximately a dyadic structure [15]. This behavior is well matched by a multi-resolution filter bank or a wavelet decomposition. An example for the former is the cortex transform [113], a flexible multi-resolution pyramid, whose filters can be adjusted within a broad range. Wavelet transforms on the other hand offer the advantage that they can be implemented in a computationally efficient manner by a lifting scheme [14]. It is believed that there are also a number of channels processing different object velocities or temporal frequencies. These include one temporal low-pass and one, possibly two, temporal band-pass mechanisms in the human visual system [27, 22], which are generally referred as sustained and transient channels, respectively.

- Contrast and adaptation: The response of the HVS depends much less on the absolute luminance than on the relation of its local variations to the surrounding background, a property known as Weber-Fechner law [87]. Contrast is a measure of this relative variation, which is commonly used in vision models. While it is quite simple to define a contrast measure for elementary patterns, it is very difficult to model human contrast perception in complex images, as it varies with the local image content [77, 78, 116]. Furthermore, the adaptation to a specific luminance level or color can influence the perceived contrast.

- Contrast sensitivity: One of the most important issues in HVS-modeling concerns the decreasing sensitivity to higher spatial frequencies. This phenomenon is parameterized by the contrast sensitivity function (CSF). The correct modeling of CSF is especially difficult for color images. Typically separability between color

and pattern sensitivity is assumed, so that a separate CSF for each channel of the color space needs to be determined and implemented. Achromatic CSFs were summarized in [5], and color CSF measurements were described in [33, 26, 64]. Take the contrast masking properties for example, it can be integrated into PSNR in a way described in [80]. The human contrast sensitivity also depends on the temporal frequency of the stimuli. Similar to the spatial CSF, the temporal CSF has low-pass and slightly band-pass shape. The interaction between spatial and temporal frequencies can be described by spatio-temporal contrast sensitivity functions, which are commonly used in vision models for video [13]. For easier implementation, they may be approximated by combinations of components separable in space and time [47, 124].

- Masking: It occurs when a stimulus that is visible by itself cannot be perceived due to the presence of another. Sometimes the opposite effect, facilitation, occurs: a stimulus that is not visible by itself can be perceived due to the presence of another. Within the framework of image processing it is helpful to consider the distortion or coding noise being masked (or facilitated) by the original image or sequence serving as background. Masking explains why similar distortions are more disturbing in certain regions. Several different types of spatial masking can be distinguished [50, 65, 114], but this distinction is not clear-cut. The terms contrast masking, edge masking, and texture masking are often used to describe masking due to strong local contrast, edges, and local activity, respectively. Temporal masking is a brief elevation of visibility thresholds due to temporal discontinuities in intensity, e.g., at scene cuts [93]. It can occur not only after a discontinuity, but also before [1].

- Pooling: It is believed that the information represented in various channels within the primary visual cortex is integrated in the subsequent brain areas. This process can be simulated by gathering the data from these channels according to rules of probability or vector summation, also known as pooling. However, little is known about the nature of the actual integration taking place in the brain, and there is no firm experimental evidence that these rules are a good description of the pooling mechanism in the human visual system [81, 22, 61]. This summation is often carried out over all dimensions in order to obtain a single distortion rating for an image or video, but in principle any subset of dimensions can be used depending on what kind of results are desired. For example, pooling over pixel locations may be omitted to produce a distortion map for every frame of an image while they are hardly noticeable elsewhere.

All aforementioned psychophysical HVS features can be used to develop objective metrics. Taking the contrast masking properties for example, it can be integrated into PSNR in a way described in [80].

Engineering approach: It is based on the extraction and analysis of high-level content features (structural elements) or distortions (blockiness, blur, etc) that can arise in the impaired signal but do not belong to the original reference signal. SSIM [111] is one of the most famous metrics in the engineering approach. It is based on the hypothesis that the HVS is highly adapted for extracting structural information

from the content of a still image or image sequence. It assumes that degradation of still images or image sequences equals to perceived structural information variation.

In particular, the structural similarity measure is constructed based on the comparisons of three components: luminance, contrast, and structure, between an original image, which is supposed to have perfect quality, and its distorted version. Supposing that x and y are two image signals, the luminance comparison function $l(x, y)$ is defined based on the comparison of the mean intensities (μ_x and μ_y) of two images, as follows in (4.3).

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{4.3}$$

in which the constant $C_1 = (K_1L)^2$, and $K_1 \ll 1$ is used to avoid instability when $\mu_x^2 + \mu_y^2$ is very close to zero. L is the dynamic range of the pixel values (e.g. 255 for 8-bit grayscale images). The contrast comparison function $c(x, y)$ takes a similar form, based on the standard deviation of the two signals, δ_x and δ_y,

$$c(x, y) = \frac{2\delta_x\delta_y + C_2}{\delta_x^2 + \delta_y^2 + C_2} \tag{4.4}$$

Again, the constant $C_2 = (K_2L)^2$, and $K_2 \ll 1$ is included to avoid instability when $\delta_x^2 + \delta_y^2$ is very close to zero. The third component, structure comparison function $s(x, y)$, is defined as:

$$s(x, y) = \frac{\delta_{xy} + C_3}{\delta_x\delta_y + C_3} \tag{4.5}$$

To avoid instability when $\delta_x\delta_y$ is very close to zero, a constant C_3 is incorporated. The general form of the SSIM index between signals x and y is:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\delta_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\delta_x^2 + \delta_y^2 + C_2)} \tag{4.6}$$

In practice, one usually requires a single overall quality measure of the entire image. A mean SSIM (MSSIM) index is used to evaluate the overall image quality:

$$MSSIM(x, y) = \frac{1}{M}\sum_{j=1}^{M} SSIM(x, y) \tag{4.7}$$

Objective quality metric validation: The goal of objective quality assessment is to design algorithms whose quality prediction is in good agreement with subjective scores obtained from human observers. There are different attributes that characterize an objective quality model in terms of its prediction performance with respect to MOS [109]. Three of these attributes are average difference, accuracy and monotonicity which are explained as follows.

- Average difference: The difference per pixel is averaged and given by the root mean squared error (RMSE):

$$RMSE = \sqrt{MSE} \tag{4.8}$$

- Accuracy: It is the ability of a metric to predict subjective ratings with minimum average error and can be determined by means of the Pearson linear correlation coefficient. For a set of N data pairs (x_i, y_i), it is defined as follows:

$$Pearson = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}} \tag{4.9}$$

where $\overline{x}$ and $\overline{y}$ are the means of the respective objective and subjective data. This assumes a linear relation between the data sets, which may not be the case. Therefore, in this thesis psychometric functions will be used to take into account the HVS behavior such as saturation effects. Then, linear correlation will be used to obtain relative comparisons between subjective and objective data.

- Monotonicity: It measures if increases (decreases) in one variable are associated with increases (decreases) in the other variable, independently of the magnitude of the increase (decrease). Ideally, differences of a metric's rating between two sequences should always have the same sign as the differences between the corresponding subjective ratings. The degree of monotonicity can be quantified by the Spearman rank-order correlation coefficient, which is defined as:

$$Spearman = \frac{\sum_{i=1}^{N}(\chi_i - \overline{\chi})(\gamma_i - \overline{\gamma})}{\sqrt{\sum_{i=1}^{N}(\chi_i - \overline{\chi})^2}\sqrt{\sum_{i=1}^{N}(\gamma_i - \overline{\gamma})^2}} \tag{4.10}$$

where χ_i is the rank of x_i and γ_i is the rank of y_i in the ordered data series. $\overline{\chi}$ and $\overline{\gamma}$ are the respective mid-ranks. The Spearman rank-order correlation makes no assumption about the relationship between x_i and y_i.

4.2.2 Stereoscopic 3D Metrics

2D image quality models are not adequate to measure 3D visual experience since various 3D characteristics are not incorporated. These 3D characteristics include at least the most positive factor (binocular depth) and negative factors (crosstalk, visual discomfort) in stereoscopic 3D. Thus, a 3D visual experience model (Figure 4.3) should be multidimensional incorporating aforementioned factors, allowing for a weighting of the factors based on perceptual importance. In fact, an effective quality metric can not be proposed without a deep understanding of the perception mechanism of stereoscopic 3D presentations, and the development of objective 3D quality models is still in its early stages.

Researchers first started with exploring whether or not traditional 2D metrics can be applied to stereoscopic quality assessment with coding artifacts [28, 9, 127]. In [28], PSNR/SSIM/VQM scores of the coded color video, as well as average

PSNR/SSIM/VQM of the rendered left and right views generated using the coded color plus depth sequences, were used in predicting perceived quality attributes of 3D. It was shown that the VQM scores and average VQM have higher correlation than others. In [9], each of 2D video objective quality metrics (SSIM, UQI, C4 and RRIQA) for left and right images was combined using average approach, main eye approach, or visual acuity approach for the quality assessment of the stereo image. No significant performance difference was observed between the three approaches. It was also noticed that among the tested metrics, RRIQA is the metric that better represents the perceived degradation of the stereo pair because of blurring. The C4 metric is the metric showing the best performance for evaluating the perceived distortion on the stereo pair. However, none of the metrics performs acceptably for all kinds of distortions. The authors of [127] further introduced some well-known 2D image quality metrics (PSNR, SSIM, MSSIM, VSNR, VIF, UQI, IFC, NQM, WSNR, PHVS, JND) and investigated their capabilities in stereoscopic image quality assessments.

As one of the most important attributes of stereoscopic images, disparity is often combined into 2D metrics for stereoscopic 3D quality metric. In [6], quality metrics for the assessment of stereopairs using the fusion (globally and locally) of 2D quality metrics (SSIM, C4) and of the disparity information were proposed. It was shown that C4 alone and enhanced SSIM with local disparity distortion measure have high correlation with the MOSs. In [127], a study on integration of the disparity information to the aforementioned well-known 2D image metrics in three ways (global correlation coefficient (GCC), mean square error (MSE), and mean absolute difference (MAD)) was presented. The experimental results demonstrated that better performance can be achieved if the disparity information and original images are combined appropriately in the stereoscopic image quality assessment.

Subsequently, a few objective metrics that take into account the characteristics of stereoscopic images have been proposed. The authors of [123] found out that the noise added to the relatively large absolute disparity affects the stereo sense more. Therefore, a metric called stereo sense assessment (SSA) based on the disparity distribution condition was proposed. The authors of [24] used the new stereo band limited contrast(SBLC) algorithm to rank stereoscopic pairs in terms of image quality. SBLC accounts for HVS sensitivity to contrast and luminance changes at regions of high spatial frequency. In [83], it was assumed that perceived distortion and depth of any stereoscopic image are strongly dependent on the local features, such as edge, flat and texture. Therefore, the blockiness and zero crossing rate within the block of the images were evaluated for artifacts and disparity, and finally integrated into a metric. In [88], the authors made use of the observation that serious difference between views in edge regions causes eye fatigue [100]. Thus, blocking artifacts and degradation in edge regions were detected using conventional video quality assessment models. In addition, it detected video quality difference between views using disparity information. The weights of three items were further obtained for the final evaluation. In [7], a cyclopean image was used for assessing the trivial monoscopic perceived distortions caused by blur, noise, contrast change etc., and a perceptual disparity map and a stereo-similarity map were defined for

assessing the perceived degradation of binocular depth cues. These maps were derived in a mulstiscale manner and measured by SSIM in each scale. Monoscopic and stereoscopic quality measurements for different scales were combined in one compound Qm and Qs separately. Coding scheme's distribution of artifacts at different depths within the stereoscopic images were modeled in a single metric [68].

However, most of the existing objective metrics as aforementioned are designed to assess quality degradations caused by lossy compression schemes. To the best of our knowledge, only one objective metric that considers non-compression quality degradations, introduced during acquisition and visualization stages of stereoscopic system, was proposed in [49]. This metric is modeled by a linear combination of three measurements, which evaluate the perceived depth, visual fatigue and temporal consistency, respectively. Based on our conducted subjective assessment regarding crosstalk perception and QoE and findings to their perceptual attributes, we further proposed objective metrics for crosstalk perception (Paper A and Paper ii) and QoE (Paper C, Paper E, Paper iii and Paper iv).

5 Thesis Contributions

A total of 19 research papers Paper A-F and Paper i-xiii have been published during the doctoral program. The relationship of all 10 first-author research papers (Paper A-E and Paper i-v) and 1 co-authorship research paper (paper F) are depicted in Figure 1.1. In this thesis Paper A-F are included.

It can be seen from Figure 1.1 that the focus is mainly on the quality issues in the simplest stereoscopic system. They are more urgent and important when compared to the quality issues in the complex stereoscopic systems. In fact, the simplest stereoscopic system is the basic and essential part of the complex system such that the quality issues in the simplest stereoscopic system continuously exist and are crucial in the complex system. In particular, crosstalk artifact which can not be eliminated completely in the simplest system was measured (Paper A, Paper B, Paper i) and modeled (Paper A, Paper ii), respectively, in top block of Figure 1.1. The impact of requisite factors in the simplest stereoscopic system to QoE was studied in Paper C and Paper D, then objective metrics were proposed in Paper C, Paper E, Paper iii and Paper iv, as shown in middle block of Figure 1.1. While the Paper v in left block of Figure 1.1 is the datasets of our conducted subjective tests in Paper A, Paper B and Paper D. The bottom block of Figure 1.1 only includes Paper F where coding artifacts in the complex stereoscopic systems were investigated by subjective assessment. The abstracts of the included papers in the thesis are first listed, then their contributions are further summarized.

5.1 Abstract of Included Papers

Paper A. Stereoscopic 3D services do not always prevail when compared to their 2D counterparts, though the former can provide more immersive experience with the help of binocular depth. Various specific 3D artifacts might cause discomfort and severely degrade the QoE. In this paper, we analyze one of the most annoying artifacts in the visualization stage of stereoscopic imaging, namely, crosstalk, by conducting extensive subjective quality tests. A statistical analysis of the subjective scores reveals that both scene content and camera baseline have significant impacts on crosstalk perception, in addition to crosstalk level itself. Based on the observed visual variations during changes in significant factors, three perceptual attributes of crosstalk are summarized as the sensorial results of the HVS. These are shadow degree, separation distance and spatial position of crosstalk. They are classified into two categories: 2D and 3D perceptual attributes, which can be described by a SSIM map and a fil-

tered depth map, respectively. An objective quality metric for predicting crosstalk perception is then proposed by combining the two maps. The experimental results demonstrate that the proposed metric has a high correlation (over 88%) when compared to subjective quality scores in a wide variety of situations.

Paper B. Crosstalk is one of the most annoying distortions in the visualization stage of stereoscopic systems. Specifically, both pattern and amount of crosstalk in multiview auto-stereoscopic displays are more complex because of viewing angle dependability, when compared to crosstalk in 2-view stereoscopic displays. Regarding system crosstalk there are objective measures to assess it in auto-stereoscopic displays. However, in addition to system crosstalk, crosstalk perceived by users is also impacted by scene content. Moreover, some crosstalk is arguably beneficial in auto-stereoscopic displays. Therefore, in this paper, we further assess how crosstalk is perceived by users with various scene contents and different viewing positions using auto-stereoscopic displays. In particular, the proposed subjective crosstalk assessment methodology is realistic without restriction of the users viewing behavior and is not limited to the specific technique used in auto-stereoscopic displays. The test was performed on a slanted parallax barrier based auto-stereoscopic display. The subjective crosstalk assessment results show their consistence to the system crosstalk meanwhile more scene content and viewing position related crosstalk perception information is provided. This knowledge can be used to design new crosstalk perception metrics.

Paper C. Stereoscopic 3D services are becoming more and more popular recently due to their capability to provide richer QoE to the end-users. In practice, stereoscopic QoE can be influenced by a complex combination of different factors. In this work, we focus on minimum stereoscopic system (including capturing and displaying stages only) and its requisite factors. A subjective stereoscopic quality assessment is conducted to investigate the influence of several requisite factors, including scene content, camera baseline, screen size and viewing position, on stereoscopic QoE. Moreover, crosstalk level is recognized as another requisite factor in the minimum stereoscopic system based on my previous work on crosstalk assessment. Thereafter, two perceptual attributes of stereoscopic QoE, namely, crosstalk perception and depth enabled visual comfort, are summarized as the sensorial results of the HVS. Their relationships to the requisite factors are explored and modeled in equations, respectively. These equations of perceptual attributes are further combined into an objective quality metric for predicting stereoscopic QoE. The experimental results demonstrate that the proposed metric has a high correlation (over 85%) when compared with subjective quality scores in a wide variety of situations.

Paper D. The stereoscopic 3D industry has fallen short of achieving acceptable QoE because of various technical limitations, such as excessive disparity, accommodation-convergence mismatch. This study investigates the effect of scene content, camera baseline, screen size and viewing location on stereoscopic QoE in a holistic approach. 240 typical test configurations are taken into account, in which a wide range of disparity constructed from the shooting conditions (scene content, camera baseline, sensor resolution/screen size) was selected from datasets, making the constructed disparities locate in different ranges of maximal disparity supported by viewing environment (viewing location). Second, an extensive subjective test is conducted using a single stimulus methodology, in which 15 samples at each viewing location were obtained. Finally, a statistical analysis is performed and the results reveal that scene content, camera baseline, as well as the interactions between screen size, scene content and camera baseline, have significant impact on QoE in stereoscopic images, while other factors, especially viewing location involved, have almost no significant impact. The generated MOS and the statistical results can be used to design stereoscopic quality metrics and validate their performance.

Paper E. Stereoscopic QoE is the result of a complex combination of different influencing factors. Previously we had investigated the effect of factors such as scene content, camera baseline, screen size and viewing position on stereoscopic QoE using subjective tests. In this paper, we propose two objective metrics for predicting stereoscopic QoE using bottom-up and top-down approaches, respectively. Specifically, the bottom-up metric is based on characterizing the significant factors of QoE directly, which are scene content, camera baseline, screen size and crosstalk level. While the top-down metric interprets QoE from its perceptual attributes, including crosstalk perception and perceived depth. These perceptual attributes are modeled by their individual relationship with the significant factors and then combined linearly to build the top-down metric. Both proposed metrics have been validated against our own database and a publicly available database, showing a high correlation (over 86%) with the subjective scores of stereoscopic QoE.

Paper F. Subjective quality assessment is widely used to understand and to study human perception of multimedia quality and as a basis for developing objective metrics to automatically predict the quality of audiovisual presentations. There are several recognized international protocols and procedures for reliable assessment of quality in multimedia systems and services, with emphasis on speech, audio and video modalities. However, the aspect of certification is not yet well understood in this context. This paper discusses various issues regarding certification of multimedia quality assessment. To be concrete, the discussion is illustrated by the procedure implemented to assess 3D video compression technologies within the MPEG effort for the definition of a 3D video coding standard. Selected

results from four laboratories, Acreo, EPFL, NTNU and UBC, which participated in the assessment are presented. This case study is used in an early attempt to define a process for certification of subjective test campaigns, based on a cross-validation of the test results across different laboratories, towards the ultimate goal of QoE certification.

5.2 Summary of Contributions

The main contributions of the thesis include the understanding, measurement (subjective tests) and eventually, modeling and prediction of (objective metrics) stereoscopic 3D quality. In particular, subjective assessment have been conducted regarding crosstalk perception and QoE under different requisite factors of the simplest stereoscopic system, since crosstalk is probably one of the most annoying distortions in 3D display and QoE contains all the aspects of stereoscopic viewing, in addition to crosstalk perception. Furthermore, an objective metric for crosstalk perception was first proposed by our understanding on its perceptual attributes, namely shadow degree, separation distance and spatial position of crosstalk. Then, by combining crosstalk perception with other perceptual attributes of QoE (binocular depth and visual discomfort), viewing experience was predicted by our proposed objective metric for QoE. A further step on quality assessment of 3DVC coding artifacts was conducted in the complex stereoscopic system. The detailed contributions of each included paper are summarized in the following.

Paper A.

- Subjective tests were conducted for crosstalk perception on stereoscopic images on polarized display, and consequently, a comprehensive database was created. Specifically, we followed the test methodologies in recommendations [42, 41, 11] and further customized them for crosstalk perception. For instance, measurement of system-introduced crosstalk was also included in the test environment, and stereo acuity vision was tested for the participants. The test stimuli for crosstalk perception varied in scene contents, camera baselines and crosstalk levels. Subjective scores were obtained in test sessions by SS method after the training sessions.
- Perceptual attributes of crosstalk were summarized as the sensorial results of the HVS, including shadow degree, separation distance and spatial position of crosstalk. The former two belong to 2D perceptual attributes, while the last one is 3D perceptual attribute. These perceptual attributes bridge the gap between low-level significant factors and high-level user perception on crosstalk.
- Subjective user perception of crosstalk is predicted using an objective metric based on a rigorous analysis of perceptual attributes of crosstalk. In particular, the 2D and 3D perceptual attributes were described by a SSIM map and a filtered depth map, respectively, followed by an objective quality metric combining these two maps.

Paper B.
- Methodologies for subjective crosstalk assessment in auto-stereoscopic displays were proposed. The proposed methodology is realistic without restriction on user viewing behaviors and is not limited to the specific technique used in auto-stereoscopic displays. Specifically, a head and score tracking system was developed for supporting the aforementioned features. A study case of crosstalk measurement on slanted parallax barrier based auto-stereoscopic display was performed, with various scene contents and arbitrary viewing positions in a specified area of 2m×2m. The subjective crosstalk assessment results show their consistence to the system crosstalk. Meanwhile, more scene content and viewing position related crosstalk perception information are provided. The results can be further used to design new crosstalk perception metrics.

Paper C.
- Influence of requisite factors in the simplest stereoscopic system to QoE was studied in subjective tests. In addition to system-introduced crosstalk level, the investigated requisite factors included scene content, camera baseline, screen size and viewing position. A comprehensive database with MOS for QoE under the aforementioned requisite factors was created.
- The perceptual attributes for QoE were identified, including crosstalk perception and depth enabled visual comfort. In particular, depth enabled visual comfort tok both binocular depth and visual discomfort into consideration, based on the understanding that both the ratio of stimuli disparity located in the comfort zone of the display system and their disparity amplitude reflects the stereoscopic QoE.
- An objective QoE metric was thereby proposed by modeling these perceptual attributes respectively and combining them linearly thereafter. Specifically, the crosstalk perception model was adopted from Paper A, while depth enabled visual comfort model was newly proposed in this paper. The experimental results demonstrate that the proposed metric has a high correlation when compared with MOS both in our own subjective database and a public available database.

Paper D.
- It is the basis for the Paper C regarding the subjective assessments. The impact of requisite factors, namely scene content, camera baseline, screen size and viewing position, on QoE were investigated in a holistic approach. Moreover, significant factors influencing QoE were identified by a statistical analysis to subjective scores obtained in the subjective tests.

Paper E.
- It is the basis for the Paper C regarding the objective metrics. A bottom-up metric was proposed based on the similarity of significant factors between disparity and QoE, while a top-down metric was proposed by interpreting QoE from its perceptual attributes,

including crosstalk perception and perceived depth. Although they exhibit similar performance, the top-down metric which is more understandable from perceptual view point was further adopted and improved in Paper C.

Paper F.

- Towards certification of 3D video quality assessment. A case study of assessing 3D video compression technologies within the MPEG effort for standardization of 3D video coding techniques was adopted for defining a process for certification of subjective test campaigns. The comprehensive database enables us a possibility to extent the previous QoE metric to include the compression artifacts in the coding stage as well.

6 Conclusion and Possible Future Work

This thesis contributes to the field of quality evaluation on stereoscopic 3D media in three aspects.

First, quality evaluation of crosstalk perception was carried out on polarized stereoscopic display, since crosstalk is one of the most annoying artifacts in the visualization stage of stereoscopic 3D and can not be completely eliminated with current technologies. The subjective tests were customized for crosstalk perception with varying independent parameters of scene content, camera baseline and crosstalk level. An objective metric for crosstalk perception was proposed based on our findings of perceptual attributes of crosstalk perception. Furthermore, subjective crosstalk assessment methodologies for auto-stereoscopic displays at arbitrary viewing positions in a specified area were suggested, supported by a head and score tracking system. Future work includes generalization of the proposed crosstalk metric to other types of stereoscopic displays, and the obtained subjective scores of the auto-stereoscopic display can be used for this purpose.

Second, an extension from crosstalk perception to QoE in the simplest stereoscopic system was studied, since QoE is often referred as a criterion for the acceptance of any commercial system and determines the success or not. In addition to crosstalk level, other requisite factors of the simplest stereoscopic system, including scene content, camera baseline, screen size and viewing position have also been investigated on their relationships to perceptual attributes of QoE. Specifically, those perceptual attributes are crosstalk perception and depth enabled visual comfort, which cover the main negative and positive aspects of stereoscopic QoE. By modeling these perceptual attributes separately and combining them thereafter, an objective QoE metric was proposed. Analysis of the obtained subjective scores and understanding the perceptual attributes should be continued to further propose updated metric with better performance.

Third, further work on the complex stereoscopic system was carried out by investigating the influence of coding artifacts on QoE. This work was done under the context of assessing 3D video compression technologies within MPEG's effort for standardizing 3D video coding techniques. However, the subjective scores can be used as ground truth dataset for proposing QoE model which will incorporate both the coding artifacts and configurations of the complex stereoscopic system. Meanwhile, since the tests were conducted at 13 test laboratories around the world with large amount of test sessions, it can be used for defining a process for certification of subjective test campaigns. Issues such as tolerable levels of variation among the

test scores and its relationship to the complexity of the test and factors involved should be further investigated for future certification.

Although this thesis and many other efforts has been dedicated to the quality assessment on stereoscopic 3D media, the current research is still at an early stage when compared to either the speech/audio quality assessment or 2D visual quality assessment. There are several potential directions for future research on stereoscopic 3D quality assessment [18, 36]. For the subjective 3D quality assessment, the directions include: a) standardized protocols for stereoscopic 3D quality evaluation are needed, considering 3D perception, source of binocular distortions, display characteristics, viewing conditions, different dimensions of perceived video quality in 3D and so on, b) user centered evaluation is needed to characterize various quality factors for 3D video, c) visual fatigue and motion sickness need long-term studies, d) quality evaluation should be conducted in realistic usage scenarios with relevant content. For objective quality metrics, the trends are: a) ground truth subjective dataset reflecting the essence of 3D QoE should be created, such as quantifying the influence of 3D distortions originating from every step within the whole processing chain, incorporating information about scene content and system configuration, b) it is not the signal itself (which is the case in 2D for quality prediction) but rather the rendered version should be analyzed in stereoscopic 3D because 3D video presents significantly different quality issues that are not encountered in 2D, c) more accurate models for 3D human visual perception are needed, d) interaction between monocular and binocular depth cues needs to be considered, e) convergence-accommodation conflict and focus of attention need to be considered.

References

[1] A. J. Ahumada Jr., B. L. Beard, and R. Eriksson. Spatiotemporal discrimination model predicts temporal masking functions. In *Human Vision and Electronic Imaging III*, volume 3299, pages 120–127, 1998.

[2] A. Alatan, Y. Yemez, U. Gudukbay, X. Zabulis, K. Muller, C. Erdem, C. Weigel, and A. Smolic. Scene representation technologies for 3DTV - a survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(11):1587–1605, 2007.

[3] M. Barkowsky, R. Cousseau, and P. Le Callet. Influence of depth rendering on the quality of experience for an autostereoscopic display. In *Quality of Multimedia Experience (QoMEX), 2009 International Workshop on*, pages 192–197, 2009.

[4] M. Barkowsky, S. Tourancheau, K. Brunnström, K. Wang, and B. Andrén. Crosstalk measurements of shutter glasses 3D displays. *SID Symposium Digest of Technical Papers*, 42(1):812–815, 2011.

[5] P. G. J. Barten. *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*. SPIE, 1999.

[6] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau. Quality assessment of stereoscopic images. *EURASIP Journal on Image and Video Processing*, 2008(1):659024, 2008.

[7] A. Boev, A. Gotchev, K. Egiazarian, A. Aksay, and G. Akar. Towards compound stereo-video quality metric: a specific encoder-based framework. In *Image Analysis and Interpretation, 2006 IEEE Southwest Symposium on*, pages 218–222, 2006.

[8] A. Boev, D. Hollosi, A. Gotchev, and K. Egiazarian. Classification and simulation of stereoscopic artifacts in mobile 3DTV content. pages 72371F–72371F–12, 2009.

[9] P. Campisi, P. L. Callet, and E. Marini. Stereoscopic images quality assessment. *EURASIP Journal on Image and Video Processing*, 2008, 2008.

[10] D. Chandler. Visual perception (introductory notes for media theory students). In *MSC portal site, University of Wales, Aberystwyth*, [available] http://www.aber.ac.uk/media/sections/image05.html.

[11] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet. New requirements of subjective video quality assessment methodologies for 3DTV. In *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2010.

[12] S. Daly, R. Held, and D. Hoffman. Perceptual issues in stereoscopic signal processing. *Broadcasting, IEEE Transactions on*, 57(2):347–361, 2011.

[13] S. J. Daly. Engineering observations from spatiovelocity and spatiotemporal visual models. In *Human Vision and Electronic Imaging III*, volume 3299, pages 180–191, 1998.

[14] I. Daubechies and W. Sweldens. Factoring wavelet transforms into lifting steps. *J. Fourier Anal. Appl*, 4:247–269, 1998.

[15] J. G. Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20(10):847–856, 1980.

[16] N. A. Dodgson. Variation and extrema of human interpupillary distance. In *Stereoscopic Displays and Virtual Reality Systems XI*, volume 5291, pages 36–46, 2004.

[17] N. A. Dodgson. Multi-view autostereoscopic 3D display. In *Stanford Workshop on 3D Imaging*, pages 1–42, 2011.

[18] T. Ebrahimi. Towards 3D visual quality assessment for future multimedia. In *keynote presentation at CORESA 2012*, [available] http://www-rech.telecom-lille1.eu/coresa2012/.

[19] M. Emoto, Y. Nojiri, and F. Okano. Changes in fusional vergence limit and its hysteresis after viewing stereoscopic TV. *Displays*, 25(2-3):67–76, 2004.

[20] K. Fliegel, S. Vítek, T. Jindra, P. Páta, and M. Klíma. Comparison of stereoscopic technologies in various configurations. In *Applications of Digital Image Processing XXXV*, volume 8499, pages 849929–849929–9, 2012.

[21] M. F. Fortuin, M. T. Lambooij, W. A. IJsselsteijn, I. Heynderickx, D. F. Edgar, and B. J. Evans. An exploration of the initial effects of stereoscopic displays on optometric parameters. *Ophthalmic and Physiological Optics*, 31(1):33–44, 2011.

[22] R. Fredericksen and R. Hess. Estimating multiple temporal mechanisms in human vision. *Vision Research*, 38(7):1023–1040, 1998.

[23] L. Goldmann, F. De Simone, and T. Ebrahimi. A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. In *Three-Dimensional Image Processing (3DIP) and Applications*, volume 7526, 2010. [available] http://mmspg.epfl.ch/3diqa, http://mmspg.epfl.ch/3dvqa.

[24] P. Gorley and N. Holliman. Stereoscopic image quality metrics and compression. In *Stereoscopic Displays and Applications XIX*, volume 6803, pages 680305–680305–12, 2008.

[25] M. Gotfryd, K. Wegner, and M. Domański. View synthesis software and assessment of its performance. *ISO/IEC JTC1/SC29/WG11 MPEG2008/M15672*, 2008.

[26] E. M. Granger and J. C. Heurtley. Visual chromaticity-modulation transfer function. *J. Opt. Soc. Am.*, 63(9):1173–1174, 1973.

[27] R. Hess and R. Snowden. Temporal properties of human visual filters: number, shapes and spatial covariation. *Vision Research*, 32(1):47–59, 1992.

[28] C. Hewage, S. Worrall, S. Dogan, and A. Kondoz. Prediction of stereoscopic video quality using objective quality models of 2D video. *Electronics Letters*, 44(16):963–965, 2008.

[29] N. Hiruma and T. Fukuda. Accommodation response to binocular stereoscopic TV images and their viewing conditions. *SMPTE J.*, 102:1137–1144, 1993.

[30] D. M. Hoffman, A. R. Girshick, K. Akeley, and S. Banks. Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *J. Vision J.*, 8:1–30, 2008.

[31] K. Hopf. An autostereoscopic display providing comfortable viewing conditions and a high degree of telepresence. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(3):359–365, 2000.

[32] H. Hori, T. Shiomi, T. Kanda, A. Hasegawa, H. Ishio, Y. Matsuura, M. Omori, H. Takada, S. Hasegawa, and M. Miyao. Comparison of accommodation and convergence by simultaneous measurements during 2D and 3D vision gaze. In *2011 international conference on Virtual and mixed reality: new trends - Volume Part I*, pages 306–314, Berlin, Heidelberg, 2011. Springer-Verlag.

[33] G. J. C. V. D. Horst and M. A. Boumran. Spatiotemporal chromaticity discrimination. *J. Opt. Soc. Am.*, 59(11):1482–1488, 1969.

[34] K.-C. Huang, J.-C. Yang, C.-L. Wu, K. Lee, and S.-L. Hwang. System-crosstalk effect on stereopsis human factor study for 3D displays. pages 75240U–75240U–8, 2010.

[35] K.-C. Huang, J.-C. Yuan, C.-H. Tsai, W.-J. Hsueh, and N.-Y. Wang. A study of how crosstalk affects stereopsis in stereoscopic displays. In *Stereoscopic Displays and Virtual Reality Systems X*, volume 5006, pages 247–253, 2003.

[36] Q. Huynh-Thu, P. Le Callet, and M. Barkowsky. Video quality assessment: From 2D to 3D - challenges and future trends. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 4025–4028, 2010.

[37] W. IJsselsteijn, H. de Ridder, and J. Vliegen. Subjective evaluation of stereoscopic images: effects of camera parameters and display duration. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(2):225–233, 2000.

[38] W. IJsselsteijn, H. d. Ridder, J. Freeman, S. E. Avons, and D. Bouwhuis. Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence. *Presence: Teleoper. Virtual Environ.*, 10(3):298–311, 2001.

[39] W. IJsselsteijn, P. Seuntiens, and L. Meesters. Human factors of 3D displays. In *3D Video Communication*, 2005.

[40] ISO/IEC JTC1/SC29/WG11 MPEG2011/N12036. Call for proposal on 3D video coding technology. In *Video and Requirement Group*, 2011.

[41] ITU-R BT.1438. Subjective assessment of stereoscopic television pictures. 2000.

[42] ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television pictures. 2002.

[43] A. Jain and J. Konrad. Crosstalk in automultiscopic 3D displays: blessing in disguise? In *Stereoscopic Displays and Virtual Reality Systems XIV*, volume 6490, pages 649012–649012–12, 2007.

[44] H. Kalva, L. Christodoulou, L. M. Mayron, O. Marques, and B. Furht. Design and evaluation of a 3D video system based on H.264 view coding. In *2006 international workshop on Network and operating systems support for digital audio and video*, NOSSDAV '06, pages 12:1–12:6, New York, NY, USA, 2006. ACM.

[45] R. Kaptein and I. Heynderickx. Effect of crosstalk in multi-view autostereoscopic 3D displays on perceived image quality. *SID Symposium Digest of Technical Papers*, 38(1):1220–1223, 2007.

[46] L. Kaufman, J. Kaufman, R. Noble, S. Edlund, S. Bai, and T. King. Perceptual distance and the constancy of size and stereoscopic depth. *Spatial Vision*, 19(5):439–457, 2006.

[47] D. H. Kelly. Spatiotemporal variation of chromatic and achromatic contrast thresholds. *J. Opt. Soc. Am.*, 73(6):742–749, 1983.

[48] P. Kerbiriou, T. Colleu, K. Mueller, R. K. Gunnewiek, R. V. D. Vleuten, S. Relier, and O. Grau. Comparative study and recommendations. In *ICT-215075 3D4YOU*, 2010.

[49] D. Kim, D. Min, J. Oh, S. Jeon, and K. Sohn. Depth map quality metric for three-dimensional video. In *Stereoscopic Displays and Applications XX*, volume 7237, pages 723719–723719–9, 2009.

[50] S. A. Klein, T. Carney, L. Barghout-Stein, and C. W. Tyler. Seven models of masking. In *Human Vision and Electronic Imaging II*, volume 3016, pages 13–24, 1997.

[51] J. Konrad, B. Lacotte, S. Member, and E. Dubois. Cancellation of image crosstalk in time-sequential displays of stereoscopic video. In *IEEE Transactions on Image Processing*, pages 897–908, 2000.

[52] F. L. Kooi and A. Toet. Visual comfort of binocular and 3D displays. *Displays*, 25:99–108, 2004.

[53] A. Krupev and A. Popova. Ghosting reduction and estimation in anaglyph stereoscopic images. In *Signal Processing and Information Technology, 2008 IEEE International Symposium on*, pages 375–379, 2008.

[54] M. Lambooij, M. Fortuin, I. Heynderickx, and W. IJsselsteijn. Visual discomfort and visual fatigue of stereoscopic displays: A review. *Journal of Imaging Science and Technology*, 53(3):30201–1–30201–14, 2009.

[55] G. Leon, H. Kalva, and B. Furht. 3D video quality evaluation with depth quality variations. In *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, pages 301–304, 2008.

[56] O. Levent. 3D video technologies: An overview of research trends. In *SPIE Electronic Imaging*, volume PM196, pages 1–116, 2011.

[57] J. S. Lipscomb and W. L. Wooten. Reducing crosstalk between stereoscopic views. In *Stereoscopic Displays and Virtual Reality Systems*, volume 2177, pages 92–96, 1994.

[58] L. Lipton. Factors affecting 'ghosting' in time-multiplexed plano-stereoscopic CRT display systems. In *True 3D Imaging Techniques and Display Technologies*, volume 761, pages 95–78, 1987.

[59] S. E. Maxwell and H. Delaney. *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Routledge Academic, 1989.

[60] D. V. Meegan, L. B. Stelmach, and W. J. Tam. Unequal weighting of monocular inputs in binocular combination: implications for the compression of stereoscopic imagery. In *Journal Experimental Psychology: Applied*, volume 7, pages 143–153, 2001.

[61] T. S. Meese and W. C. B. Probability summation for multiple patches of luminance modulation. *Vision Research*, 40(16):2101–2113, 2000.

[62] L. Meesters, W. IJsselsteijn, and P. Seuntiens. A survey of perceptual evaluations and requirements of three-dimensional tv. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(3):381–391, 2004.

[63] A. K. Moorthy, C.-C. Su, A. Mittal, and A. C. Bovik. Subjective evaluation of stereoscopic image quality. *Signal Processing: Image Communication*, 2012. [available] http://live.ece.utexas.edu/research/quality/live_3dimage_phase1.html.

[64] K. T. Mullen. The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *The Journal of Physiology*, 359(1):381–400, 1985.

[65] M. J. Nadenau, J. Reichel, and M. Kunt. Performance comparison of masking models based on a new psychovisual test method with natural scenery stimuli. *Signal Processing: Image Communication*, 17(10):807–823, 2002.

[66] S. Nagata. The binocular fusion of human vision on stereoscopic displays - field of view and environment effects. *Ergonomics*, 39(11):1273–1284, 1996. PMID: 8888639.

[67] 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 *Stereoscopic Displays and Virtual Reality Systems X*, volume 5006, pages 195–205, 2003.

[68] R. Olsson and M. Sjostrom. A depth dependent quality metric for evaluation of coded integral imaging based 3D-images. In *3DTV Conference, 2007*, pages 1–4, 2007.

[69] S. Pala, R. Stevens, and P. Surman. Optical cross-talk and visual comfort of a stereoscopic display used in a real-time application. pages 649011–649011–12, 2007.

[70] S. E. Palmer. *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge, Massachussetts, 1999.

[71] C.-C. Pan, Y.-R. Lee, K.-F. Huang, and T.-C. Huang. Crosstalk evaluation of shuttertype stereoscopic 3D display. *SID Symposium Digest of Technical Papers*, 41(1):128–131, 2010.

[72] S. Pastoor. Human factors of 3D images: Results of recent research at Heinrich-Hertz-Institut Berlin. In *International Display Workshop*, volume 3, pages 69–72, 1995.

[73] S. Patel, H. Bedell, D. Tsang, and M. Ukwade. Relationship between threshold and suprathreshold perception of position and stereoscopic depth. In *J. Opt. Soc. Am. A Opt. Image Sci. Vis.*, volume 26, pages 847–861, 2009.

[74] R. Patterson. Human factors of 3D displays. *Journal of the Society for Information Display*, 15(11):861–871, 2007.

[75] R. Patterson. Invited paper: Human factors of stereoscopic displays. *SID Symposium Digest of Technical Papers*, 40(1):805–807, 2009.

[76] R. Patterson and W. L. Martin. Human stereopsis. In *Hum Factors*, volume 34, pages 669–692, 1992.

[77] E. Peli. Contrast in complex images. *J. Opt. Soc. Am. A*, 7(10):2032–2040, 1990.

[78] E. Peli. In search of a contrast metric: Matching the perceived contrast of gabor patches at different phases and bandwidths. *Vision Res.*, 37(23):3217–3224, 1997.

[79] J. Poikonen and J. Paavola. Error models for the transport stream packet channel in the DVB-H link layer. In *Communications, 2006 IEEE International Conference on*, volume 4, pages 1861–1866, 2006.

[80] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin. On between-coefficient contrast masking of DCT basis functions. In *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2007.

[81] J. Quick, R.F. A vector-magnitude model of contrast detection. *Kybernetik*, 16:65–67, 1974.

[82] W. Richards. Stereopsis and stereoblindness. In *Experimental Brain Research*, volume 10, pages 380–388, 1970.

[83] Z. Sazzad, S. Yamanaka, Y. Kawayokeita, and Y. Horita. Stereoscopic image quality prediction. In *Quality of Multimedia Experience (QoMEX), 2009 International Workshop on*, pages 180–185, 2009.

[84] A. Schertz. Source coding of stereoscopic television pictures. In *Image Processing and its Applications, 1992., International Conference on*, pages 462–464, 1992.

[85] C. M. Schor and C. W. Tylera. Spatio-temporal properties of panum's fusional area. *Vision Research*, 21(5):683–692, 1981.

[86] C. M. Schor and I. Woods. Disparity range for local stereopsis as a function of luminance spatial frequency. In *Hum Factors*, volume 23, pages 1649–1654, 1983.

[87] W. Schreiber. In *Fundamentals of Electronic Imaging Systems*, page 332. New York: Springer, 1993.

[88] J. Seo, D. Kim, and K. Sohn. Compressed stereoscopic video quality metric. In *Stereoscopic Displays and Applications XX*, volume 7237, pages 723712–723712–11, 2009.

[89] P. J. H. Seuntiens, L. M. J. Meesters, and W. A. IJsselsteijn. Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation. *TAP*, 3(2):95–109, 2006.

[90] P. Seuntiëns. Visual experience of 3DTV. *Doctoral thesis, Eindhoven University of Technology.*, 2006.

[91] P. Seuntiëns, L. Meesters, and W. IJsselsteijn. Perceptual attributes of crosstalk in 3D images. *Displays*, 26(4–5):177–183, 2005.

[92] I. Sexton and P. Surman. Stereoscopic and autostereoscopic display systems. *Signal Processing Magazine, IEEE*, 16(3):85–99, 1999.

[93] A. Seyler and Z. Budrikis. Detail perception after scene changes in television image presentations. *Information Theory, IEEE Transactions on*, 11(1):31–43, 1965.

[94] S. Shestak, D. Kim, and S. Hwang. Measuring of graytogray crosstalk in a LCD based timesequential stereoscopic display. *SID Symposium Digest of Technical Papers*, 41(1):132–135, 2010.

[95] Y. Q. Shi and H. Sun. In *Image and Video Compression for Multimedia Engineering - Fundamentals, Algorithms, and Standards, Second Edition*. CRC Press, 2008.

[96] T. Shibata, J. Kim, D. M. Hoffman, and M. S. Banks. The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of Vision*, 11(8), 2011.

[97] M. Siegel. Perceptions of crosstalk and the possibility of a zoneless autostereoscopic display. In *Stereoscopic Displays and Virtual Reality Systems VIII*, volume 4297, pages 34–41, 2001.

[98] A. Smolic, K. Mueller, N. Stefanoski, J. Ostermann, A. Gotchev, G. Akar, G. Triantafyllidis, and A. Koz. Coding algorithms for 3DTV - a survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(11):1606–1621, 2007.

[99] G. W. Snedecor and W. G. Cochran. *Statistical Methods*. Iowa State University Press, 1989.

[100] L. B. Stelmach and W. J. Tam. Stereoscopic image coding: Effect of disparate image-quality in left- and right-eye views. *Signal Processing: Image Communication*, 14:111–117, 1998.

[101] Stereoscope. http://en.wikipedia.org/wiki/charles_wheatstone.

[102] Stereoscopy. http://en.wikipedia.org/wiki/stereoscopy#realistmanual.

[103] M. Tanimoto. Overview of FTV (free-viewpoint television). In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 1552–1553, 2009.

[104] A. Tikanmaki, A. Gotchev, A. Smolic, and K. Miller. Quality assessment of 3D video in rate allocation experiments. In *Consumer Electronics, 2008. ISCE 2008. IEEE International Symposium on*, pages 1–4, 2008.

[105] K. Ukai and Y. Kato. The use of video refraction to measure the dynamics properties of the near triad in observers of a 3D display. In *Ophthalmic Physiol. Opt.*, volume 22, pages 385–388, 2002.

[106] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, 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. In *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, pages 109–114, 2012. [available] ftp://ftp.ivc.polytech.univ-nantes.fr/NAMA3DS1_COSPAD1/.

[107] C. van Berkel and J. A. Clarke. Characterization and optimization of 3D-LCD module design. In *Stereoscopic Displays and Virtual Reality Systems IV*, pages 179–186, 1997.

[108] D. Vatolin. Understanding requirements for high-quality 3D video: A test in stereo perception. 2011.

[109] Video Quality Experts Group. Final report from the video quality experts group on the validation of objective models of video quality assessment. 2000. http://www.vqeg.org.

[110] L. Wang, K. Teunissen, Y. Tu, L. Chen, P. Zhang, T. Zhang, and I. Heynderickx. Crosstalk evaluation in stereoscopic displays. *J. Display Technol.*, 7(4):208–214, 2011.

[111] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[112] Z. Wang, H. R. Sheikh, and A. C. Bovik. Objective video quality assessment. In *Handbook of Video Databases: Design and Applications*, pages 1041–1078. CRC Press, 2003.

[113] A. B. Watson. The cortex transform: rapid computation of simulated neural images. *Comput. Vision Graph. Image Process.*, 39(3):311–327, 1987.

[114] A. B. Watson, R. Borthwick, and M. Taylor. Image quality and entropy masking. In *Human Vision and Electronic Imaging II*, volume 3016, pages 2–12, 1997.

[115] S. Winkler. Perceptual video quality metrics - a review. In *Digital Video Image Quality and Perceptual Coding*, pages 155–179. CRC Press, 2005.

[116] S. Winkler and P. Vandergheynst. Computing isotropic local contrast from oriented pyramid decompositions. In *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, volume 4, pages 420–424, 1999.

[117] A. Woods. Understanding crosstalk in stereoscopic displays. In *Three-Dimensional Systems and Applications*, 2010.

[118] A. Woods, T. Docherty, and R. Koch. Image distortions in stereoscopic video systems. In *Stereoscopic Displays and Applications IV*, volume 1915, pages 36–48, 1993.

[119] A. J. Woods. How are crosstalk and ghosting defined in the stereoscopic literature? pages 78630Z–78630Z–12, 2011.

[120] M. Wopking. Viewing comfort with stereoscopic pictures: An experimental study on the subjective effects of disparity magnitude and depth of focus. *Journal of The Society for Information Display*, 3, 1995.

[121] H. Yamanoue, M. Okui, and F. Okano. Geometrical analysis of puppet-theater and cardboard effects in stereoscopic HDTV images. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16(6):744–752, 2006.

[122] H. Yamanoue, M. Okui, and I. Yuyama. A study on the relationship between shooting conditions and cardboard effect of stereoscopic images. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(3):411–416, 2000.

[123] J. Yang, C. Hou, Y. Zhou, Z. Zhang, and J. Guo. Objective quality assessment method of stereo images. In *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009*, pages 1–4, 2009.

[124] J. Yang and W. Makous. Spatiotemporal separability in contrast sensitivity. *Vision Research*, 34(19):2569–2576, 1994.

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

[126] Y. Yeh and L. Silverstein. Limits of fusion and depth judgment in stereoscopic color displays. In *Hum Factors*, volume 32, pages 45–60, 1990.

[127] J. You, G. Jiang, L. Xing, and A. Perkis. Quality of visual experience for 3D presentation - stereoscopic image. In M. Mrak, M. Grgic, and M. Kunt, editors, *High-Quality Visual Experience*, Signals and Communication Technology, pages 51–77. Springer Berlin Heidelberg, 2010.

Part II

Included Papers

A Assessment of Stereoscopic Crosstalk Perception

Author

Liyuan Xing
Junyong You
Touradj Ebrahimi
Andrew Perkis

Journal

IEEE Transactions on Multimedia (TMM), 2012.

Abstract

Stereoscopic three-dimensional (3D) services do not always prevail when compared to their two-dimensional (2D) counterparts, though the former can provide more immersive experience with the help of binocular depth. Various specific 3D artefacts might cause discomfort and severely degrade the quality of experience (QoE). In this paper, we analyse one of the most annoying artefacts in the visualization stage of stereoscopic imaging, namely, crosstalk, by conducting extensive subjective quality tests. A statistical analysis of the subjective scores reveals that both scene content and camera baseline have significant impacts on crosstalk perception, in addition to crosstalk level itself. Based on the observed visual variations during changes in significant factors, three perceptual attributes of crosstalk are summarized as the sensorial results of the human visual system (HVS). These are shadow degree, separation distance and spatial position of crosstalk. They are classified into two categories: 2D and 3D perceptual attributes, which can be described by a structural similarity (SSIM) map and a filtered depth map, respectively. An objective quality metric for predicting crosstalk perception is then proposed by combining the two maps. The experimental results demonstrate that the proposed metric has a high correlation (over 88%) when compared to subjective quality scores in a wide variety of situations.

A.1 Introduction

Stereoscopic three-dimensional (3D) imaging is based on simultaneously capturing a pair of two-dimensional (2D) images, and then separately delivering them to respective eyes. Consequently, 3D perception is generated in the human visual system (HVS). Although stereoscopic 3D services introduce a new modality (binocular depth) that can offer increasingly richer experience (immersion and realism) to the end-users, they do not always prevail when compared to their 2D counterparts. One of the major drawbacks of stereoscopic 3D services is visual discomfort, which can potentially cause users to feel uncomfortable and severely degrade the viewing experience.

The importance of various causes and aspects of visual discomfort is clarified in [13]. Especially, 3D artefacts are considered to be one of the most prominent factors contributing to visual discomfort. Such artefacts can be introduced in each stage from the acquisition to the restitution in a typical 3D processing chain [3]. In particular, crosstalk is one of the most annoying distortions in the visualization stage of a stereoscopic imaging system [19]. Crosstalk is produced by imperfect view separation that causes a small proportion of one eye image to be seen by the other eye as well. Crosstalk artefacts are usually perceived as ghosts, shadows, or double contours by human subjects.

Nowadays, crosstalk exists in almost all stereoscopic displays. However, the mechanisms behind occurrence of crosstalk can be significantly different across different stereoscopic display technologies. These mechanisms have been analysed in order to characterize and measure the components contributing to crosstalk. Therefore, crosstalk reduction can be achieved by reducing the effect of one or more of these components. Since it is not possible to completely eliminate crosstalk of displays with current technologies, researchers attempt to conceal crosstalk of a 3D presentation using image processing methods before display. Such methods are usually categorized into crosstalk cancelation. Crosstalk cancelation does not always perform efficiently in all situations. These issues have been widely investigated in the literature, e.g. see a review in [23]. However, neither of the aforementioned methods can completely eliminate crosstalk artefacts.

Therefore, it is beneficial to study how users perceive crosstalk of 3D presentations. Comparatively few research efforts have been devoted to this topic. In [16], a visibility threshold of crosstalk for different amounts of disparity and image contrast ratios in gray scale patches is provided. It shows that the visibility of crosstalk increases with increasing image contrast and disparity. However, a stereoscopic presentation is the result of a combination of different contrasts and disparities per pixel over an entire image and it is more practical to know how much crosstalk can be perceived when it is visible. Therefore, the authors of [19] investigated more realistic scenarios where natural scenes varying in crosstalk levels (0, 5, 10, and 15%) and camera baselines (0, 4, and 12cm) affect the perceptual attributes of crosstalk (perceived image distortion, perceived depth, and visual discomfort). However, only two, rather similar, natural scenes were used in their experiments. More scene contents with different depth structures and image contrasts should be taken into

account when designing a subjective experiment, because depth structure of scene content together with camera baseline can in principle determine the disparity, one of the most major factors impacting crosstalk visibility [16]. Moreover, the authors of [7] found out that monocular cues of images also play an important role in the crosstalk perception, in addition to contrast ratio and disparity. In [14], it is shown that edges and high contrast of computer-generated wire-frames make crosstalk more visible when compared to natural images. This means that crosstalk can be more efficiently concealed on images with more texture or details. These observations partially support a hypothesis that scene content is an important factor impacting users' perception of crosstalk. Although other artefacts, e.g. blur and vertical disparity as investigated in [12], may also have impact on the crosstalk perception, they can be often corrected by post processing techniques.

Although subjective test is the most reliable way to evaluate the perceived quality, it is time-consuming, expensive, and unsuitable for real-time applications. To deal with these drawbacks, objective metrics that can predict the human subjects' judgment with a high correlation are desired. To develop good objective metrics, the perception mechanisms need to be well understood and taken into account. However, this is usually fairly difficult. Therefore, development of objective 3D quality models is still in its early stages. Researchers first started with exploring whether or not traditional 2D metrics can be applied to stereoscopic quality assessment [1, 26]. Subsequently, a few objective metrics [15, 6, 17] that take into account the characteristics of stereoscopic images have been proposed. However, most of the existing objective metrics are designed to assess quality degradations caused by lossy compression schemes. To the best of our knowledge, only one objective metric that considers non-compression quality degradations, induced during acquisition and display stages of stereoscopic media, has been proposed in [11]. This metric is modelled by a linear combination of three measurements, which evaluate the perceived depth, visual fatigue and temporal consistency, respectively.

In this paper, subjective tests [24] have been conducted to collect the evaluation scores on crosstalk perception of a wide range of 3D stimuli, including different scene contents, camera baselines and crosstalk levels. Thereby, a comprehensive database of crosstalk perception for a wide variety of situations has been created. Furthermore, based on a statistical analysis of subjective scores, scene content, camera baseline and crosstalk level are found to have significant impacts on the perception of crosstalk. By changing the amplitude of the significant factors, three perceptual attributes of crosstalk in the HVS have been observed. These perceptual attributes are further used to design an objective quality metric [25] for crosstalk perception.

The main contributions of the paper are twofold: first, our subjective tests provide a comprehensive database for crosstalk perception in stereoscopic images. Second, users' subjective perception of crosstalk is predicted using an objective metric based on a rigorous analysis of perceptual attributes of crosstalk.

The remainder of this paper is organized as follows. In Section A.2, we present the subjective tests on crosstalk perception as well as a statistical analysis of the subjective scores. In Section A.3, perceptual attributes of crosstalk are explained by

an observation on the visual variations of stimuli when several significant factors change. Furthermore, a perceptual objective metric for crosstalk perception is proposed by describing the perceptual attributes of crosstalk and the experimental results are reported in Section A.4. Finally, concluding remarks are given in Section A.5.

A.2 Subjective Tests on Crosstalk Perception

Several recommendations for subjective evaluation of visual stimuli have been issued by the International Telecommunication Union (ITU), e.g. the widely used ITU-R BT.500 [10] for television pictures. For subjective evaluation of stereoscopic television pictures, ITU-R BT.1438 [9] has made a few first steps, but it still lacks many details. The authors of [4] have summarized the lacks in the form of additional requirements. In this subjective test, we followed these methodologies and further customized them for the crosstalk perception. In the following, we will provide some details about laboratory environment where the subjective tests were conducted, how test stimuli were prepared, which test method was adopted, as well as what results were obtained from the subjective tests.

A.2.1 Laboratory Environment

A.2.1.1 Display System

Polarization technique was used to present 3D images, as illustrated in Figure A.1. Specifically, two Canon XEED SX50 projectors with resolution of 1280×960 were placed on a Chief ASE-2000 Adjusta-Set Slide Stacker. The stacker can be adjusted with $+/-7^\circ$ swivel, $+/-20^\circ$ tilt and $+/-7^\circ$ leveling ranges. Two Cavision linear polarizing glass filters with size of 4in×4in were installed orthogonally in front of the projectors. In this way, two views were projected and superimposed onto the backside of a 2.40m×1.20m silver screen. The projected distance between the projectors and the silver screen was about 2 meters, forming a projected region occupying the central area of the silver screen with a width of 1.12m and height of 0.84m. Images up-sampled by a bicubic interpolation method were displayed in full screen mode. The subjects equipped with polarized glasses were asked to view 3D images on the opposite side of the silver screen. The viewing distance was set to about five times the image height (0.84m×5), as suggested in [10]. The Field of View (FOV) was thus 15°.

A.2.1.2 Alignment of Display System

Prior to the tests, the display system was calibrated to align the two projectors. In particular, the positions of two projectors were adjusted to guarantee that the center points of projectors, projected region, and silver screen, positioned in the same horizontal line (center horizontal line as shown in Figure A.1) and the line was perpendicular to the silver screen. Moreover, the angles of stackers and the

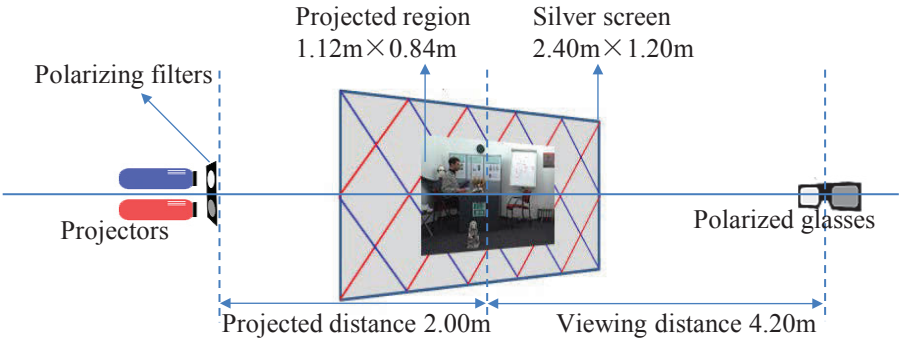


Figure A.1: The polarized display system used in subjective tests.

keystones of the projectors were adjusted with the help of projected Hermann grid images. The adjustment of display system was finished once the two Hermann grid images from the left and right projectors were exactly overlapped.

A.2.1.3 Measurement of System-introduced Crosstalk

After the alignment, the system-introduced crosstalk was measured immediately. As introduced in [23], the terminology and mathematical definitions of crosstalk are diverse and sometimes contradictory. We adopt the definition of system-introduced crosstalk as the degree of the unexpected light leakage from the unintended channel to the intended channel. In particular, we measured the leakage in a situation when the left and right test images have the maximum difference in brightness. The system-introduced crosstalk is measured mathematically as follows,

$$P_l = \frac{Lx_{GL}(WB) - Lx_{GL}(BB)}{Lx_D(WB) - Lx_D(BB)} \tag{A.1}$$

where WB denotes a pair of test images (the left image is in white completely whilst the right in black), and BB is another image pair both in black. Lx_D denotes the luminance measured on silver screen and Lx_{GL} denotes the luminance after the right lens of polarized glasses which is cling to the silver screen. Therefore, P_l denotes the system-introduced crosstalk from the left channel to the right, which is approximately 3% in our experiments. The consistence of the system-introduced crosstalk of polarized display has also been verified over the display, between projectors, and among different combinations of brightness between left and right test images.

A.2.1.4 Room Conditions

The test room had the length of 11.0m, width of 5.6m and height of 2.7m. During the subjective tests, all the doors and windows of the test room were closed and covered by black curtains. In addition, the lights in the room were turned off

(a) Book Arrival

(b) Champagne (c) Dog (d) Love Bird

(e) Outdoor (f) Pantomime (g) Newspaper

Figure A.2: Visual samples of the selected scenes.

except for one reading lamp on a desk in front of the subject, which was used to illuminate the keyboard when entering subjective scores. In this way, subjects could concentrate on the 3D perception, as opposed to entering the scores using the keyboard.

A.2.2 Test Stimuli

Scene content and camera baseline are requisite factors for stereoscopic imaging and also affect users' perception on crosstalk. Therefore, scene content, camera baseline and crosstalk level were selected as three observed factors in the subjective tests of crosstalk perception. In particular, three camera baselines and four crosstalk levels were applied to six scene contents, which resulted in 72 test stimuli, in total.

A.2.2.1 Scene Content

Seven multi-view sequences (one for training) from the MPEG [8] were chosen as representative scene contents, as shown in Figure A.2. These scene contents cover a wide range of depth structures, contrasts, colors, edges and textures, which were considered as potential factors impacting users' perception of crosstalk. In particular, a wide range of depth structures were obtained by including both indoor and outdoor scenes.

Table A.1: Number of the selected cameras from left to right and the resulting camera baselines

Content	Camera	Camera baseline (mm)
Book.	04-01	58,114,172
Cham.	38-41	50,100,150
Dog.	38-41	50,100,150
Love.	01-04	39,77,116
Out.	04-01	72,135,204
Pant.	38-41	50,100,150
News.	00-03	46,93,139

A.2.2.2 Camera Baseline

Three camera baselines were formed from four consecutive cameras. The leftmost camera always served as the left eye view and the other three cameras took turns as the right eye views for 3D images. In this way, three 3D images with different camera baselines were generated for each scene. Table A.1 gives more information about the selected cameras and the resulting camera baselines of the 3D images.

A.2.2.3 Crosstalk Level

In order to simulate different levels of system-introduced crosstalk for different displays, crosstalk artefacts were added to three 3D image pairs, to each of which four different crosstalk levels were introduced using the algorithm developed in [2]. This algorithm can be summarized by the following equations,

$$\begin{cases} R_c^p = R_o + p \times L_o \\ L_c^p = L_o + p \times R_o \end{cases} \tag{A.2}$$

where L_o and R_o denote the original left and right views, L_c^p and R_c^p are the distorted views by simulating system-introduced crosstalk distortions, and the parameter p is to adjust the level of crosstalk distortion. According to equation (A.2), the simulating algorithm keeps a consistent characteristic of the system-introduced crosstalk of polarized display by applying the same leakage percentage p to all pixels in the entire image, both the left and right views, and different brightness of all pixels.

In our experiments, the system-introduced crosstalk P is 3%, which should be added to the simulated crosstalk in image pairs in equation (A.2). Therefore, the overall system-introduced crosstalk perceived by the users is defined as follows,

$$\begin{cases} R_c^{(P+p)} = R_c^p + P \times L_c^p = (1 + P \times p) \times R_o + (P + p) \times L_o \\ L_c^{(P+p)} = L_c^p + P \times R_c^p = (1 + P \times p) \times L_o + (P + p) \times R_o \end{cases} \tag{A.3}$$

where $R_c^{(P+p)}$ is the overall system-introduced crosstalk combining both the system-introduced crosstalk P and the simulated crosstalk p. As crosstalk aroused by stereoscopic techniques usually ranges from 0 to 15% [16], and the image quality might be very low if the crosstalk level is large, e.g. over 15% [19], the parameter p was set to 0, 5%, 10% and 15%, respectively, in our subjective tests. Thus, the overall crosstalk levels in our experiments were actually $P+p$, i.e., 3%, 8% 13% and 18%, respectively. As the maximum pixel value change tuned by $P \times p$ is only 1.1475 ($255 \times 3\% \times 15\%$), its effect can be ignored. Therefore, the overall system-introduced crosstalk is simulated as following based on an additive rule,

$$\begin{cases} R_c^{(P+p)} = R_o + (P+p) \times L_o \\ L_c^{(P+p)} = L_o + (P+p) \times R_o \end{cases} \tag{A.4}$$

Equation (A.4) indicates that the different simulated crosstalk levels can be applied to a stereoscopic display with a consistent system-introduced crosstalk level. Therefore, the crosstalk level will refer to the overall crosstalk level $P + p$ in this work.

A.2.3 Test Methodology

A.2.3.1 Single Stimulus

Among different methodologies for subjective quality assessment of Standard Definition TeleVision (SDTV) pictures in ITU-R BT. 500 [10], three widely used methodologies are Double Stimulus Continuous Quality Scale (DSCQS), Double Stimulus Impairment Scale (DSIS), and Single Stimulus (SS). In this study, as several camera baselines for each scene have been taken into account, it is difficult to choose an original 3D image as the reference. Therefore, we adopted the SS method. The SS method was also used in assessing the quality levels of stereoscopic images with varying camera baselines in the literature [19, 5]. In addition, in order for subjects to have sufficient time to generate their 3D perception and have an extensive exploration of still 3D images, a minor modification was made on the SS method such that the subjects could freely decide the viewing time for each image as in [19].

A.2.3.2 Test Interface

In order to support the adaptive SS methodology, a special interface was developed to conveniently display the stereoscopic images in a random order. In addition, a subject could conveniently and freely decide when he/she moved to the next image pairs by pressing 'Ctrl' key on the keyboard. The score of the current image pairs was recorded by pressing a 'Numerical' key instead of writing on an answer sheet. Other special considerations, such as displaying in full screen, disabling unnecessary keys, updating the scores and so on, were also included in the developed interface.

A.2.3.3 Subjects

Before the training sessions, visual perception related characteristics of the subjects were collected, including pupillary distance (measured by a ruler), normal or corrected binocular vision (tested by the Snellen chart), color vision (tested by the Ishihara), and stereo vision (tested by the TV-04 and TV-07 in ITU-R BT. 1438 [9]).

A total of 28 subjects participated in the tests, consisting of 15 males and 13 females, aged from 23 to 46 years old. The binocular vision of all the subjects was above 0.80 with the mean of 1.05 and the standard deviation of 0.28. Although 7 subjects had monocular vision differences of either 0.4 or 0.2, all the subjects could perceive the binocular depth.

A.2.3.4 Training Sessions

Subjects participated in both training and test sessions individually. During the training sessions, an example of five categorical adjectival levels (see Table A.2) was shown to the subject in order to benchmark and harmonize their measuring scales. The Book Arrival scene was selected by expert viewers in such a way that each quality level was represented by an example image and that these example images could cover a full range of quality levels within the set of test stimuli. When each example image was displayed, the operator verbally explained the corresponding quality level to the subject. In addition, a detailed explanation of every scale was provided to subjects in form of written instructions (see Table A.2). Subjects were encouraged to view the representative examples as long as they wished and asked questions if they needed any further clarifications. The training sessions would continue until subjects could understand and distinguish the five different quality levels.

A.2.3.5 Test Sessions

During the test sessions, subjects were first presented with three dummy 3D images from the Book Arrival content, which were not used in the training sessions. These dummy images were used to stabilize subjects' judgment, and the corresponding scores were not included in the subsequent data analysis. Following the dummy images, 72 test images were randomly shown to the subjects. A new 3D image was shown after a subject had entered his/her score for the previous one. During the test period, the subjects were not allowed to ask questions in order to avoid any interruption during the entire session.

A.2.4 Subjective Results Analysis

The subjective scores of the 72 test stimuli given by 28 subjects are analysed in this subsection. Particularly, we aim to analyse the relationship between crosstalk perception and three potential significant factors, including scene content, camera baseline and crosstalk level.

Table A.2: Explanations of five categorial adjectival levels and their training examples from Book Arrival

	Explanation	**Examples (baseline, crosstalk)**
5	**Imperceptible**: you cannot see any crosstalk or you can perceive very slightly only when you pay special attention to a certain region.	0mm,3%
4	**Perceptible but not annoying**: you can see a little bit of crosstalk at a first glance, but the quality of the whole image is still good.	58mm,3%
3	**Slightly annoying**: there is obvious crosstalk. However you can accept viewing such quality, reluctantly.	114mm,8%
2	**Annoying**: the 3D perception still can be formed however you refuse to accept viewing such quality in daily life.	114mm,13%
1	**Very annoying**: the 3D perception is hardly formed and you feel uncomfortable.	172mm,18%

A.2.4.1 Normality Test and Outlier Removal

In order to apply arithmetic mean value as Mean Opinion Scores (MOS) and use parametric statistical analysis methods, such as ANalysis Of VAriance (ANOVA), the normality of subjective scores across subjects needs to be validated. The β_2 test recommended in [10] based on calculating the kurtosis coefficient was adopted for a normality test. We classified the β_2 test results into three groups: normal ($2 \leq \beta_2 \leq 4$), close to normal ($1 \leq \beta_2 < 2$ or $4 < \beta_2 \leq 5$), and abnormal ($\beta_2 < 1$ or $\beta_2 > 5$). If the total proportion of normal and close to normal was more than 80%, we assumed that the subjective scores in our tests subject to the normal distribution. The results showed that the majority of stimuli (55 over 72) were normal distributed and (11 over 72) were close to normal, while others (6 over 72) were not. Therefore, we can assume that the subjective scores subject to the normal distribution. A screening test of subjects was also performed according to a guideline in [10]. Subjects who had produced votes significantly distant from the average scores should be removed. Consequently, one outlier was detected and the corresponding results were excluded from the following analysis.

A.2.4.2 Observations

After removing the outlier, MOS and 95% Confidence Interval (CI) were computed and plotted as a function of camera baseline and crosstalk level for all the six scene contents separately, as shown in Figure A.3. A number of observations can be made based on the results in those plots.

Generally speaking, the MOS values decrease as the level of crosstalk distortions increases. However, the decreasing degree of MOS for Dog and Pantomime is not

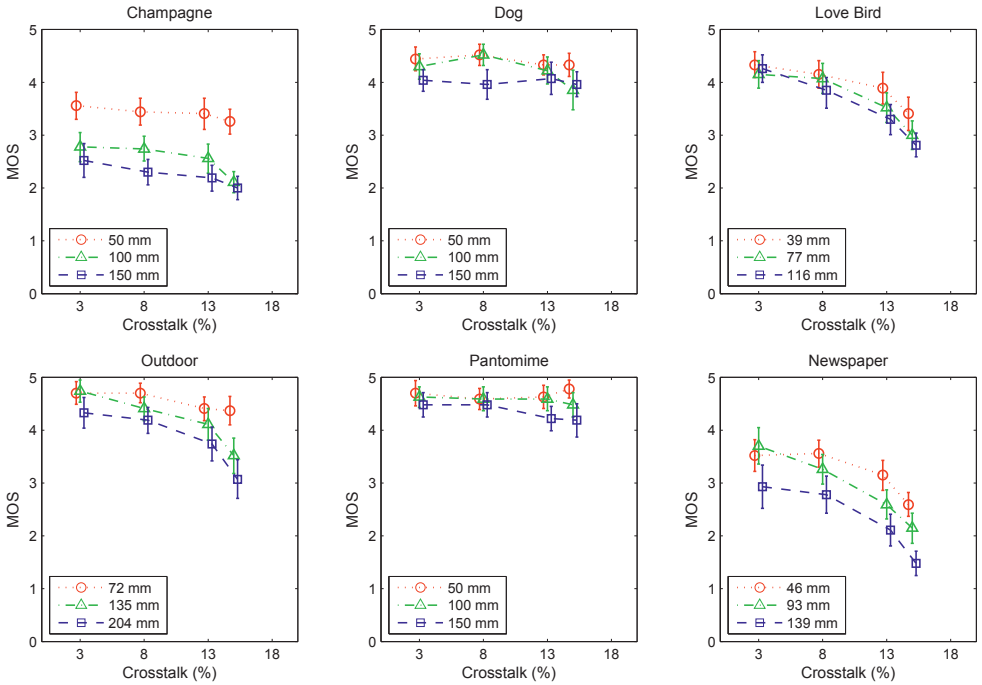


Figure A.3: MOS and CI (significance of 95%) of subjective scores on crosstalk perception for scene contents.

significant as those in other four scenes. When considering the impact of camera baseline, we can observe that there is a general tendency of reduction of the MOS values of crosstalk perception with increasing camera baseline. However, while this tendency is significant for the near indoor scenes (Champagne and Newspaper), it is less significant for others, especially for Dog and Pantomime. Therefore, we can summarise the observations as follows:

- *observation i*: crosstalk level and camera baseline have an impact on crosstalk perception.

- *observation ii*: the impact of crosstalk level and camera baseline on crosstalk perception varies with scene content.

In addition, the individual curves in Figure A.3 show that even the highest MOS values of Champagne and Newspaper are still below 4, which indicates that the system-introduced crosstalk is more perceptible in close up scenes. Furthermore, we also noticed that there exist exceptions where MOS values increase with the increasing of camera baseline and crosstalk level. Hence, crosstalk perception might be influenced by other perceptual attributes, such as perceived depth.

Table A.3: Impact of crosstalk level(CL) and camera baseline(CB) on crosstalk perception for each Scene

	Cham.	**Dog.**	**Love.**	**Out.**	**Pant.**	**News.**
CL	√	√	√	√	√	√
CB	√	√	√	√	−	√
CL*CB	−	√	√	√	−	√

A.2.4.3 Statistical Analysis

In order to verify the observations and evaluate the impact of the independent variables (scene content, camera baseline, crosstalk level) on the dependent variable (crosstalk perception), we utilized ANOVA to analyse the subjective scores obtained in our tests. ANOVA is a general technique that can be used to test the equality hypothesis of means among two or more groups. These groups are classified by factors (independent variables whose settings are controlled and varied by the operator) or levels (the intensity settings of a factor). An N-way ANOVA treats N factors and the null hypothesis includes: *i*) there is no difference in the means of each factor; *ii*) there is no interaction between n-factors ($2 \leq n \leq N$). The null hypothesis is verified using the F-test and can be easily judged by the p-value. When the p-value is smaller than 0.05, the null hypothesis is rejected, which means there is a significant difference in means. In particular, there is a significant difference between the levels of a factor such that the factor has significant effect; or the difference between the levels of one factor is not same for the levels of other factors such that there is an interaction between different factors.

We used Statistical Package for the Social Sciences (SPSS) statistics toolbox for our analysis. Tables A.3-A.5 show the ANOVA results for different factors. In these results, '√' indicates that the corresponding factor has a significant effect on the crosstalk perception or multiple factors have interactions in terms of the impact on the crosstalk perception, and '−' means the factor has no significant effect, or there is no interaction between multiple factors.

When considering the *observation i*, we first tested the impact of crosstalk level and camera baseline on crosstalk perception for each scene content. As shown in Table A.3, both crosstalk level and camera baseline have a significant impact on the crosstalk perception in each scene content, generally speaking. However, an exception is that camera baseline has no significant impact on crosstalk perception for Pantomime. In addition, for most scenes except for Champagne and Pantomime, the crosstalk level and camera baseline have interaction in terms of the impact on crosstalk perception.

Regarding the *observation ii*, the impact of scene content on crosstalk perception between every two scenes has been reported in Table A.4. It can be seen that there is a significant difference between scene contents in terms of crosstalk perception for most scene content pairs. However, there are two exceptional pairs, Champagne and Newspaper, as well as Outdoor and Dog. In other words, there is no significant

Table A.4: Impact of scene content(SC) on crosstalk perception between every two scenes.

	Cham.	**Dog.**	**Love.**	**Out.**	**Pant.**	**News.**
Cham.	n/a	√	√	√	√	−
Dog.	√	n/a	√	−	√	√
Love.	√	√	n/a	√	√	√
Out.	√	−	√	n/a	√	√
Pant.	√	√	√	√	n/a	√
News.	−	√	√	√	√	n/a

Table A.5: Impact of crosstalk level (CL), camera baseline (CB) and scene content (SC) on crosstalk perception for all the scenes

CL	**CB**	**SC**	**CL*CB**	**CL*SC**	**CB*SC**	**CL*CB*SC**
√	√	√	√	√	√	−

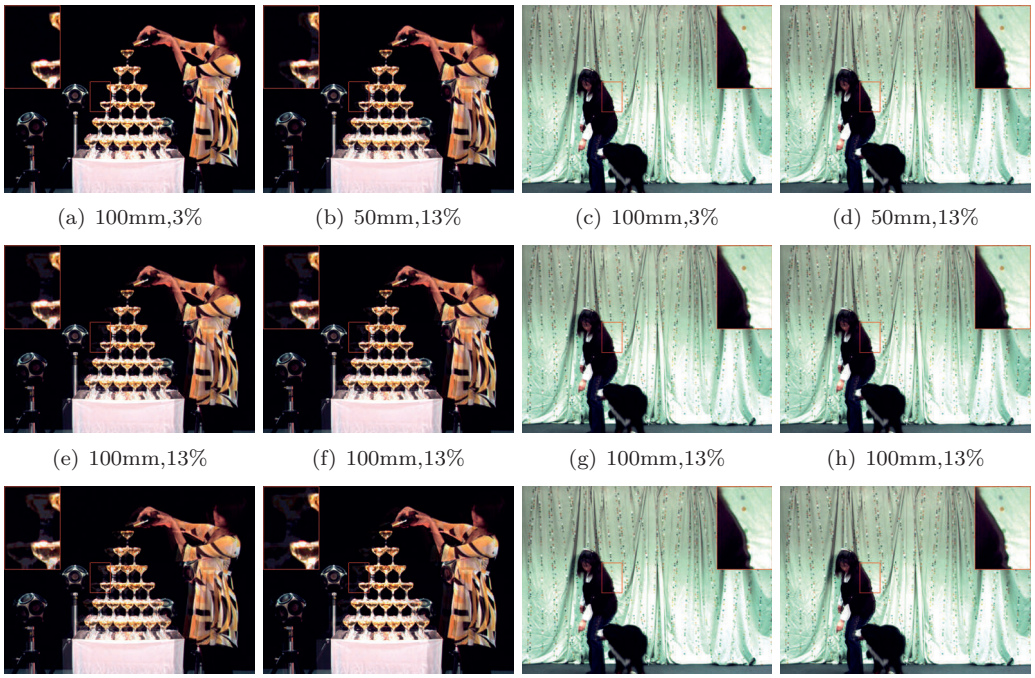
difference between Champagne and Newspaper when their crosstalk perceptions are considered. The same argument also applies to Outdoor and Dog, although it may seem that Pantomime and Dog are similar when judging from Figure A.3.

All these observations can be further verified if we consider three factors together for crosstalk perception on the whole test stimuli. Table A.5 shows that crosstalk level, camera baseline and scene content have significant impacts on crosstalk perception, respectively, and they have 2-factors interactions in terms of the impact on crosstalk perception. However, 3-factors interaction does not have a significant impact on crosstalk perception.

A.3 Understanding of Crosstalk Perception

After identifying the significant factors, their relationship with the perceptual attributes of crosstalk can be modelled. Because the perceptual attributes of crosstalk are the sensorial results of HVS and closer to perceptive viewpoint, the gap between low-level significant factors and high-level users' perception on crosstalk can be bridged.

Ten test stimuli with different amplitudes of the significant factors were selected to represent the perceptual attributes of crosstalk, as shown in Figure A.4. The red rectangular regions highlight the selected regions in the images for the sake of discussion of the crosstalk, and have been enlarged and placed on a top right or left corner of each image. These stimuli consist of two scene contents (Champagne and Dog), which were applied five combinations of camera baselines and crosstalk levels, respectively. Specifically, the selected scene contents have comparatively large differences in depth structures and image contrast.



(a) 100mm,3% (b) 50mm,13% (c) 100mm,3% (d) 50mm,13%

(e) 100mm,13% (f) 100mm,13% (g) 100mm,13% (h) 100mm,13%

(i) 100mm,18% (j) 150mm,13% (k) 100mm,18% (l) 150mm,13%

Figure A.4: Left eye view for scene contents Champagne and Dog with different combinations of camera baseline and crosstalk level.

When these test stimuli were perceived on a stereoscopic display in a certain order of changing significant factors, we summarized the visual variations of crosstalk to its perceptual attributes, which in turn are shadow degree, separation distance and spatial position of crosstalk. Shadow degree and separation distance are 2D perceptual attributes existing in single eye view and they are still maintained in 3D perception. On the other hand, spatial position emphasizes the perceptual attribute of crosstalk in 3D perception when the left and right views are fused.

A.3.1 2D Perceptual Attributes

A.3.1.1 Shadow Degree of Crosstalk

We define it as the distinctness of crosstalk against the original view. If the shadow degree increases, crosstalk becomes more annoying. When viewing the Champagne and Dog presentations from top downwards in the first and third columns, it can be noticed that the shadow degree of crosstalk becomes stronger with the increase of the crosstalk level. It indicates that crosstalk level relates to shadow degree of crosstalk. Moreover, the shadow degree is more visible in the Champagne pre-

sentations when compared to the Dog presentations. This is due to the different contrast structures in Champagne and Dog presentations. Thus, the contrast of scene content also relates to shadow degree of crosstalk. In fact, the contrast of scene content and crosstalk level reflect the shadow degree of crosstalk mutually, which implies that the 2-factors interaction between crosstalk level and contrast of scene content has a relationship with the shadow degree of crosstalk.

A.3.1.2 Separation Distance of Crosstalk

We define it as the distance of crosstalk separated from the original view. Crosstalk is more annoying with increasing the separation distance. When viewing the Champagne and Dog presentations from top downwards in the second and fourth columns, it can be noticed that the separation distance of crosstalk becomes larger with the increase of camera baseline. It indicates that camera baseline reflects the separation distance of crosstalk, which shows that camera baseline has a relationship with separation distance. Moreover, the separation distance of crosstalk is more visible in Champagne presentations as opposed to Dog. This is due to different relative depth structures in Champagne and Dog presentations, thus depth of scene content also relates to separation distance of crosstalk. Actually, the camera baseline and relative depth structure of scene content together, namely, disparity, determine the separation distance of crosstalk. This confirms that the 2-factors interaction between camera baseline and depth of scene content relates to separation distance of crosstalk.

A.3.1.3 Interaction between 2D Perceptual Attributes

If we pay attention to the change of crosstalk level and camera baseline together when viewing the Champagne and Dog presentations from left to right in the first and third rows, it can be noticed that the shadow degree and separation distance of crosstalk interact mutually. It reflects that the interaction between crosstalk level and camera baseline has a relationship with the interaction between 2D perceptual attributes. Moreover, less shadow degree and separation distance of crosstalk can be perceived with the Dog presentations when the same camera baseline and crosstalk level changes were applied as that of Champagne because of the difference of scene content including both contrast and relative depth structure. Thus, scene content relates to interaction between 2D perceptual attributes. Furthermore, this also confirms that the impact of crosstalk level and camera baseline on crosstalk perception varies with the scene content. Thus, the 3-factors interaction between crosstalk level, camera baseline and scene content has a relationship with the interaction between 2D perceptual attributes.

A.3.2 3D Perceptual Attribute

Spatial position of crosstalk is defined as the impact of crosstalk position in 3D space on perception when the left and right views are fused and 3D perception is

Table A.6: Relationship between perceptual attributes of crosstalk and significant factors: crosstalk level (CL), camera baseline (CB), contrast of scene content (SC_C), depth of scene content (SC_D), both contrast and depth of scene content (SC_CD)

Perceptual attributes	**Related factors**
2D: Shadow degree	CL, SC_C, CL*SC_C
2D: Separation distance	CB, SC_D, CL*SC_D
Interaction between 2D perceptual attributes	CL*CB, SC_CD, CL*CB*SC_CD
3D: Spatial position	SC_D in visible crosstalk region

generated. Specifically, we observed that spatial position of crosstalk only impacts the visible crosstalk satisfying requirements of shadow degree and separation distance of crosstalk. In our experiments, the crosstalk of foreground objects usually has more impact on perception than background objects due to the fact that the foreground objects are closer to the test subjects and have larger disparity because of parallel camera arrangement and rectification. Therefore, relative depth structure of scene content in the region of visible crosstalk relates to the perceptual attribute spatial position of crosstalk. Additionally, focus of attention might also have an important role behind the observation from our experiments. However, as in the data evaluated in this work, foreground objects were always also a priori the focus of attention. In future work, we will further investigate the influence of focus of attention on the crosstalk perception.

A.3.3 Summary

Table A.6 lists the relationship between the perceptual attributes and related factors as explained earlier. As can be seen from the table, the 2D perceptual attributes include all the significant factors in Table A.5, which indicates that 2D perceptual attributes can characterize low-level significant factors while in a more perceptual level of HVS. Moreover, the table also shows that 2D perceptual attributes alone are not enough to explain the visual perception of crosstalk. Thus, 3D perceptual attribute should be modeled to predict the users' perception on crosstalk. It indicates that an objective metric proposed directly from the significant factors in Table A.5 is not comprehensive. However, selecting those stimuli with distinct visual variations corresponding to the significant factors indeed reduces the complexity and facilitates observation of the perceptual attributes.

A.4 Objective Quality Metric

An objective metric for crosstalk perception can be developed based on modelling 2D and 3D perceptual attributes of crosstalk. In this section, we will explain what kinds of existing maps can reflect the perceptual attributes, how these maps are

combined to construct a perceptual objective metric, and the experimental results of the metric.

A.4.1 2D Perceptual Attributes Map

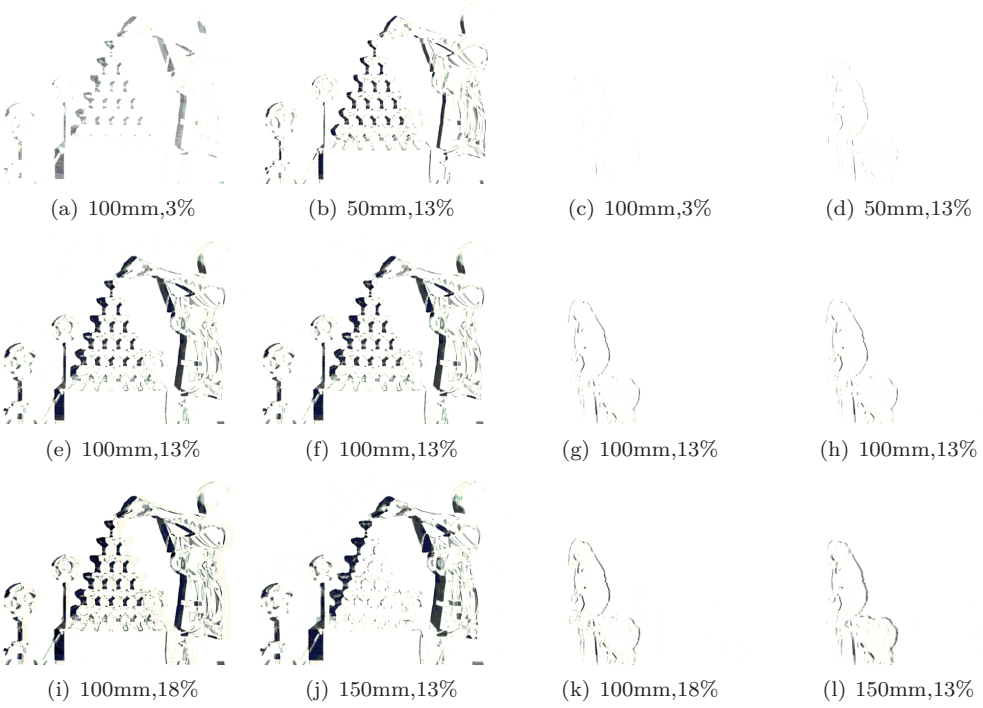
The 2D perceptual attributes were illustrated in Figure A.4 using the left eye view with crosstalk added distortion as in equation (A.4). It can be noticed that shadow degree, separation distance of crosstalk, and their interaction are most visible in the edge region with high contrast. The Structural SIMilarity (SSIM) quality measure proposed by Z. Wang et al. [22] can describe the 2D perceptual attributes of crosstalk to some extent. SSIM assumes that the measurement on structural information provides a good estimation of the perceived image quality because the HVS is highly adapted to extract structural information from a visual scene.

A Matlab implementation of the SSIM is accessible from [20]. In addition, an SSIM map of a test image is also provided, which allows a closer look at specific regions instead of the entire image. Considering the combination of 2D and 3D perceptual attributes in a single objective metric, the SSIM map with quality measure on all the pixels is preferred, as opposed to the SSIM with a single quality measure for the entire image. SSIM is constructed based on the comparisons of three components: luminance, contrast, and structure, between an original image without any distortions and its degraded version. In our case, the original image is the one shown on the stereoscopic display without any crosstalk L_o, and the distorted version is the one with both system-introduced and simulated crosstalk L_c. Finally, the SSIM map is defined as follows,

$$L_s = SSIM(L_o, L_c) \tag{A.5}$$

where SSIM denotes the SSIM algorithm and L_s is the generated SSIM map of the left eye view.

Figure A.5 is a representative illustration of SSIM map derived from the crosstalk distorted Champagne and Dog presentations in Figure A.4. In the SSIM map, 0 (black) at a pixel means the largest difference between the original and crosstalk added image and 1 (white) denotes no difference. For Champagne, it can be seen that when the crosstalk level is larger in the first column, the shadow degree of crosstalk represented by the SSIM map is darker. Also, when the camera distance is larger in the second column, the separation distance of crosstalk represented by the SSIM map becomes wider. In addition, their mutually interaction is also described by the SSIM map when the shadow degree and separation distance change synchronously in the first and third rows. The same situation exists with the Dog presentation. Moreover, it can be seen that different shadow degrees and separation distances of crosstalk for Champagne and Dog with the same camera baseline and crosstalk level can also be expressed by the SSIM map, which means that the scene content difference is also characterized. Thus, the SSIM map can reflect these 2D perceptual attributes, namely, shadow degree of crosstalk, separation distance of crosstalk and their interactions.



(a) 100mm,3% (b) 50mm,13% (c) 100mm,3% (d) 50mm,13%

(e) 100mm,13% (f) 100mm,13% (g) 100mm,13% (h) 100mm,13%

(i) 100mm,18% (j) 150mm,13% (k) 100mm,18% (l) 150mm,13%

Figure A.5: Illustrations of SSIM map on Champagne and Dog.

A.4.2 3D Perceptual Attribute Map

Spatial position of crosstalk describes users' perception of crosstalk in 3D space, which can be characterized because visible crosstalk of foreground objects should have more impact on perception than background objects. Therefore, in order to form a 3D perceptual attribute map, depth structure of scene content and region of visible crosstalk should be combined.

Relative depth structure of scene content can be represented by the depth map. Depth estimation algorithms are usually performed in two approaches: *i*) from one single image using monocular cues; and *ii*) from stereo or multi images using stereo (triangulation) cues. The latter is usually more accurate but requires corresponding intrinsic and extrinsic camera parameters of the stereo images. Since the performance of the metric relies on the accuracy of the depth estimation algorithm, we adopt the latter approach and the Depth Estimation Reference Software (DERS) [21] version 4.0 was employed in this paper. The depth map of the original right eye view R_o is calculated as follows,

$$R_{dep} = DERS(R_o) \tag{A.6}$$

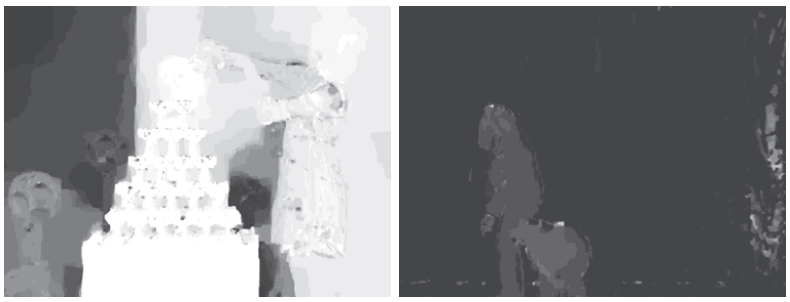


Figure A.6: Illustrations of depth map of Champagne and Dog when camera baseline is 150mm.

Figure A.7: Illustrations of filtered depth map of Champagne and Dog when camera baseline is 150mm and crosstalk level is 3%.

where $DERS$ denotes the DERS algorithm proposed in [20] and R_{dep} is the generated depth map of the right view. R_{dep} is normalized to represent a relative 3D depth in which 0 denotes the farthest depth value and 255 the nearest. Figure A.6 gives an example of the depth map of Champagne and Dog. The farthest and nearest depth values are 7.7m and 2.0m for Champagne, and 8.2m and 2.5m for Dog, respectively. However, they are both normalized by a same factor 5.7m. It can be seen that the foreground object champagne is much brighter than that in the Dog presentation. Therefore, the foreground of Champagne is much closer to its nearest depth plane than that of Dog.

The region of visible crosstalk is also defined based on the SSIM map, because we observed that crosstalk is more visible in the regions where the pixel value of SSIM map is smaller than a threshold. A threshold 0.977 was obtained experimentally from our experiments. Therefore, the following equation is used to define the filtered depth map as 3D perceptual attribute map,

$$R_{pdep}(i,j) = \left\{ \begin{array}{ll} R_{dep}(i,j) & if L_s(i,j) < 0.977 \\ 0 & if L_s(i,j) \geq 0.977 \end{array} \right\} \tag{A.7}$$

where i and j are the pixel index, and R_{pdep} denotes the filtered depth map corresponding to the visible crosstalk region of left eye image, as illustrated in Figure A.7 .

A.4.3 Objective Metric for Crosstalk Perception

As aforementioned, the 2D and 3D perceptual attributes can be represented by the SSIM and filtered depth maps. Therefore, the overall crosstalk perception is supposed to be an integration of the two maps. Since 3D perceptual attributes discover that visible crosstalk of foreground objects has more impacts on perception than background objects, more weights should be assigned to the visible crosstalk of foreground than background. In other words, SSIM map should be further weighted by filtered depth map. Thus, the integration is performed in the following equation,

$$C_{pdep} = L_s \times (1 - R_{pdep}/255) \tag{A.8}$$

$$V_{pdep} = AVG(C_{pdep}) \tag{A.9}$$

where C_{pdep} and V_{pdep} denote the combined map and the quality value predicted by the objective metric, respectively. AVG denotes the average operation. In the equation (A.8), the filtered depth map R_{pdep} is normalized into the interval $[0, 1]$ first by the maximum depth value 255, and then subtracted it from 1 to comply with the meaning of SSIM map that a lower pixel value in SSIM map means a larger crosstalk distortion. When two pixels with identical values in the SSIM map locate in the foreground and background, respectively, the C_{pdep} value of the foreground pixel will be smaller than the background pixel after combining with the filtered depth map.

A.4.4 Experimental Results

The performance of an objective quality metric can be evaluated by a comparison with respect to the MOS values obtained in subjective tests. The proposed metric was compared with traditional 2D metrics V_{psnr} and V_{ssim} as well as other three metrics V_{dep}, V_{pdis} and V_{dis}, which combine the 2D and 3D perceptual attributes in different approaches as in the following equations.

$$V_{psnr} = PSNR(L_o, L_c) \tag{A.10}$$

$$V_{ssim} = AVG(L_s) \tag{A.11}$$

$$V_{dep} = AVG(L_s \times (1 - R_{dep}/255)) \tag{A.12}$$

$$R_{dis} = SSDMF(R_o, L_o) \tag{A.13}$$

$$R_{pdis}(i,j) = \left\{ \begin{array}{ll} R_{dis}(i,j) & if L_s(i,j) < 0.977 \\ 0 & if L_s(i,j) \geq 0.977 \end{array} \right\} \tag{A.14}$$

$$V_{pdis} = AVG(L_s \times (1 - R_{pdis}/255)) \tag{A.15}$$

$$V_{dis} = AVG(L_s \times (1 - R_{dis}/255)) \tag{A.16}$$

where V_{psnr} and V_{ssim} are the 2D metrics calculated between the original and crosstalk added left image, instead of original left and right images. V_{dep} is a combination of SSIM map L_s and the depth map R_{dep} instead of the filtered depth map R_{pdep} as in the metric V_{pdep}. It means that R_{dep} weights the entire image while R_{pdep} only weights the region of visible crosstalk in the image. R_{dis} denotes the disparity map of the right eye image, which is the result of a combination of relative depth structure of scene content and camera settings, such as camera baseline. Since R_{dis} also contains the information about relative depth structure of scene content, we attempt to compare the performance of the different metrics based on the disparity and depth maps, respectively. In the equations, the filtered disparity map R_{pdis} was obtained from R_{dis} using the same approach as R_{pdep} from R_{dep}, and metrics V_{pdis} and V_{dis} followed the same combination as in building V_{pdep} and V_{dep}, respectively. Particularly, we adopted a stereo correspondence algorithm using the Sum of Squared Difference plus Min Filter ($SSDMF$) to estimate the disparity map [18] in equation (A.13). The disparity map is a gray image with black denoting the smallest disparity 0 pixel and white being the largest 255 pixel.

For evaluating each metric V, root mean squared error (RMSE), Pearson correlation coefficient, and Spearman rank-order correlation coefficient have been selected as the evaluation criteria. They are calculated between objective values MOS_p after a nonlinear regression using Equation (A.17), suggested by the VQEG, and the subjective scores MOS.

$$MOS_p = b_1/(1 + exp(-b_2 \times (r(V) - b_3))) \tag{A.17}$$

where b_1, b_2 and b_3 are the regression coefficients, $r(V)$ is the raw value calculated from metric V, and exp is the exponential function. The main purpose of equation (A.17) is to unify $r(V)$ for each metric to the range of MOS. Table A.7 reports the evaluation results.

According to the evaluation results, the objective metric for crosstalk perception V_{pdep} can achieve a higher correlation against the subjective MOS values when compared to traditional 2D metrics (V_{psnr} and V_{ssim}), and other metrics (V_{dep}, V_{pdis} and V_{dis}). The performance of the proposed metric is better than V_{psnr} and V_{ssim}, which indicates that the metric taking 3D characteristics into consideration can give a better prediction of crosstalk perception than 2D metrics. However, different combinations of 3D characteristics might have different prediction capabilities. It can be seen from Table A.7 that the metric V_{dis} has a worse performance than its counterpart V_{ssim}. Moreover, the performance of V_{pdep} and V_{pdis} is better than the

corresponding metrics V_{dep} and V_{dis}, respectively. This indicates that weighting the region of visible crosstalk only might be in accordance with users' perception. However, V_{pdis} exhibit slightly poorer performance when compared to the proposed metric V_{pdep}, which implies that relative depth instead of absolute depth is more suitable for the weighting operation. Therefore, V_{pdep} that models the perceptual attributes of crosstalk has the best performance. As the Pearson correlation of the proposed metric V_{pdep} is 88.4%, it is promising for evaluating the crosstalk perception of stereoscopic images.

In order to have a closer look at the proposed metric of crosstalk perception V_{pdep} in our subjective dataset, we validated its performance on different scene contents. Figure A.8 shows the scatter plot of the MOS values versus predicted quality values MOS_p on different scene contents. Based on the experimental results, the performance of the proposed metric does not have a significant difference between scene content while the impairments levels can significantly influence the performance. In particular, the proposed metric has better performance in predicting crosstalk perception of stereoscopic images with low and high impairments than images with medium impairments. We think that the performance difference might originate from the filtered depth map where the dominating perception on the maximum crosstalk of different impairments levels should be considered. However, this conclusion needs to be further verified in future work.

A.5 Conclusion

In this paper, we have conducted subjective tests for stereoscopic crosstalk perception with varying parameters of scene content, camera baseline and crosstalk level. The statistical results show that crosstalk level, camera baseline and scene content have significant impacts on crosstalk perception, respectively, and they have 2-factors interactions in terms of the impact on crosstalk perception. Moreover, the perceptual attributes (shadow degree, separation distance and spatial position) of crosstalk are summarized by observing the visual variations when the significant factors (crosstalk level, camera baseline and scene content) change. These perceptual attributes are the sensorial results of the HVS and classified into two categories: 2D and 3D perceptual attributes. Subsequently, an objective metric for

Table A.7: Evaluation results of different metrics on subjective dataset

Metrics	**RMSE**	**Pearson**	**Spearman**
V_{psnr}	0.465	0.821	0.763
V_{ssim}	0.461	0.825	0.784
V_{pdep}	0.382	0.884	0.859
V_{dep}	0.448	0.836	0.844
V_{pdis}	0.416	0.860	0.808
V_{dis}	0.574	0.709	0.688

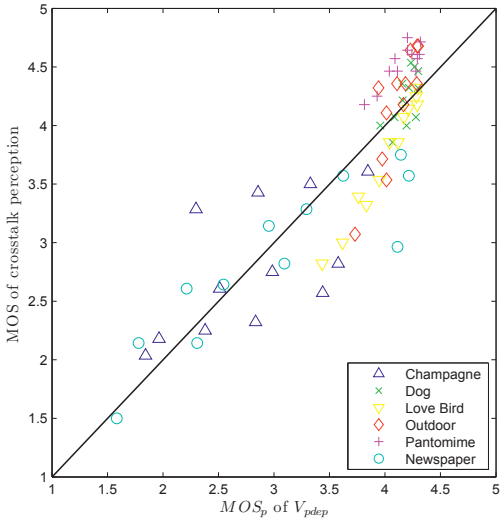


Figure A.8: Scatter plot of MOS of crosstalk perception versus predicted values MOS_p.

crosstalk perception has been proposed by combining SSIM map and filtered depth map. The experimental results with respect to our subjective evaluation scores have demonstrated promising performance of this metric, achieving more than 88% correlation with the MOS results. The performance of the proposed quality metrics is better than traditional 2D models and other compared metrics with different combination methods.

Acknowledgment

The authors would like to thank Professor Leif Arne Rønningen for kindly allowing us to use his lab Caruso for subjective tests and Dr. Jie Xu for his helpful discussions and suggestions throughout this work.

References

[1] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau. Quality assessment of stereoscopic images. *EURASIP Journal on Image and Video Processing*, 2008(1):659024, 2008.

[2] A. Boev, D. Hollosi, and A. Gotchev. Software for simulation of artefacts and database of impaired videos. In *Mobile 3DTV Project report, No. 216503.*, [available] http://mobile3dtv.eu.

[3] A. Boev, D. Hollosi, A. Gotchev, and K. Egiazarian. Classification and simulation of stereoscopic artifacts in mobile 3DTV content. pages 72371F–72371F–12, 2009.

[4] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet. New requirements of subjective video quality assessment methodologies for 3DTV. In *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2010.

[5] L. Goldmann, F. De Simone, and T. Ebrahimi. A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. In *Three-Dimensional Image Processing (3DIP) and Applications*, volume 7526, 2010. [available] http://mmspg.epfl.ch/3diqa, http://mmspg.epfl.ch/3dvqa.

[6] P. Gorley and N. Holliman. Stereoscopic image quality metrics and compression. In *Stereoscopic Displays and Applications XIX*, volume 6803, pages 680305–680305–12, 2008.

[7] K.-C. Huang, J.-C. Yuan, C.-H. Tsai, W.-J. Hsueh, and N.-Y. Wang. A study of how crosstalk affects stereopsis in stereoscopic displays. In *Stereoscopic Displays and Virtual Reality Systems X*, volume 5006, pages 247–253, 2003.

[8] ISO/IEC JTC1/SC29/WG11. M15377, M15378, M15413, M15419. 2008.

[9] ITU-R BT.1438. Subjective assessment of stereoscopic television pictures. 2000.

[10] ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television pictures. 2002.

[11] D. Kim, D. Min, J. Oh, S. Jeon, and K. Sohn. Depth map quality metric for three-dimensional video. In *Stereoscopic Displays and Applications XX*, volume 7237, pages 723719–723719–9, 2009.

[12] F. L. Kooi and A. Toet. Visual comfort of binocular and 3D displays. *Displays*, 25:99–108, 2004.

[13] M. Lambooij, M. Fortuin, I. Heynderickx, and W. IJsselsteijn. Visual discomfort and visual fatigue of stereoscopic displays: A review. *Journal of Imaging Science and Technology*, 53(3):30201–1–30201–14, 2009.

[14] L. Lipton. Factors affecting 'ghosting' in time-multiplexed plano-stereoscopic CRT display systems. In *True 3D Imaging Techniques and Display Technologies*, volume 761, pages 95–78, 1987.

[15] R. Olsson and M. Sjostrom. A depth dependent quality metric for evaluation of coded integral imaging based 3D-images. In *3DTV Conference, 2007*, pages 1–4, 2007.

[16] S. Pastoor. Human factors of 3D images: Results of recent research at Heinrich-Hertz-Institut Berlin. In *International Display Workshop*, volume 3, pages 69–72, 1995.

[17] Z. Sazzad, S. Yamanaka, Y. Kawayokeita, and Y. Horita. Stereoscopic image quality prediction. In *Quality of Multimedia Experience (QoMEX), 2009 International Workshop on*, pages 180–185, 2009.

[18] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47(1-3):7–42, 2002.

[19] P. Seuntiëns, L. Meesters, and W. IJsselsteijn. Perceptual attributes of crosstalk in 3D images. *Displays*, 26(4–5):177–183, 2005.

[20] SSIM implementation. http://www.ece.uwaterloo.ca/∼z70wang/research/ssim/.

[21] M. Tanimoto, T. Fujii, K. Suzuki, and et al. Reference softwares for depth estimation and view synthesis. In *ISO/IEC JTC1/SC29/WG11 MPEG2008/M15377*, 2008.

[22] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[23] A. Woods. Understanding crosstalk in stereoscopic displays. In *Three-Dimensional Systems and Applications*, 2010.

[24] L. Xing, T. Ebrahimi, and A. Perkis. Subjective evaluation of stereoscopic crosstalk perception. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 77441V–77441V–9, 2010.

[25] L. Xing, J. You, T. Ebrahimi, and A. Perkis. A perceptual quality metric for stereoscopic crosstalk perception. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 4033–4036, 2010.

[26] J. You, G. Jiang, L. Xing, and A. Perkis. Quality of visual experience for 3D presentation - stereoscopic image. In M. Mrak, M. Grgic, and M. Kunt, editors, *High-Quality Visual Experience*, Signals and Communication Technology, pages 51–77. Springer Berlin Heidelberg, 2010.

B Subjective Crosstalk Assessment Methodology for Auto-stereoscopic Displays

Author

Liyuan Xing
Jie Xu
Kim Skildheim
Touradj Ebrahimi
Andrew Perkis

Conference

IEEE International Conference on Multimedia & Expo (ICME), 2012.

Abstract

Crosstalk is one of the most annoying distortions in the visualization stage of stereoscopic systems. Specifically, both pattern and amount of crosstalk in multi-view autostereoscopic displays are more complex because of viewing angle dependability, when compared to crosstalk in 2-view stereoscopic displays. Regarding system crosstalk there are objective measures to assess it in auto-stereoscopic displays. However, in addition to system crosstalk, crosstalk perceived by users is also impacted by scene content. Moreover, some crosstalk is arguably beneficial in auto-stereoscopic displays. Therefore, in this paper, we further assess how crosstalk is perceived by users with various scene contents and different viewing positions using auto-stereoscopic displays. In particular, the proposed subjective crosstalk assessment methodology is realistic without restriction of the users viewing behavior and is not limited to the specific technique used in auto-stereoscopic displays. The test was performed on a slanted parallax barrier based auto-stereoscopic display. The subjective crosstalk assessment results show their consistence to the system crosstalk meanwhile more scene content and viewing position related crosstalk per-

ception information is provided. This knowledge can be used to design new crosstalk perception metrics.

B.1 Introduction

The human visual system (HVS) reconstructs three-dimensional (3D) perception from two-dimensional (2D) left and right retinal images. A stereoscopic 3D system mimics the HVS behavior by simultaneously capturing a pair of 2D images from slightly different positions, and then separately delivering them to respective eyes. Consequently, 3D perception is generated in the HVS. This simple principle of stereoscopic 3D was first demonstrated by Charles Wheatstone in his published drawings in 1833.

Stereoscopic display is a key device in such a stereoscopic 3D system, and various techniques are used to (de)multiplex the different views, such as wavelength, polarization, time and angle, as reviewed in [7]. The former three techniques are usually adopted in stereoscopic displays which require wearing viewing anaglyph glasses, polarized glasses or shutter glasses, respectively. The last angle based technique is often implemented in auto-stereoscopic displays which do not require any glasses. Instead, an optical filter is added in front of the screen (e.g. parallax barriers, lenticular lenses).

However, none of the aforementioned techniques can avoid crosstalk, which is one of the most annoying distortions in the visualization stage [12]. In particular, crosstalk is produced by imperfect view separation that causes a small proportion of one eye image to be seen by the other eye as well. Crosstalk artefacts are usually perceived as ghosts, shadows, or double contours by human subjects.

Since the mechanisms behind occurrence of crosstalk are significantly different across these techniques, e.g. see a review in [15], crosstalk in certain (auto-)stereoscopic displays exhibits different characteristics both in pattern and amount. Usually the crosstalk in auto-stereoscopic displays is more complex when compared to stereoscopic displays because of multi-view and viewing angle dependability. Specifically, objective crosstalk measurement is developed to derive system crosstalk from the measurement of the sensors' output luminance (e.g. camera [2], Fourier optics instrument [3]) or perceived luminance by the observers [1] versus viewing angle. It shows that the system crosstalk in auto-stereoscopic display has the following three features which distinguish itself from the one in stereoscopic display: a) the crosstalk depends on the observation positions; b) most of the crosstalk comes from neighbor views but also other views; c) the amount of crosstalk from neighbor views varies along the horizontal and vertical axes of the screen.

In addition to system crosstalk, the perceived crosstalk by end users, when a stereoscopic media is displayed, is impacted by more factors. Characteristics of displayed scene content, such as binocular parallax [12, 16, 10], depth structure [16], image contrast [16, 10, 8], edges [16, 8], texture and details [16, 8] are found to have an impact on users' perception of crosstalk. Moreover, some crosstalk is argued to have benefit in multi-view auto-stereoscopic displays in [6]. Therefore, we are interested in assessing the crosstalk perception when both the system crosstalk features of auto-stereoscopic displays and characteristics of scene content are taken into consideration. Consequently, a subjective crosstalk assessment methodology for auto-stereoscopic displays is proposed. This methodology is realistic without

restriction of subjects' viewing behavior and not limited to the specific technique used in auto-stereoscopic displays. Thereby, a comprehensive database of crosstalk perception on auto-stereoscopic display for a wide variety of situations has been created, which can be used to design position and scene content sensitive crosstalk perception metrics.

The remainder of this paper is organized as follows. In next section, assessment set up is described. In Section B.3, subjective crosstalk assessment is carried out. The subjective results are analyzed in Section B.4. Finally, concluding remarks are given in Section B.5.

B.2 Assessment Set Up

Several equipments and systems are needed to construct the realistic crosstalk assessment environment.

B.2.1 Auto-stereoscopic Display

It should be noted that our subjective crosstalk test methodology can be used for any auto-stereoscopic display with all kinds of implementation techniques. In this test, we adopted the 'Tridelity MV4200' [13] auto-stereoscopic display which is available in our lab. The Tridelity display is a slanted parallax barriers auto-stereoscopic display and supports either 5 or 9 views. In our case, we displayed the test stimuli by choosing the 5 views with pattern of $\cdots$ V3, V4, V5, V1, V2, V3, V4, V5, V1, V2, V3 $\cdots$ horizontally distributed and having the native resolution of 1936×1360 pixels. In theory, the worst crosstalk happens when V5 and V1 are seen by one eye at the same time since the difference between these two views is the most significant. The suggested optimal viewing distance of the display is 3.5m [13] with neighbor views interval to be about 62mm.

We have checked that the Tridelity display follows the three features of system crosstalk in auto-stereoscopic display as summarized in Section B.1. Moreover, we have observed that both view interval and system crosstalk pattern from neighbor views vary along the viewing distance. Figure B.1 illustrates the captured images by camera when the displayed image having totally white image at V3 while totally black image at the other 4 views. It can be seen from the left image that interval between the white light beams from V3 becomes larger when the viewing distance is farther. Specifically, we also measured the interval between neighbor views which is 47mm, 63mm, and 72mm at the viewing distance of 2.6m, 3.6m, and 4.6m, respectively. At these viewing distances, we further captured the system crosstalk introduced from the neighbor views, as you can see in the right images from top downwards in Figure B.1. It shows that the system crosstalk (black) at near viewing distance is worst and becomes best around the optimal viewing distance but becomes worse again at far viewing distance.

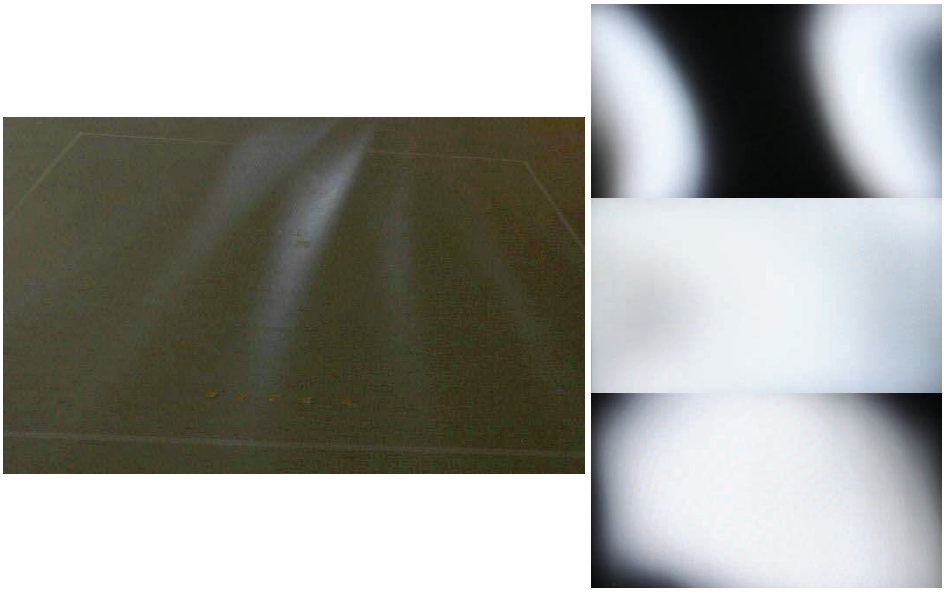


Figure B.1: Captured images by camera, view interval and crosstalk pattern.

B.2.2 Head Tracking System

With the rapid development of interactive human machine interface, several approaches with open sources available online can support the tracking of head position. For example, the Wiimote Virtual Reality Desktop [14] and Kinect Prime Sensor^{TM} NITE 1.3 Framework [9] can both be developed to a head tracking system. In order not to add any disturbance to the subjects, we decided to use Kinect system which does not require wearing anything by the subjects.

The Kinect system consists of a Kinect sensor and the software. In particular, the Kinect sensor is a horizontal bar with one RGB camera and two depth sensors. Both the output RGB and depth sensing videos of the Kinect sensor are at a frame rate of 30 Hz and in resolution of 640×480 pixels. The software enables advanced gesture recognition, facial recognition and tracking. The sensor can maintain tracking through an extended range of approximately 0.8m-6m. The resolution is about 1.3mm and 7.6mm per pixel at viewing distance of 0.8m and 4.5, respectively. The skeleton tracking is adopted in our tracking system. The basic assumption of skeleton tracking is that the user's upper body is mostly inside the field of view. Specifically, the skeleton tracking algorithm uses the label map output by a user segmentation process to generate a skeleton. A calibration pose is used to adjust the skeleton model to the user's body proportions within a few seconds and activate the head tracking accordingly. The output of the skeleton tracking is the positions and orientations of the skeleton joints. In our case, we only use the head position in world coordinates in units of mm. If the user is not visible for more than 10 seconds,

Figure B.2: Visual samples of the selected scenes. From top left to bottom right: Cafe, Outdoor, Pantomime, Dog, Love Bird, Poznan Street, Balloons, Poznan Hall, Book Arrival. Among which, the top row is for training.

the user is considered lost. Therefore, we restricted the subjects in a certain region in front of the auto-stereoscopic display as detailed in Subsection B.2.5.

B.2.3 Score Tracking

In addition to head tracking, we need track the score that the subjects input at corresponding head position as well. A Bluetooth Numpad was adopted such that the subjects can hold the Numpad while they move in the viewing region. When the subjects input a score using the number keyboard, both the pressed score and the current head position are recorded simultaneously by the developed software.

B.2.4 Scene Content

Based on our previous experience, scene content covering a wide range of depth structures, contrasts, colors, edges and textures is considered as a potential factor impacting users' perception of crosstalk. Correspondingly, nine multi-view sequences (three for training) from the MPEG [4] were chosen as representative scene contents, as shown in Figure B.2.

B.2.5 Lab Environment

auto-stereoscopic display is restricted in a 2.0m×2.0m square region. Specifically, suppose the world coordinates system is defined as follows, the origin point O is

the display center, the XY plane is the display plane with X axis horizontal and Y axis vertical to the floor, and Z axis is vertical to the XY plane and facing to the subjects. The square moving region has X from -1.0m to 1.0m and Z from 2.5m to 4.5m. This region was chosen based on both the optimal viewing distance of our auto-stereoscopic display and the tacking range of the Kinect sensor. Both the display and Kinect were put on a table at the same display plane and the distance between their centers to the floor is 2.4m and 2.1m, respectively.

The sitting chair for subjects with five wheels can be moved freely on the floor. Moreover, the height of the chair can be adjusted by the subjects easily. The subjects who sit on the chair should have their eyes in the same height as the display center. We used 5 same-height anchor points hanged from ceil to calibrate the height of the subjects' eyes. Four anchor points marks the four corners of the moving region and one more at the middle of back corners.

During the subjective tests, all the doors and windows of the test room were closed and covered by black curtains. In addition, the lights in the room were turned off except for two reading lamps behind the auto-stereoscopic display as a background. We used the EyeOne Display2 to measure color temperature and illuminance of background, which were 2500K and 63lux, respectively. Moreover, the measured color temperature, gamma, and luminance on the screen were 6500K, 2.2 and $95.5cd/m^2$, respectively.

B.3 Subjective Crosstalk Assessment

This section provides the details about the revised Single Stimulus (SS) test methodology that we used to conduct the subjective crosstalk assessment.

B.3.1 Single Stimulus

The most important reason that the double stimulus methods cannot be adopted in this study is that there is no existence of the reference pictures for crosstalk perception, since there is always system crosstalk by the auto-stereoscopic display and perceptible to the subjects. The SS method was also used in assessing the quality levels of stereoscopic images with varying camera baselines in the literature [11, 16] when it is difficult to choose the reference.

B.3.2 Training Sessions

Each subject had to complete a training session to get an idea of how to evaluate the test stimuli. As listed in Table B.1, five categorical adjectival levels with illustrated examples were used to benchmark and harmonize the subjects' measuring scales. These examples were selected by expert viewers in such a way that each quality level was represented and that these examples could cover a full range of quality levels within the set of test stimuli. The scale is discrete from 1 to 5 in a way of lowest to highest quality. When each example was displayed, the operator verbally

Table B.1: Explanations of five categorical adjectival levels and their training examples

	Explanation	**Example**
1	Very annoying: Crosstalk is so much, that the 3D perception is hardly to be formed and you feel a little uncomfortable.	Café-Worst
2	Annoying: Crosstalk is much, although the 3D perception still can be formed but you refuse to accept viewing such quality in daily life.	Café-Best
3	Slightly annoying: there is obvious crosstalk. However you can accept viewing such quality reluctantly.	Outdoor-Worst
4	Perceptible but not annoying: you can see a little bit crosstalk at a first glance, but the quality of the whole image is still good.	Outdoor-Best
5	Imperceptible: you cannot see any crosstalk or you can perceive very slightly only when you pay special attention to a certain region.	Pantomime-Best

explained the corresponding quality level to the subject as described in Table B.1. For example, when the training picture Cafe was shown, the subject was told to move his position to find a position where he perceived the worst crosstalk, and quality level 'very annoying' was explained. Then, he was told to move to a position where he perceived the best crosstalk and quality level 'annoying' was stated. The same procedure was repeated for training pictures Outdoor and Pantomime for other three quality levels 'slightly annoying', 'perceptible but not annoying' and 'imperceptible'.

Subjects were encouraged to view the representative examples as long as they wished and asked questions if they needed any further clarifications. Specifically, it was emphasized that the subjects should ignore color changes and other abnormalities that were not related to crosstalk. The training sessions would continue until subjects could understand and distinguish the five different quality levels.

B.3.3 Test Sessions

The head tracking system was activated by the calibration pose as mentioned earlier when the test session started. The test session for every subject consisted of 6 sub sessions. In each sub session, one multi-view image, randomly chosen from the test stimuli (Dog, Love Bird, Poznan Street, Balloons, Poznan Hall, Book Arrival), was shown 7 minutes. During the sub session, the subject should

- Move his body freely in the viewing region without following any static patterns or searching for high or low values patterns, but keep head not slanted.
- Judge the quality without thinking so much and trust his first feeling.

- Enter the score using Numpad while keep his back on the chair as he did during training and his head still in the exact position where the score is corresponding to.
- Cover as many positions as possible in the moving region.

These sub sessions were continued with one minute break in between. During the test period, the subjects were not allowed to ask questions in order to avoid any interruption during the entire session.

B.3.4 Subjects

Before the training sessions, a visual screening test was performed to the subjects, including normal or corrected vision acuity larger than 1.0 (tested by the Snellen chart), no colorblind vision (tested by the Ishihara), and stereo vision smaller than 30 arcsecond (tested by the Randot SO-002). Pupillary distance of subjects (measured by a ruler) was also collected. A total of 25 qualified subjects (15 males and 10 females, aged from 21 to 50 years old) took part in the tests.

B.4 Subjective Results Analysis

The viewing position related subjective scores of the 6 scene contents given by 25 subjects are analyzed in this section. Particularly, we observed the relationship between the subjective and objective crosstalk measurement both in raw data and interpolated data ways.

B.4.1 Two Analysis Ways

Since all the subjects had their own moving trajectory in the viewing region and did not cover almost any same position among the subjects, the concept mean opinion score (MOS), confidential interval (CI) and screening of subjects defined in the ITU-R BT.500 [5] as common methods of results analysis cannot be applied in this case anymore. Therefore, we proposed two analysis ways which are based on raw data and interpolated data respectively to observe the relationship between the subjective and objective crosstalk measurement.

The raw data approach is based only on the raw subjective scores obtained from the subjects, without any interpolation to generate interpolated subjective scores, while the interpolated data approach is based also on the interpolated subjective scores from the raw subjective scores. By interpolation, a surface passes through all the raw subjective scores is estimated, therefore, we are able to calculate the values of the surface at any position between the known raw subjective scores. In particular, we adopt the cubic interpolation, which is the simplest method that offers true continuity between the raw subjective scores. The first and second rows of Figure B.3 and B.4 are the raw data scatter as a function of viewing positions for all the six scene contents separately, and the interpolated data surface from raw

subjective scores correspondingly. For both approaches, we further plot the slices at viewing distance of 3.5m, as shown in the third and fourth rows of Figure B.3 and B.4, where the green curves are the direct slice curve, while the red curves are the $SinN(a_1 \times sin(b_1 \times x + c_1) + \cdots + a_N \times sin(b_N \times x + c_N))$ regression curves from the direct slice data. In particular, the value N is 3, 4 or 5 for different scene contents depending on the period experimentally.

B.4.2 Subjective versus Objective Crosstalk Measurement

In theory, the scores are viewing position related and follow certain distribution according to characteristics of the auto-stereoscopic display. As it can be seen from the raw data scatter and interpolated data surface in the first and second rows of Figure B.3 and B.4, the best (red) and worst (blue) subjective scores occur periodically based on the viewing position, especially obvious on the interpolated data surface. In particular, when the Z distance is larger, the period exhibited in X axis becomes larger for all the scene content in Figure B.3 and B.4, which is because of the interval between neighbor views becomes larger as measured in Subsection B.2.1. Moreover, the crosstalk is more annoying (blue) at the near Z distance than far Z distance as you can see from scene content Balloons, Poznan Hall and Book Arrival. It also can be noticed in the right images from top downwards in Figure B.1. Therefore, the subjective crosstalk assessment is consistent with the objective crosstalk measurement. Additionally, Figure B.3 and B.4 shows the obvious difference among the subjective scores in various scene contents. Thus, the subjective crosstalk assessment is more comprehensive than the objective crosstalk measurement, which is also scene content related.

If we further look at the slice of the raw data scatter and interpolated data surface at viewing distance of 3.5m, more details regarding the horizontal distribution can be seen. Both the green curves at third and fourth rows of Figure B.3 and B.4 are not smooth, we believe, it is because of the sparse collected raw data. According to the theory and objective crosstalk measurement, there should be a perfect periodical curve having 6.45 cycles in 2 meters (6.45≈2m/62mm/5). Thus, SinN regression is used on the raw and interpolated data respectively to approve the perfect periodical curve. It can be seen that most of the red curves have 6 or 7 cycles, which is also consistent with the objective measurement.

B.5 Conclusion and Further Work

In this paper, we have proposed a novel and realistic subjective crosstalk assessment methodology on auto-stereoscopic display with consideration of different scene contents and viewing positions. The subjective results show their consistence to the characteristics of auto-stereoscopic display, while provide more detailed crosstalk perception information in terms of viewing position and scene content when compared to the objective crosstalk measurement. Furthermore, the resulting interpolated surface can be used to design of crosstalk perception metrics for auto-

stereoscopic display. However, based on the analysis on the collected data, it indicates that this methodology can be improved in ways of increasing the head tracking accuracy, decreasing the discrepancy among subjects by training, having post processing (e.g. region based MOS and outlier removal) under the condition of relatively dense raw data.

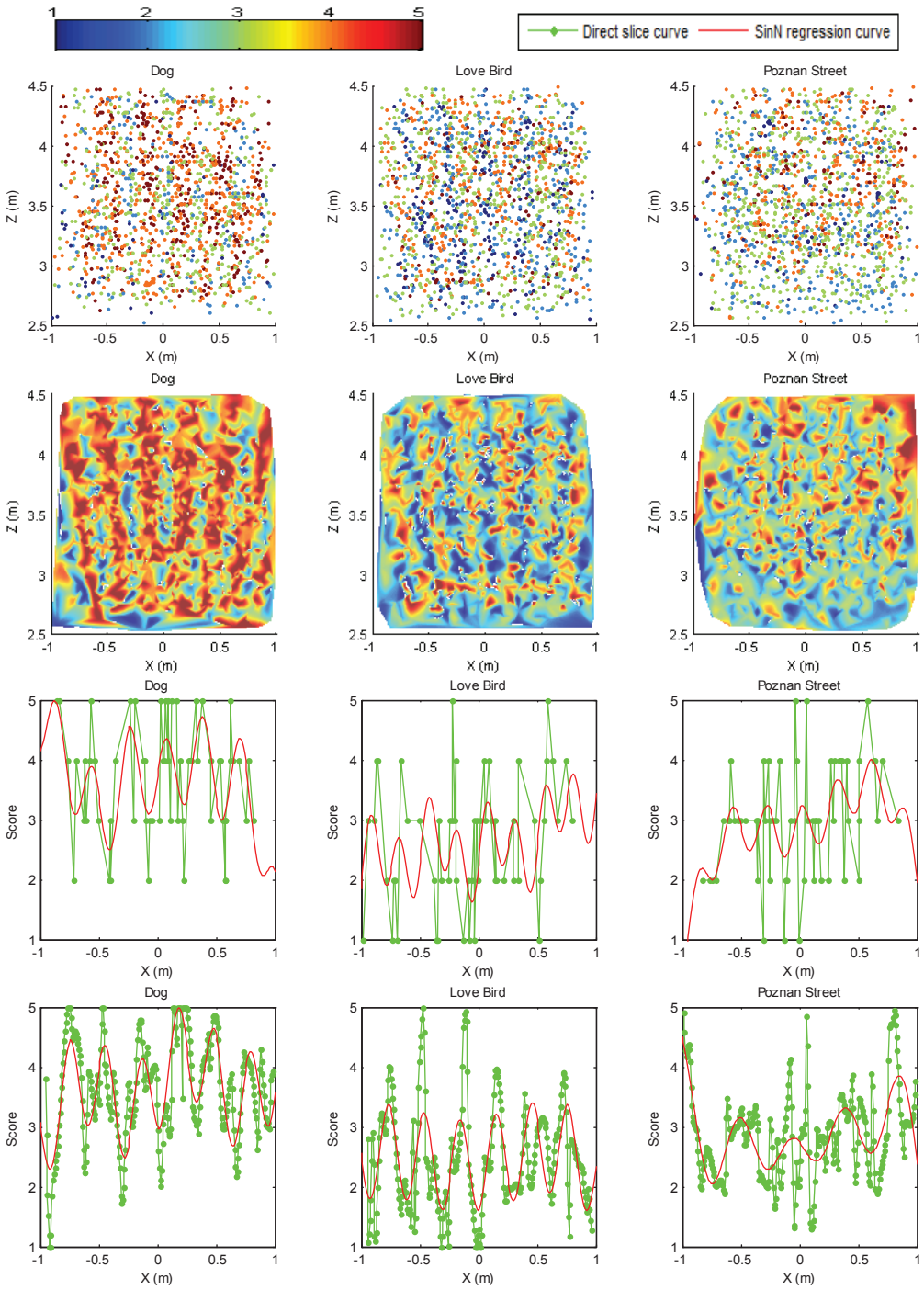


Figure B.3: Visualization of subjective scores on crosstalk perception for scene contents. From top to bottom: raw data scatter, interpolated data surface, slice at distance of 3.5m on raw data, slice at distance of 3.5m on interpolated data.

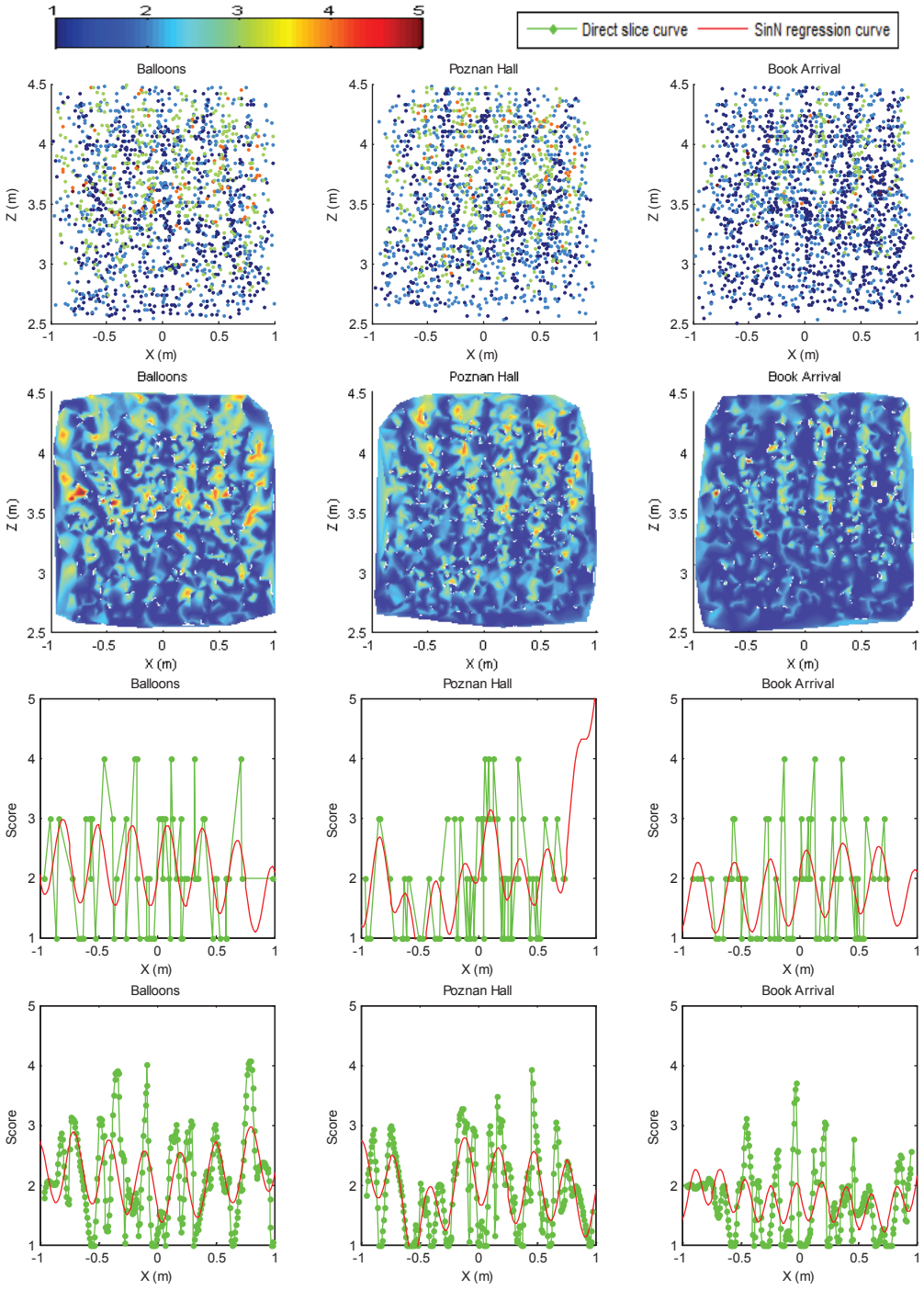


Figure B.4: Visualization of subjective scores on crosstalk perception for scene contents. From top to bottom: raw data scatter, interpolated data surface, slice at distance of 3.5m on raw data, slice at distance of 3.5m on interpolated data.

References

[1] M. Barkowsky, P. Campisi, P. Le Callet, and V. Rizzo. Crosstalk measurement and mitigation for autostereoscopic displays. In *Crosstalk measurement and mitigation for autostereoscopic displays*, San José, United States, 2010.

[2] A. Boev, A. Gotchev, and K. Egiazarian. Crosstalk measurement methodology for auto-stereoscopic screens. In *3DTV Conference, 2007*, pages 1–4, 2007.

[3] P. Boher, T. Leroux, T. Bignon, and V. Collomb-Patton. A new way to characterize autostereoscopic 3D displays using fourier optics instrument. In *Stereoscopic Displays and Applications XX*, volume 7237, pages 72370Z–72370Z–12, 2009.

[4] ISO/IEC JTC1/SC29/WG11. M15377, M15378, M15413, M15419. 2008.

[5] ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television pictures. 2002.

[6] A. Jain and J. Konrad. Crosstalk in automultiscopic 3D displays: blessing in disguise? In *Stereoscopic Displays and Virtual Reality Systems XIV*, volume 6490, pages 649012–649012–12, 2007.

[7] O. Levent. 3D video technologies: An overview of research trends. In *SPIE Electronic Imaging*, volume PM196, pages 1–116, 2011.

[8] L. Lipton. Factors affecting 'ghosting' in time-multiplexed plano-stereoscopic CRT display systems. In *True 3D Imaging Techniques and Display Technologies*, volume 761, pages 95–78, 1987.

[9] NITE. http://andrebaltazar.files.wordpress.com/2011/02/nite-controls-1-3-programmers-guide.pdf.

[10] S. Pastoor. Human factors of 3D images: Results of recent research at Heinrich-Hertz-Institut Berlin. In *International Display Workshop*, volume 3, pages 69–72, 1995.

[11] P. J. H. Seuntiens, L. M. J. Meesters, and W. A. IJsselsteijn. Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation. *TAP*, 3(2):95–109, 2006.

[12] P. Seuntiëns, L. Meesters, and W. IJsselsteijn. Perceptual attributes of crosstalk in 3D images. *Displays*, 26(4–5):177–183, 2005.

[13] Tridelity. http://www.tridelity.com/3d-display-mv4210va.3d-display-mv4200.0.html.

[14] Wiimote. http://channel9.msdn.com/coding4fun/articles/wiimote-virtual-reality-desktop.

[15] A. Woods. Understanding crosstalk in stereoscopic displays. In *Three-Dimensional Systems and Applications*, 2010.

[16] L. Xing, T. Ebrahimi, and A. Perkis. Subjective evaluation of stereoscopic crosstalk perception. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 77441V–77441V–9, 2010.

C Stereoscopic Quality of Experience: Subjective Assessment and Objective Modeling

Author

Liyuan Xing
Junyong You
Touradj Ebrahimi
Andrew Perkis

Journal

Submitted to IEEE Transactions on Multimedia (TMM), 2013.

Abstract

Stereoscopic three-dimensional (3D) services have become more popular recently amid promise of providing immersive quality of experience (QoE) to the end-users. In practice, stereoscopic QoE can be influenced by a complex combination of different factors. In this work, we focus on an analysis of stereoscopic QoE based on the simplest stereoscopic imaging system (including indispensable capturing and displaying stages) and its requisite factors. A series of subjective stereoscopic quality assessments have been conducted to investigate the influence of several requisite factors on stereoscopic QoE, including scene content, camera baseline, screen size and viewing position. Thereby, a comprehensive database for stereoscopic QoE was available. Moreover, based on our previous work on crosstalk assessment, crosstalk level has been recognized as another requisite factor in the simplest stereoscopic systems. Thereafter, two perceptual attributes of stereoscopic QoE, namely, crosstalk perception and depth enabled visual comfort, are summarized as the sensorial results of the human visual system (HVS). Their relationships with the explored requisite factors are analyzed and modeled in a mathematical approach. The models of perceptual attributes are further combined into an objective quality metric for predicting stereoscopic QoE. The experimental results demonstrate that the

proposed metric has a high correlation (over 85%) when compared with subjective quality scores in a wide variety of situations.

C.1 Introduction

Stereoscopic three-dimensional (3D) media services have recently become popular thanks to their capability to provide binocular depth which is expected to be the next big step forward in the evolution of media after addition of color and sound information. Offering the binocular depth by stereoscopic technology can bring new experience to viewers, rather than a simple enhancement in the quality sense.

However, stereoscopic 3D services do not always prevail when compared to their two-dimensional (2D) counterparts. Nowadays, stereoscopic imaging technology is usually based on simultaneously capturing a pair of two-dimensional (2D) images using stereo cameras, and then separately delivering them to respective eyes by a screen based stereoscopic display. Consequently, 3D perception is generated in the brain through the human visual system (HVS). Although the fundamental principle of stereoscopic 3D imaging technology sounds straightforward, it is rather difficult to implement in practice. The simplest stereoscopic system, consisting of the basic capturing and displaying stages without coding, transmission and representation steps, usually fails to provide satisfying viewing experience due to various binocular artefacts. Binocular artefacts that can potentially degrade the stereoscopic quality of experience (QoE) include depth plane curvature, keystone-distortion, cardboard effect in the capturing stage and convergence-accommodation rivalry, interocular crosstalk, shear distortion, puppet theater effect, picket fence effect in the displaying stage [6]. Under the limitation of current 3D imaging techniques, most binocular artefacts cannot be eliminated completely.

Due to the existence of binocular artefacts, stereoscopic QoE might be influenced significantly by different visual factors, such as characteristics of content (e.g. contrast, depth structure), configuration of capturing and displaying systems (e.g. camera baseline, system-introduced crosstalk, screen size, screen resolution, pixel width, viewing position), selection of image processing mechanism (e.g. compression, conversion, view synthesis), perceptual-physiological limitations of HVS (e.g. panum's fusional area, convergence-accommodation mismatch), cognitive-emotional factors of viewer (e.g. expectation, attitude, viewing context, culture), as reviewed in [10]. Stereoscopic QoE can be influenced by a complex combination of aforementioned factors. Moreover, these influence factors might also have strong interactions between each other. In order to provide the best experience to costumers of 3D services, it is crucial and beneficial to investigate the impact of the influence factors on the overall stereoscopic QoE qualitatively and quantitatively. Subjective quality testing methodologies are often utilized to perform such tasks.

Some subjective quality tests [25, 29, 17, 18, 24, 3] are in their principle focused around how the distortions introduced by image processing technologies affect the stereoscopic QoE. Compression related factors [18, 24, 3] are widely studied. Research efforts have also been dedicated to studying the influence of different capture and display configurations on stereoscopic QoE independently, such as camera baseline [25, 24, 11, 8], screen size [12] and viewing position [21, 15]. Though these studies strengthen our knowledge about stereoscopic perception mechanism, a comprehensive understanding of how the influence factors in the simplest stereoscopic

system, namely, requisite factors, affect stereoscopic QoE is still missing. Although the grid patterns distorted by the stereoscopic display system are visualized for a rectilinear grid in front of the camera system under various settings of most requisite factors in [28], further influence of requisite factors on the perceived quality for human subjects needs to be evaluated quantitatively.

Nevertheless, a subjective test is usually expensive in time and labor, and unsuitable for real-time applications. To overcome these drawbacks, objective metrics that can automatically predict the perceived quality are desired. Such metrics can be employed to monitor, benchmark, and optimize a stereoscopic media system. However, an effective quality metric cannot be proposed without a deep understanding of the perception mechanism of stereoscopic 3D presentations.

Development of objective quality models for stereoscopic media is still in its early stages. Researchers first started with exploring how well 2D quality metrics can be applied to stereoscopic quality assessment [4, 36]. Subsequently, a few objective metrics taking into account the characteristics of stereoscopic images have been proposed [20, 9, 22]. However, most objective stereoscopic metrics are designed to assess quality degradations caused by lossy compression schemes. To model the non-compression quality degradations, the authors of [16] proposed a metric linearly combining three quality measures, including perceived depth, visual fatigue and temporal consistency. We have also proposed a metric for modeling crosstalk perception [32, 35] based on the perceptual attributes of crosstalk. Specifically, perceptual attributes can be considered as the sensorial results of the HVS, which are useful for better understanding of perception mechanism in stereoscopic vision and further employed in the development of the objective quality metric.

In this work, subjective tests [33] have been conducted to investigate the influence of some requisite factors on stereoscopic QoE, including scene content, camera baseline, screen size and viewing position, in different conditions. Particularly, both the characteristics of test stimuli and the limitations of HVS have been taken into consideration explicitly when designing the test configurations. Moreover, another requisite factor, crosstalk level, has been recognized. Based on our understanding of requisite factors and stereoscopic QoE, two perceptual attributes, including crosstalk perception and depth enabled visual comfort, have been identified. Both perceptual attributes have been modeled mathematically and then combined into an objective metric to predict stereoscopic QoE. The metric has been validated against two databases containing wide different conditions of requisite factors and shown proposing performance.

The main contributions of this paper are twofold. First, the conducted subjective tests in the simplest stereoscopic system provide a comprehensive database (Q2S-QoE database) for QoE mechanism investigation. This database will be open to the public as long as the paper is accepted. Second, an objective quality metric has been proposed to predict the stereoscopic QoE based on an appropriate analysis of several requisite factors and perceptual attributes.

The remainder of the paper is organized as follows. In Section C.2, the details of the subjective tests are presented. Section C.3 explains the requisite factors and perceptual attributes of QoE, as well as their relationships. The objective QoE

metric and experimental results are presented in Section C.4. Finally, concluding remarks are given in Section C.5.

C.2 Subjective QoE Tests on Stereoscopic Images

Several recommendations for subjective quality evaluation of visual stimuli have been standardized by the International Telecommunication Union (ITU), e.g. the widely referenced ITU-R BT.500 [14] for television picture quality assessment. For subjective evaluation of stereoscopic television pictures, ITU-R BT.1438 [13] has made a few preliminary steps, but it still lacks of necessary details. The authors of [7] have summarized the lacks in a form of additional requirements. In our subjective tests, we followed these methodologies and further customized them for assessment of stereoscopic QoE. In the following, we will provide the details about the test laboratory environment where the evaluations were conducted, how the test configurations were designed, which test method was adopted, and what results were obtained from these tests.

C.2.1 Laboratory Environment

C.2.1.1 Display System

The polarization technique, as illustrated in Figure C.1, was used to present stereoscopic images in the tests. Specifically, two Canon XEED SX50 projectors with resolution of 1360×768 were placed on a chief ASE-2000 Adjusta-Set slide stacker. The stacker can be adjusted within $+/-7^\circ$ swivel, $+/-20^\circ$ tilt and $+/-7^\circ$ leveling ranges. Two Cavision linear polarizing glass filters with size of 4in×4in were installed orthogonally in front of the projectors. In this way, two views were projected and superimposed onto the backside of a 2.40m×1.20m silver screen. The distance between the projectors and the silver screen was about 3m, forming a projected region occupying the central area of the silver screen with a width of 1.98m and height of 1.06m. Images with different resolutions were displayed in their actual sizes in order to simulate different screen sizes. The subjects equipped with polarized glasses were asked to view the stereoscopic images on the opposite side of the silver screen. Both the configurations of screen size and viewing positions are designed in a holistic approach, as will be explained in detail in the following subsections.

C.2.1.2 Alignment of Display System

Prior to the tests, the display system was calibrated to align the two projectors. In particular, the positions of the two projectors were adjusted to guarantee that the center points of projectors, projected region and silver screen, positioned on the same horizontal line (central horizontal line as shown in Figure C.1) and the line was perpendicular to the silver screen. Moreover, the angles of stackers and the keystones of the projectors were adjusted with the help of projected Hermann grid

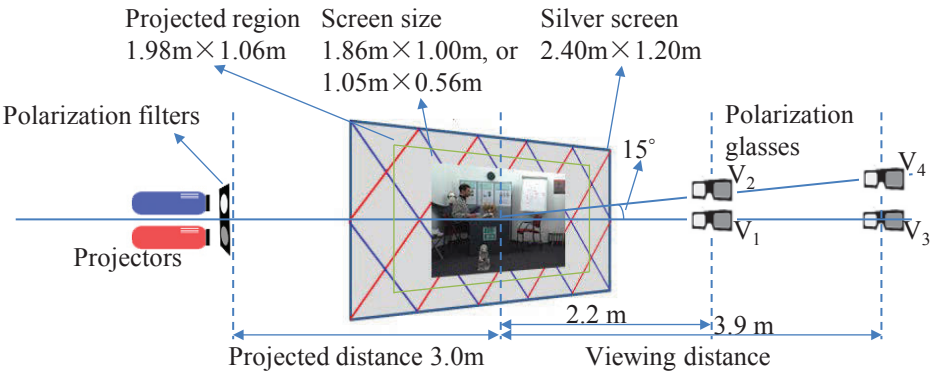


Figure C.1: The polarized display system.

images. The adjustment of display system was finished once the two Hermann grid images from the left and right projectors were exactly overlapped.

C.2.1.3 Measurement of System-introduced Crosstalk

After the alignment, system-introduced crosstalk was measured. As defined in [27], the terminology and mathematical definitions of crosstalk are diverse and sometimes contradictory. We adopt the definition of system-introduced crosstalk as the degree of the unexpected light leakage from the unintended channel to the intended channel. In particular, we measured the leakage in a situation where the left and right test images have the maximum difference in brightness. The system-introduced crosstalk is measured mathematically as follows,

$$P_l = \frac{Lx_{GL}(WB) - Lx_{GL}(BB)}{Lx_D(WB) - Lx_D(BB)} \tag{C.1}$$

where WB denotes a pair of test images (the left image is completely in white, whilst the right in black), and BB is another image pair both in black. Lx_D denotes the luminance measured on silver screen and Lx_{GL} denotes the luminance after the right lens of polarized glasses, which is cling to the silver screen. P_l denotes the system-introduced crosstalk from the left channel to the right, which was approximately 3% in our experiments.

C.2.1.4 Room Conditions

The test room had the length of 11.0m, width of 5.6m and height of 2.7m. During the subjective tests, all the doors and windows of the test room were closed and covered by black curtains. In addition, the lights in the room were turned off except for one reading lamp to shed light on the answering sheet when entering subjective scores.

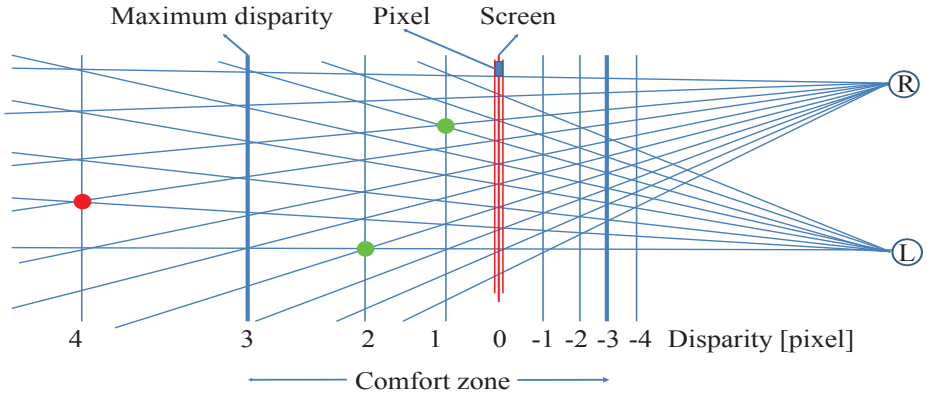


Figure C.2: The comfort zone of a display system and generated points of stimulus in the space.

C.2.2 Test Design

Binocular disparity is especially invoked in stereoscopic displays and presented to viewers for inducing the perception of stereoscopic depth. However, current stereoscopic displays often have convergence-accommodation mismatch that the screen is always accommodated regardless wherever the convergence point is located. This convergence-accommodation mismatch is not a natural function of the HVS and can further cause viewers discomfort. Therefore, the comfort zone for stereoscopic displays [19] is usually limited. The authors of [7] pointed out that such comfort zone can change across stereoscopic displays and the maximum disparity of comfort zone for certain displays can be calculated. As can be seen in Figure C.2, two green points located at the positions within the maximum disparity of comfort zone can be viewed comfortably, while the red point beyond the maximum disparity might cause uncomfortable viewing experience. A test stimulus with different disparity levels between the left and right images can generate plenty of green or red points at different positions in the space. Thus, we believe that the generated points located in different ranges of comfort zone result in different viewing experiences. Based on the calculated maximum disparity of the display system in our subjective quality experiments, different kinds of stimuli with different disparity levels were generated to cover a full range of QoE levels. In total 240 test cases (10 scene contents $\times$ 3 camera baselines $\times$ 2 screen sizes $\times$ 4 viewing positions) have been included, as detailed in the following subsection.

C.2.2.1 Maximum Disparity of Display

The maximum disparity of our display system is calculated using the following equations as in [7] and the viewing positions are further selected according to the

maximum disparity.

$$Z_f = d - \frac{1}{1/d + dof}, Z_b = \begin{cases} \frac{1}{1/d - dof} - d, & if \quad d < 5 \\ \infty, & if \quad d \geq 5 \end{cases} \tag{C.2}$$

$$D_f = \frac{Z_f \times e}{(d - Z_f) \times p_w}, D_b = \frac{Z_b \times e}{(Z_b + d) \times p_w} \tag{C.3}$$

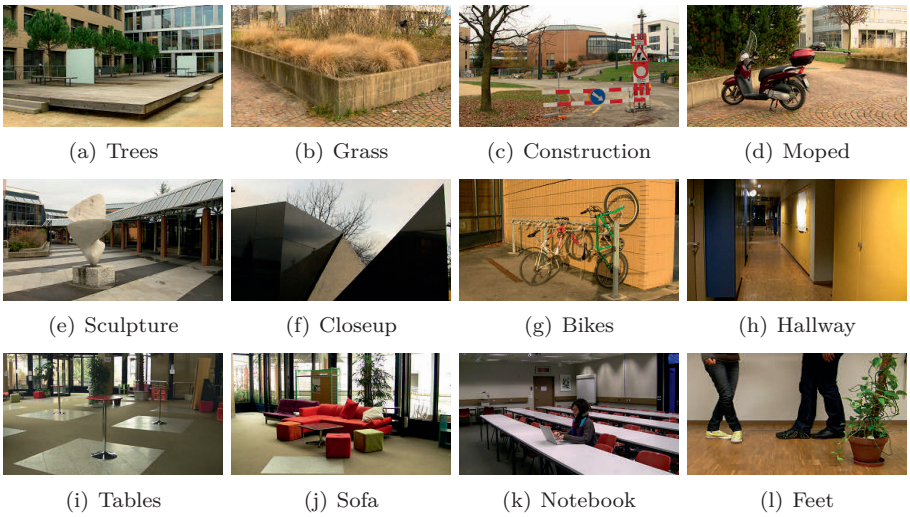
where d, e, p_w and dof are the viewing distance, pupillary distance, pixel width of the display and depth of focus (DOF), respectively, Z_f and Z_b denote the foreground and background distances of the comfort viewing zone in meter, and D_f and D_b are the corresponding maximum foreground and background disparity levels in terms of pixel. When d is below 5, D_f is equal to D_b. The values of e and dof are usually assumed to be 65mm and 0.2diopter, respectively. Additionally, p_w is also fixed in our display system, which is 1.46mm calculated by dividing the screen width (1.98m) by the horizontal resolution (1360). Therefore, the viewing distance is the only variable determining the maximum disparity.

We designed two viewing distances 2.2m and 3.9m resulting in the maximum disparity levels to be 20 and 35 pixels, respectively. Additionally, another two positions at 15° away from the center were used in order to investigate the impact of viewing position on stereoscopic QoE assessment. Subsequently, the four viewing positions were named as Near Center (V1), Near Side (V2), Far Center (V3), and Far Side (V4), respectively, as illustrated in Figure C.1.

C.2.2.2 Maximum Disparity of Test Stimuli

Test stimuli with a wide range of disparity levels were generated from different scene contents, camera baselines and image resolutions. These three factors determine disparity. Particularly, camera baseline reflects the general disparity information of all objects in a scene, while the disparity difference between different objects is dependent on relative depth structures in the scene content. Moreover, disparity is usually measured in a unit of pixel rather than metric unit; it is, thus, affected by the resolution of imaging sensor.

Five indoor and seven outdoor scenes, as shown in Figure C.3, were selected from a publicly available 3D image quality EPFL-QoE database [1, 2]. Among these scenes, the contents Trees and Grass were selected as training example and a dummy test stimulus, respectively. All the scene contents were originally captured at the resolution of 1920×1080 and stored in high quality without any perceptible compression distortions. The main consideration of combining outdoor and indoor scenes is to assess different depth ranges of scene content for various disparity structures. In addition, these scene contents also contain various complex features in contrast, textures and colors. The captured images had been applied by spatial and color alignment using a relative vertical and horizontal translation based on point correspondences and histogram matching, respectively, in the EPFL-QoE database. Thus no further post-processing was not required in our experiments.



(a) Trees (b) Grass (c) Construction (d) Moped

(e) Sculpture (f) Closeup (g) Bikes (h) Hallway

(i) Tables (j) Sofa (k) Notebook (l) Feet

Figure C.3: Snapshots of 12 scenes selected from the EPFL-QoE database.

For each scene content, three camera baselines (10cm, 20cm and 30cm) and two resolutions (1280×720 and 720×405) were selected to constitute different disparities, which were located in different ranges of maximum disparity levels (20 and 35 pixels) of our display. In particular, three baselines were shot originally, but two resolutions were down-sampled from the original one of 1920×1080. Consequently, two screen sizes corresponding to these two resolutions were simulated because all the images were presented in their actual sizes. Since the viewers watched the images on the screen directly, the influence factor of resolution can be actually replaced by screen size and it was divided into two categories, namely, Large (1.86m×1.00m) and Small (1.05m×0.56m). When the two screen sizes were combined with the two designed viewing distances, four fields of views (FOV) were consequently formed which were 15° (Small, Far), 27° (Small, Near; Large, Far) and 46° (Large, Near).

Table C.1 lists the constructed maximum disparity for each scene content at different camera baselines and screen sizes. It can be seen from the table that scene contents in the configurations of (10cm, Large), (10cm, Small) and (20cm, Small) were mostly located in the range of maximum display disparity of 35 pixels. This indicates that these stimuli can be viewed without any discomfort at a far viewing distance. Other configurations were beyond 35 pixels but to different extents of uncomfortable viewing. In near viewing distance, scene content in the configuration of (10cm, Small) is the only one mostly located in its comfort zone (20 pixels).

Table C.1: Maximum disparity of different scene contents at three camera baselines and two screen sizes

Content	Large			Small		
	10cm	**20cm**	**30cm**	**10cm**	**20cm**	**30cm**
Construction	16	36	54	9	20	31
Moped	29	65	59	17	37	55
Sculpture	19	43	64	11	24	36
Closeup	20	46	72	11	26	40
Bikes	26	50	77	14	28	44
Hallway	50	103	156	28	58	88
Tables	23	49	78	13	28	44
Sofa	23	45	69	13	25	39
Notebook	36	77	131	20	43	74
Feet	35	72	134	20	40	75

C.2.3 Test Methodology

C.2.3.1 Single Stimulus

Among different methodologies for subjective quality assessment of SDTV pictures in ITU-R BT. 500 [14], three widely used methods are Double Stimulus Continuous Quality Scale (DSCQS), Double Stimulus Impairment Scale (DSIS), and Single Stimulus (SS). In this study, as several camera baselines for each scene have been considered, the original reference 3D images were not available. Therefore, the SS method was used, which has also been widely adopted in assessing stereoscopic image quality levels with varying camera baselines in [25, 29, 8].

C.2.3.2 Subjects

A total of 30 subjects (18 males and 12 females, aged from 23 to 47 years old) recruited from the Norwegian University of Science and Technology participated in the tests. Before the training sessions, a screen test was performed to collect visual characteristics of subjects, including pupillary distance (measured by a ruler), normal or corrected binocular vision (tested by the Snellen chart), color vision (tested by the Ishihara), and stereo vision (tested by the TV-04 and TV-07 in ITU-R BT. 1438 [13]). All the subjects had binocular visual acuity larger than 1.0, stereo vision smaller than 30 arcsecond and no colorblind subjects were detected. The pupillary distance of subjects varied from 62 to 74 millimeters.

C.2.3.3 Training Session

In the tests, all the subjects were divided into 15 groups, each of which consisted of two subjects. All of them participated in both training and test sessions. During the training sessions, two subjects seated at the Far Center and Far Side viewing

Table C.2: Explanations of five categorical adjectival levels and training examples of the Trees scene

Levels	**Explanation**	**Examples** (baseline, size)
Excellent	Positive≫Negative: Enjoy the experience very much.	(10cm, Large)
Good	Positive>Negative: Accept viewing such quality happily in daily life.	(20cm, Large)
Fair	Positive≈Negative: Accept viewing such quality but reluctantly.	(30cm, Small)
Poor	Positive<Negative: Refuse to accept viewing such quality.	(40cm, Small)
Bad	Positive≪Negative: Feel headache and do not want to view 3D again.	(50cm, Large)

positions. A 5-scale adjectival categorical measure with a presentation example from the Tree content at each quality level was used to benchmark and harmonize the measure scale between subjects, as shown in Table C.2. When an example image pair was presented, the test operator verbally explained the corresponding quality level to the subjects. The subjects were told to judge quality by comparing the positive 3D effect with the negative. The positive effect refers to the binocular depth, immersiveness and reality, while the negative effect can be caused by discomfort arising from 3D fusion limitations, visible binocular artifacts and geometric distortions. In addition, a detailed definition of each scale was provided to the subjects in form of written instructions (see Table C.2). The subjects were encouraged to view the representative examples as long as they wished and ask questions if they needed any further clarifications. An answering sheet marked with the corresponding quality level (diagonal line on the continual scale) was shown to the subjects when the example image pairs were displayed in order for them to become familiar with the same procedure to be performed in the real test sessions. The training sessions would continue until the subjects could understand and distinguish the five different quality levels.

C.2.3.4 Test Session

Since the number of test configurations (240) was too large for a single session, each test session was split into two sub-sessions. During each sub-session, 60 test images were randomly presented to two subjects (A and B) seated at respective positions. If starting from the positions of Far Center (subject A) and Far Side (subject B) in the first sub-session, they would switch to the positions Near Side (subject A) and Near Center (subject B), respectively, in the second sub-session. They could also start from Near positions and switch to Far positions subsequently. Additionally, at the beginning of each sub-session, the subjects were presented by 3

dummy versions derived from the Grass scene in an order of (50cm, Large), (10cm, Small) and (30cm, Large), to calibrate their quality judgment. The voting on the dummy images was excluded from the subsequent analysis. During the whole test period, the subjects were not allowed to ask questions anymore to guarantee that the test was not interrupted. Each sub-session lasted about 15 minutes, in which each image was shown for 10 seconds, plus another 5 seconds to enter the score on the answering sheet. To avoid visual fatigue, a 5 minutes break was inserted between two sub-sessions. In total 15 testing samples were obtained for each viewing position.

C.2.4 Observations on the Subjective Scores

The subjective quality scores on the stereoscopic 3D images were observed first by plotting figures, indicating their relationship with scene content, camera baseline, screen size and viewing position.

C.2.4.1 Normality Test and Outlier Detection

Quantitative quality of the test stimuli can be performed by averaging the voted quality scores across the participated subjects. Therefore, Mean Opinion Score (MOS) is often used in current quality assessment methodologies, in which the subjective scores are assumed to be subject to the normal distribution. In this work, a β_2 test [14] based on the kurtosis coefficient was employed for normality validation. We classified the β_2 test results into three categories: normal distribution ($2 \leq \beta_2 \leq 4$), close to normal distribution ($1 \leq \beta_2 < 2$ or $4 < \beta_2 \leq 5$) and non-normal distribution ($\beta_2 < 1$ or $\beta_2 > 5$). If the ratio of normal and close to normal distributions was more than 80% out of the total test stimuli, we can reasonably assume that the subjective scores follow a normal distribution. The results showed that the majority of stimuli (159 of 240) were normally distributed and (48 of 240) were close to normal, while the rest (33 of 240) were not. Therefore, the subjective scores across the subjects can be assumed to be reasonably subject to the normal distribution. In addition, subjects who produced votes significantly distant from the average scores should be removed from result analysis. Thus, an outlier detection was performed according to the guidelines described in [14], but no outliers were detected in our experiments.

C.2.4.2 MOS and Observations

The MOS values and 95% Confidence Interval (CI), computed from the raw quality scores, are plotted in Figure C.4 with respect to the camera baselines under different viewing conditions for the 10 scene contents. It can be seen that the CIs are usually very small indicating a high consistence of the quality scores across subjects. Generally speaking, the MOS values significantly vary across different scene contents and decrease when increasing the camera baseline. Therefore, both scene content and camera baseline may have strong impact on QoE individually.

Moreover, when increasing the camera baseline, the decreasing slope of MOS in Moped, Hallway, Tables and Notebook scenes is more significant than that in other scenes. This indicates that the differences on MOS among camera baselines are not same in each scene contents, which implies an interaction between camera baseline and scene content in QoE assessment. By taking the factor of screen size into account, it can be seen that the difference on MOS between Large and Small sizes with the scenes of Construction and Bikes is larger than that with Closeup and Tables. Thus, there might also be an interaction between screen size and scene content. For the contents of Moped, Bikes, Hallway, Sofa, Notebook and Feet, the difference on MOS between Large and Small sizes at different camera baselines is relatively large. Thereby, an interaction might exist between screen size and camera baseline or even scene content. However, no clear difference on MOS has been observed between the viewing distances of Near and Far, Center and Side, since the MOSs of different viewing positions stick to each other in all scene contents, camera baselines and screen sizes. Therefore, viewing positions can be considered to have no significant impact on QoE.

C.3 Exploring the Relationship between Requisite Factors and QoE

The indicated relationship between above requisite factors and stereoscopic QoE can be further confirmed by extracting significant factors from a statistical analysis. Moreover, a separate requisite factor, crosstalk level, will be explored in this section. Subsequently, the perceptual attributes of stereoscopic QoE can be employed to bridge the gap between low-level requisite factors and high-level users' viewing experience.

C.3.1 Requisite Factors

C.3.1.1 Significant Factors in Subjective Tests

In order to verify the above observations and evaluate the impact of the independent variables (scene content, camera baseline, screen size, viewing position) on the dependent variable (stereoscopic QoE), we have employed ANalysis Of VAriance (ANOVA) to analyze the subjective quality scores. ANOVA is a useful tool to test the equality hypothesis of means between two or more sample groups using p value.

We have used the Statistical Package for the Social Sciences (SPSS) statistics (version 17.0) for result analysis. Figure C.5 shows the ANOVA analysis results under an entire null hypothesis. In the figure, SC, CB and LS denote the within-subjects factors including scene content, camera baseline and screen size, respectively, NF and CS denote the between-subjects factors, including two groups of viewing positions, Near versus Far and Center versus Side. The symbol * indicates existence of interaction between two factors, and the horizontal red line equal to 0.95 represents the threshold of significant difference. It can be seen from Figure

C.5 that SC, CB, LS*SC, LS*CB, SC*CB and LS*SC*CB are beyond the significant difference threshold. Thus, these are significant factors for stereoscopic QoE. It means that scene content and camera baseline have significant impact on QoE individually, and there are 2-factors and 3-factors interactions between screen size, scene content and camera baseline in terms of stereoscopic QoE. However, other factors have no significant impact or interaction on stereoscopic QoE. The ANOVA analysis results are in accordance with the observations in the previous section.

C.3.1.2 Crosstalk Level

Nowadays, system-introduced crosstalk exists in almost all stereoscopic screen displays. Although crosstalk reduction and cancelation technologies are often adopted to eliminate the crosstalk artifact, neither of them can completely eliminate it. Therefore, crosstalk level of display is another requisite factor in current stereoscopic display systems. The crosstalk levels of various stereoscopic displays are different. As measured in Subsection C.2.1.3, the crosstalk level in our projected polarized display is approximately 3%. While in Hyundai S465D display, the crosstalk level is about 1.3% in the max white brightness. These are two polarized displays employed in Q2S-QoE and EPFL-QoE databases, respectively. In these polarized displays, the consistence of the system-introduced crosstalk has also been verified over the display, between projectors, and among different combinations of brightness between left and right test images. The system-introduced crosstalk can be simulated by the algorithm developed in [5], as summarized in the following equation:

$$\begin{cases} R_c = R_o + P \times L_o \\ L_c = L_o + P \times R_o \end{cases} \tag{C.4}$$

where L_o and R_o denote the original left and right views shown on the stereoscopic display, L_c and R_c are the perceived images influenced by the system-introduced crosstalk of stereoscopic display, and the parameter P is the crosstalk level.

C.3.2 Perceptual Attributes

Perceptual attributes are the sensorial results of HVS, thereby they can more accurately represent QoE perception rather than the requisite factors. It was pointed out in [25] that the overall QoE is a trade-off between perceived image distortion, perceived depth and visual strain. Particularly, crosstalk is one of the most annoying distortions degrading stereoscopic image perception in the visualization stage. In addition, perceived depth and visual strain condition each other in ways of strong perceived depth usually brings more visual strain while weak perceived depth often related to less visual strain. We named this attribute as depth enabled visual comfort and found that it can be represented by a derivation from comfort zone to some extent.

C.3.2.1 Crosstalk Perception

It can be assumed that crosstalk can always be perceived during the subjective tests (i.e., the Q2S and EPFL QoE tests) under two conditions. One is that both displays in Q2S-QoE and EPFL-QoE databases have system-introduced crosstalk levels, which were measured as 3% and 1.3%, respectively. The other is that the camera baseline of the test stimuli in Q2S-QoE and EPFL-QoE tests is adequately large and the maximum reaches 300mm and 600mm, respectively, because it is known from the subjective tests in [25, 29] that higher crosstalk levels are more visible at larger camera base distances. Thus, it is inevitable that the participants perceived the visible crosstalk when assessing the perceived quality in the Q2S-QoE and EPFL-QoE tests. Moreover, the perceived visual stimuli can be simulated by equation (C.4).

As implied in [35], crosstalk perception further relies on its three perceptual attributes, namely shadow degree, separation distance, and spatial position of crosstalk. In particular, shadow degree and separation distance of crosstalk are the distinctness and distance of crosstalk against from the original view, respectively. They are 2D perceptual attributes which can be perceived via single eye view. While spatial position of crosstalk is 3D perceptual attribute, which is the impact of crosstalk position in 3D space on perception when the left and right views are fused and 3D perception is generated. Furthermore, these 2D perceptual attributes interact mutually and the 3D perceptual attribute only impacts the visible crosstalk satisfying requirements of 2D perceptual attributes.

C.3.2.2 Depth Enabled Visual Comfort

A visual stimulus located in the comfort zone of a stereoscopic display indicates that it is below the perceptual-physiological limitations of HVS and cannot cause uncomfortable viewing experience. However, visual comfort does not always exist for stimuli with different disparity levels due to the limited comfort zone of current stereoscopic displays. Therefore, we believe that the stimuli located in different parts in the comfort zone will raise different viewing experiences.

As illustrated in Figure C.2, the points of a specified image pairs (stimulus) on particular display can be categorized into two groups: inside (green points) and outside (red points) of the comfort zone of display. The inside points can be viewed without discomfort, while the outside points can be introduced with visual discomfort. It indicates that the visual comfort should be related to the ratio of the points located in the comfort zone of the display system. However, the ratio is not a single indicator. If all the points of a stimulus are located in the comfort zone, viewers usually prefer those points closer to the maximum disparity, since they can provide stronger depth perception while not causing any discomfort. This implies that visual comfort can also be dependent on disparity amplitude of those points inside the comfort zone. Therefore, we believe that the depth enabled visual comfort of stereoscopic QoE should reflect both the ratio of points located in the comfort zone of the display system and their disparity amplitude.

C.3.3 Analysis of Relationship between Requisite Factors and Perceptual Attributes

In our previous work [35], we have found out that crosstalk level, camera baseline and scene content have significant impacts on crosstalk perception, and also they have 2-factors interactions between each other in terms of the impact on crosstalk perception. Specifically, shadow degree of crosstalk is related to crosstalk level and contrast of scene content and their interaction, while separation distance of crosstalk is related to camera baseline and depth of scene content and their interaction. Thereby, the interaction between 2D perceptual attributes depends on interaction between camera baseline and crosstalk level, interaction between contrast and depth of scene content, and the 3-factors interaction among camera baseline, crosstalk level and scene content. Moreover, spatial position of crosstalk has a relationship to the depth of scene content in the visible crosstalk region.

When depth enabled visual comfort is considered, it is a mutual result between test stimuli and stereoscopic display. Relevant factors of test stimuli include depth of scene content, camera baseline and image resolution, since they can determine the disparity in pixel unit. On the other hand, the comfort zone of a stereoscopic display is usually determined by the viewing distance, pupillary distance, depth of focus (DOF) and pixel width of the display, as shown in equation (C.2). However, only pixel width of screen and viewing position are variable, since pupillary distance and DOF of an individual viewer are usually assumed to be constant. In addition, as viewing distance is found to have no significant effect to QoE in Subsection C.3.1.1, we have set the viewing distance to be 3 times of the display height in our experiments, which is also conform to standard suggestions. Moreover, the QoE significant factor screen size is exhibited in the screen resolution and pixel width, and it is necessary to make the image resolution of test stimuli same to the screen resolution.

The aforementioned relationship between perceptual attributes of QoE and requisite factors is summarized in Table C.3. It can be seen that the relationship between a single perceptual attribute and its factors is relatively clear. However, it is still unclear how these perceptual attributes contributed to the stereoscopic QoE, which will be explored in the next section by proposing objective metric and making use of the subjective scores.

C.4 Towards a Stereoscopic QoE Metric

An objective quality metric for stereoscopic QoE in the simplest stereoscopic system can be developed by modeling the aforementioned perceptual attributes. This section explains the derivation of the metric and the validation of the metric with respect to Q2S-QoE and EPFL-QoE databases, respectively.

Table C.3: Relationship between QoE perceptual attributes and requisite factors: crosstalk level (CL), camera baseline (CB), size of screen(S_S), resolution of screen (S_R), pixel width of screen (S_PW), viewing position (VP), contrast of scene content (SC_C), depth of scene content (SC_D), both contrast and depth of scene content (SC_CD)

Perceptual attributes		**Related factors**
Crosstalk perception	Shadow degree	CL, SC_C, CL*SC_C
	Separation distance	CB, SC_D, CL*SC_D
	Shadow degree * Separation distance	CL*CB, SC_CD, CL*CB*SC_CD
	Spatial position	SC_D in visible crosstalk region
Depth enabled visual comfort	Disparity of test stimuli	CB, SC_D, S_R, and their 2-factors and 3-factors interaction
	Comfort zone of stereoscopic display	S_PW(S_S/S_R), VP

C.4.1 Crosstalk Perception

We have modeled the crosstalk perception by combining two maps, namely, a structural similarity (SSIM) map and a filtered depth map, representing the 2D and 3D perceptual attributes of crosstalk, respectively [35]. In this work a filtered disparity map with relative depth is employed to replace the filtered depth map with absolute depth in [35] in order to represent spatial position more appropriately. Although these two maps often exhibit similar performance of crosstalk perception, the advantage of using disparity map is that camera intrinsic and extrinsic parameters which are often not available in stereo image pairs are not required when estimating the disparity map. Therefore, the proposed metric in this work can be applied in practical systems.

C.4.1.1 SSIM Map

The SSIM quality measure proposed by Z. Wang et al. [26] assumes that the measurement of structural information provides a good estimation of the perceived image quality. The structural similarity measure is constructed based on comparison of three components: luminance, contrast, and structure, between an original image in perfect quality and its distorted version. In our case, the original image is the one showed on the stereoscopic display without any crosstalk L_o, and the distorted version is the one perceived by the viewer with both system-introduced and simulated crosstalk L_c. Thus, the SSIM index map can be defined as follows,

$$L_s = SSIM(L_o, L_c) \tag{C.5}$$

where SSIM denotes the SSIM model in [26] and L_s is the generated SSIM index map of the left eye view.

C.4.1.2 Disparity Map

Disparity defines the difference captured by two cameras instead of eyes in computer stereovision, and a disparity map can be estimated based on two-frame stereo correspondence algorithms. In this work, we adopted a method using the Sum of Squared Difference plus Min Filter (SSDMF) [23] to estimate the disparity information.

$$R_{dis} = SSDMF(R_o, L_o) \tag{C.6}$$

where R_{dis} denotes the disparity map computed by SSDMF. The disparity map is a gray image with black pixels denoting the smallest disparity 0 whilst white pixels being the largest 255.

Subsequently, a filtered disparity map is defined based on the SSIM index map, as following in equation (C.7):

$$R_{pdis}(i,j) = \begin{cases} R_{dis}(i,j) & if \quad L_s(i,j) < 0.977 \\ 0 & if \quad L_s(i,j) \geq 0.977 \end{cases} \tag{C.7}$$

where i and j are the pixel indices, R_{pdis} denotes the filtered disparity map corresponding to the visible crosstalk region of left eye image.

C.4.1.3 Metric for Crosstalk Perception

Based on the above analysis, crosstalk perception can be considered as an integration of the SSIM index map and the filtered disparity map. Consequently, a crosstalk perception index is defined, as following in equation (C.8):

$$V_{cp} = AVG(L_s \times (1 - R_{pdis}/255)) \tag{C.8}$$

where AVG and array multiply $.*$ denote the average operation and the element-by-element multiplication, respectively. V_{cp} is the final predicted value of the objective metric for crosstalk perception. As can be seen in the equation, the filtered disparity map R_{pdis} is normalized into the interval [0, 1] first by dividing the maximum depth value 255, and then by subtracting it from 1 in order to be consistent with the physical meaning of SSIM index map that a lower pixel value in SSIM index map indicates a larger crosstalk distortion.

C.4.2 Metric for Depth Enabled Visual Comfort

As analyzed in Subsection C.3.2.2, both the ratio of points located in the comfort zone of a display system and the disparity amplitude of those points should be modeled in presenting depth enabled visual comfort perceptual attribute. Specifically, the maximum disparity of comfort zone can be calculated using equations (C.2) and (C.3) and the disparity amplitudes of points can be represented by the

disparity levels of the corresponding pixel pairs in stimuli, namely, disparity map. Subsequently, the metric is defined in the following equation:

$$V_{devc} = \frac{\sum R_{dis}(i,j)}{D \times Res}, \quad if \quad R_{dis}(i,j) \leq D \tag{C.9}$$

where i and j are pixel indices, Res is the resolution of the stimulus, D denotes the maximum disparity of comfort zone, and $\sum$ denotes the sum operation. V_{devc} can express both the ratio and disparity amplitude of those points in the comfort zone by summing up the disparity amplitude of the points in the comfort zone, and then divided it by the resolution of the stimulus and the maximum disparity of the comfort zone. In particular, the maximum disparity D in the denominator is used for normalization and its values for the Hunydai display and our rear projected system are 42.3, 26.8 and 15.0 pixels, respectively, as shown in Table C.4. Moreover, V_{devc} reaches its maximum 1 when all the pixels of a stimulus have a same disparity equal to D, indicating maximal perceived depth and visual comfort. On the other hand, it reaches the minimum 0 while the stimuli has zero disparity or only disparity outside the comfort zone, where the former indicates no perceived depth and the latter no visual comfort.

C.4.3 Objective Metric for Stereoscopic QoE

The overall stereoscopic QoE is an integrated result of crosstalk perception and depth enabled visual comfort. Thus, the objective metric for predicting the overall QoE is developed by combining them linearly, as explained in equation (C.10):

$$V_{QoE} = (1 - \alpha) \times V_{cp} + \alpha \times V_{devc} \tag{C.10}$$

where α is a weight in the interval of [0, 1] to balance the crosstalk perception and depth enabled visual comfort in the overall stereoscopic QoE VQoE. In particular, an optimal value of α can be calculated by cross validation, as explained in the next subsection. For a certain α, when the disparity of a stimulus increases, the crosstalk perception becomes more visible whilst the V_{cp} decreases towards 0. However, V_{devc} may first increase in prior to that most of the disparity reaches D and then decrease afterwards.

C.4.4 Experimental Results

The performance of an objective quality metric can be evaluated by a comparison with respect to the MOS values obtained in subjective quality tests. The proposed metric has been validated against both the Q2S-QoE and EPFL-QoE databases. The requisite factors of the two databases vary in different levels increasing the complexity and robustness of the database, as listed in Table C.4, and thus can provide a better verification of the proposed metric.

As the metric employs the weight α, a cross validation was adopted to demonstrate the performance of the proposed metrics. A 3 runs of 3-fold cross validation

Table C.4: Overview of Q2S and EPFL QoE databases with most of the requisite factors in the simplest stereoscopic system

Database	**EPFL-QoE[8]**	**Q2S-QoE[33]**
Test cases	$54 = 9SC \times 6CB \times 1SS \times 1VP$	$240 = 10SC \times 3CB \times 2SS \times 4VP$
Scene content (SC)	Captureed(JPEG)	Captureed(JPEG)
Camera baseline (CB)	100-600mm	100-300mm
Polarized display	Hyundai S465D	Rear projected system (Figure C.1)
Crosstalk level	1.3%	3%
Screen size (SS)	$1.01m \times 0.57m$	$1.86m \times 1.00m$; $1.05m \times 0.56m$
Screen Resolution	1920×1080	1280×960; 720×405
Pixel width	0.53mm	1.46mm
Viewing position (VP)	2m	3.9m, 2.2m, Center and Side
Assumed viewing position	1.71m	1.68m, 1.00m
Maximal disparity	42.3 pixel	26.8 pixel;15.0 pixel

was applied to an integrated database combining the Q2S-QoE (240 test cases) and EPFL-QoE (54 test cases) databases. Specifically, in a K-fold cross validation, the samples in the integrated database are randomly partitioned into K subsamples. An arbitrary single subsample has been extracted from the K subsamples and used as the test set for evaluating the metric, while the rest K-1 subsamples have been used for training the weight. The cross validation process is then repeated K times (the folds), in which each of the K subsamples is used exactly once for the test. K results from the folds are then averaged to produce a single estimation. An N runs of K-fold cross validation involves running K-fold cross validation N times and the average of the N results is taken as the final estimation result.

The proposed metric was compared with two commonly used 2D image quality metrics V_{psnr} and V_{ssim}, as well as other stereoscopic QoE metrics V_{dis}, V_{sdis}, V_{sdiss} and V_{cdis} proposed in our previous work [30, 31, 34] using the following equations.

$$V_{psnr} = PSNR(L_o, L_c) \tag{C.11}$$

$$V_{ssim} = AVG(L_s) \tag{C.12}$$

$$V_{dis} = AVG(1 - R_{dis}/255) \tag{C.13}$$

$$V_{sdis} = AVG(L_s \times (1 - R_{dis}/255)) \tag{C.14}$$

$$V_{sdiss} = (1-\beta) \times AVG(L_s \times (1 - R_{dis}/255)) + \beta \times S \tag{C.15}$$

where V_{psnr} and V_{ssim} are the 2D metrics calculated between the original and crosstalk added left image, instead of original left and right images. V_{dis} is the average disparity of the stimuli which represents the scene content and camera baseline without taking into account the properties of the display. V_{sdis} made use of the SSIM map L_s to integrate the system-introduced crosstalk, and V_{sdiss} further takes the significant factor screen size into consideration with a weighted combination. The values of S have been set to 1.86, 1.05 and 1.01 corresponding to different screen sizes in the test databases, respectively.

To evaluate these metrics, Root Mean Squared Error (RMSE), Pearson correlation coefficient, and Spearman rank-order correlation coefficient have been selected as the criteria. They are calculated between the objective values MOS_p after a nonlinear regression using equation (C.16), suggested by the Video Quality Experts Group (VQEG), against the subjective scores MOS.

$$MOS_p = b_1/(1 + exp(-b_2 \times (r(V) - b_3))) \tag{C.16}$$

where b_1, b_2 and b_3 are the regression coefficients, $r(V)$ is the predicted quality value calculated by a metric V, and exp is the exponential function. The main purpose of the nonlinear regression is to unify $r(V)$ to the range of the MOS values.

Table C.5 lists the evaluation results of all the objective metrics in respect to the integrated database. When the parameter α or β has been used, a 3 runs of 3-fold cross validation was applied. In particular, the parameters were optimized in the training set and the evaluation results were calculated in the test set. Corresponding to the results in Table C.5, α was optimized to be 1 for 7 times, 0.87 for 1 time and 0.84 for 1 time, and β was optimized to 0.05 for 5 times and 0.04 for 4 times.

According to the results, the proposed metric V_{QoE} and two other metrics V_{devc} and V_{sdiss} exhibit similar performance and are significantly better than V_{psnr}, V_{ssim} and V_{cp} while slightly better than V_{dis} and V_{sdis}. This shows that 2D metrics (V_{psnr}, V_{ssim}) without considering 3D characteristics and the metric V_{cp} based on crosstalk perception alone cannot provide a good prediction of stereoscopic QoE. Moreover,

Table C.5: Evaluation results of different metrics

Metrics	**RMSE**	**Pearson**	**Spearman**
V_{QoE}	0.463	**0.854**	0.801
V_{cp}	0.778	0.528	0.407
V_{devc}	0.469	**0.852**	0.800
V_{psnr}	0.860	0.289	0.324
V_{ssim}	0.821	0.419	0.344
V_{dis}	0.513	0.820	0.744
V_{sdis}	0.510	0.821	0.747
V_{sdiss}	0.456	**0.859**	0.805

it can be seen that V_{dis} and V_{sdis} without taking screen size into consideration have a slightly worse performance than V_{sdiss}, which indicates that the compensation of screen size is necessary but its effect is limited. This phenomenon can also be observed from the value (0.05 or 0.04) of parameter β. Moreover, V_{dis} and V_{sdis} have similar performance indicating that combining SSIM map L_s with disparity map R_{dis} cannot significantly improve their performance. This might be due to the low crosstalk level of the used displayed systems and the small difference between them. Such a guess was verified again by V_{QoE}, where V_{cp} and V_{devc} have been combined while V_{cp} has been assigned almost no weight. This phenomenon implies that crosstalk perception is not quite perceptible in these two databases, while it does not suggest that crosstalk perception has no contributions to stereoscopic QoE. When system-introduced crosstalk achieves certain level, crosstalk perception often plays a significant role in the overall QoE perception. Furthermore, V_{QoE} is more tenable and convincing when compared to V_{sdiss}, although they have similar performance. Since V_{QoE} is constructed by the perceptual attributes of QoE, while V_{sdiss} is composed by some significant factors, thus V_{QoE} is more understandable from perceptual view point. Therefore, we have chosen V_{QoE} as the best metric for stereoscopic QoE prediction instead of V_{devc} and V_{sdiss}.

C.5 Conclusions

In this study, we have investigated the influence of four requisite factors on stereoscopic QoE assessment, including scene content, camera baseline, screen size and viewing position, by conducting a series of subjective quality tests. Typical test configurations have been designed to cover a full range of stereoscopic QoE changes in terms of disparity coverage between maximum disparity supported by displays and constructed disparity by test stimuli. The observation from the MOS result plots and the ANOVA statistical results demonstrated that scene content, camera baseline, as well as the interactions between screen size, scene content and camera baseline, have significant impacts on stereoscopic QoE, while other factors, especially viewing position related, have virtually no significant impact. Furthermore, the system-introduced crosstalk level has been identified to be a significant factor of crosstalk perception, which together with depth enabled visual comfort are two perceptual attributes of stereoscopic QoE. Based on the understanding of these perceptual attributes, an objective metric for predicting stereoscopic QoE has been proposed and validated against two stereoscopic quality databases by a cross validation, showing a high correlation (over 85%) with the subjective quality evaluation. However, the robustness of the proposed metric to other stereoscopic displays with different system-introduced crosstalk levels and screen sizes need to be further validated. Moreover, this study has been limited in the simplest stereoscopic imaging system, potential extensions to include all stages in the processing chain of 3D signals and the associated artefacts, e.g. compression artefacts, will also be investigated in future work.

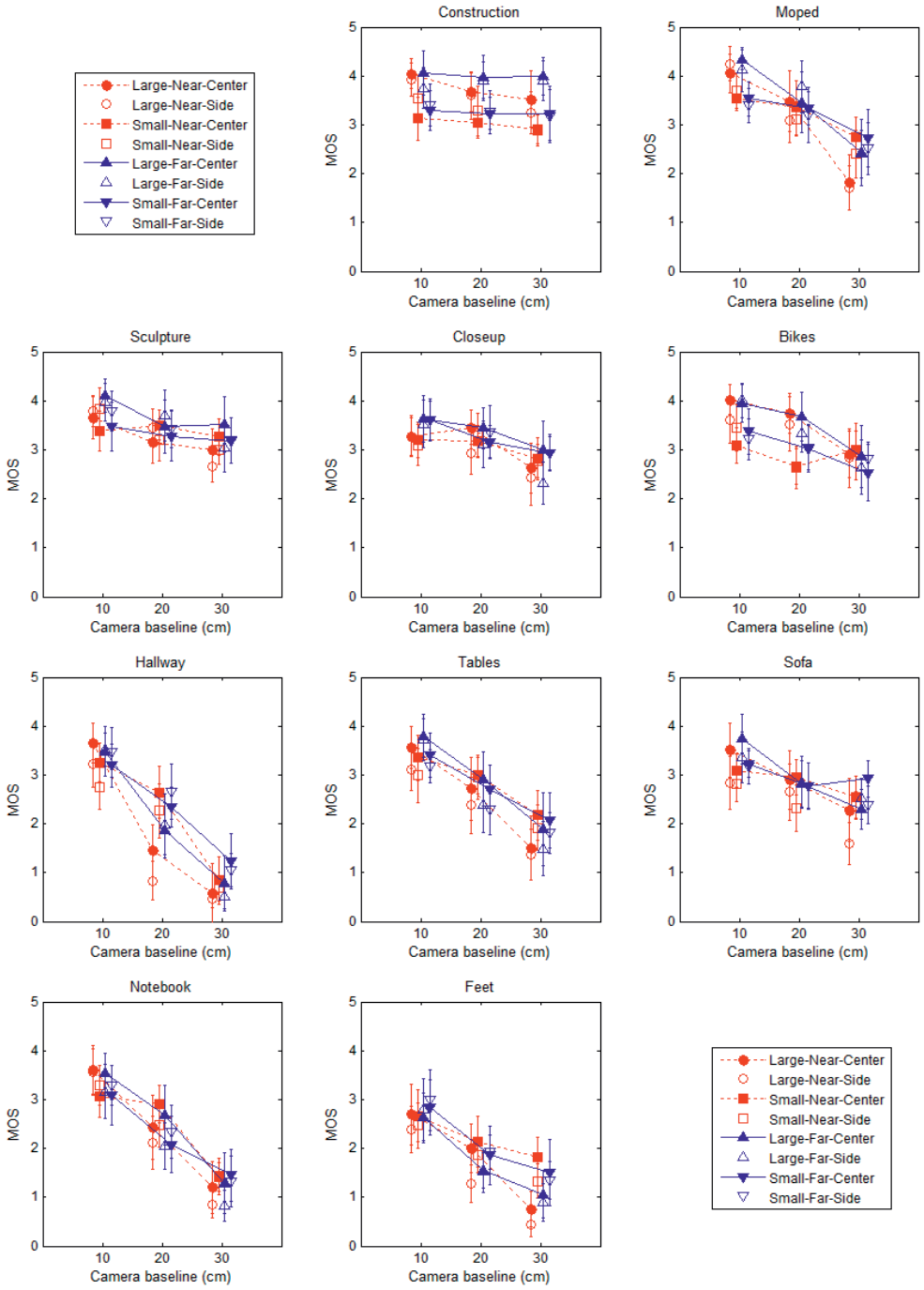


Figure C.4: MOS and CI (significance of 95%) of subjective scores for all test images.

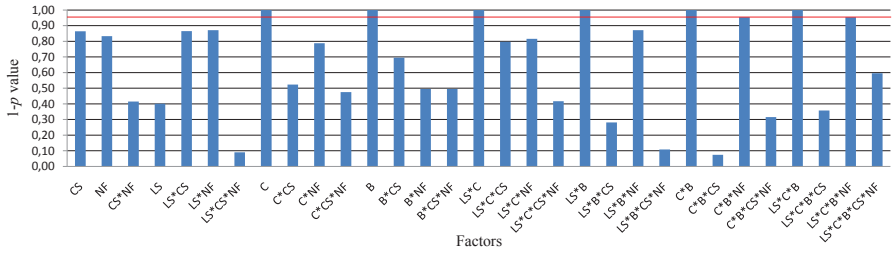


Figure C.5: Significant difference histogram of a 5-way ANOVA.

References

[1] 3D Image Quality Assessment. http://mmspg.epfl.ch/3diqa. 2010.

[2] 3D Video Quality Assessment. http://mmspg.epfl.ch/3dvqa. 2010.

[3] P. Aflaki, M. Hannuksela, J. Hakkinen, P. Lindroos, and M. Gabbouj. Subjective study on compressed asymmetric stereoscopic video. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 4021–4024, 2010.

[4] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau. Quality assessment of stereoscopic images. *EURASIP Journal on Image and Video Processing*, 2008(1):659024, 2008.

[5] A. Boev, D. Hollosi, and A. Gotchev. Software for simulation of artefacts and database of impaired videos. In *Mobile 3DTV Project report, No. 216503.*, [available] http://mobile3dtv.eu.

[6] A. Boev, D. Hollosi, A. Gotchev, and K. Egiazarian. Classification and simulation of stereoscopic artifacts in mobile 3DTV content. pages 72371F–72371F–12, 2009.

[7] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet. New requirements of subjective video quality assessment methodologies for 3DTV. In *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2010.

[8] L. Goldmann, F. De Simone, and T. Ebrahimi. A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. In *Three-Dimensional Image Processing (3DIP) and Applications*, volume 7526, 2010. [available] http://mmspg.epfl.ch/3diqa, http://mmspg.epfl.ch/3dvqa.

[9] P. Gorley and N. Holliman. Stereoscopic image quality metrics and compression. In *Stereoscopic Displays and Applications XIX*, volume 6803, pages 680305–680305–12, 2008.

[10] J. Häkkinen, T. Kawai, J. Takatalo, T. Leisti, J. Radun, A. Hirsaho, and G. Nyman. Measuring stereoscopic image quality experience with interpretation based quality methodology. In *Image Quality and System Performance V*, volume 6808, pages 68081B–68081B–12, 2008.

[11] W. IJsselsteijn, H. de Ridder, and J. Vliegen. Subjective evaluation of stereoscopic images: effects of camera parameters and display duration. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(2):225–233, 2000.

[12] W. IJsselsteijn, H. d. Ridder, J. Freeman, S. E. Avons, and D. Bouwhuis. Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence. *Presence: Teleoper. Virtual Environ.*, 10(3):298–311, 2001.

[13] ITU-R BT.1438. Subjective assessment of stereoscopic television pictures. 2000.

[14] ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television pictures. 2002.

[15] L. Kaufman, J. Kaufman, R. Noble, S. Edlund, S. Bai, and T. King. Perceptual distance and the constancy of size and stereoscopic depth. *Spatial Vision*, 19(5):439–457, 2006.

[16] D. Kim, D. Min, J. Oh, S. Jeon, and K. Sohn. Depth map quality metric for three-dimensional video. In *Stereoscopic Displays and Applications XX*, volume 7237, pages 723719–723719–9, 2009.

[17] S. Kishi, S. H. Kim, T. Shibata, T. Kawai, J. Häkkinen, J. Takatalo, and G. Nyman. Scalable 3D image conversion and ergonomic evaluation. In *Stereoscopic Displays and Applications XIX*, volume 6803, pages 68030F–68030F–9, 2008.

[18] K. Klimaszewski, K. Wegner, and M. Domanski. Distortions of synthesized views caused by compression of views and depth maps. In *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009*, pages 1–4, 2009.

[19] B. Mendiburu. *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*. Focal Press/Elsevier, 2009.

[20] R. Olsson and M. Sjostrom. A depth dependent quality metric for evaluation of coded integral imaging based 3D-images. In *3DTV Conference, 2007*, pages 1–4, 2007.

[21] S. Patel, H. Bedell, D. Tsang, and M. Ukwade. Relationship between threshold and suprathreshold perception of position and stereoscopic depth. In *J. Opt. Soc. Am. A Opt. Image Sci. Vis.*, volume 26, pages 847–861, 2009.

[22] Z. Sazzad, S. Yamanaka, Y. Kawayokeita, and Y. Horita. Stereoscopic image quality prediction. In *Quality of Multimedia Experience (QoMEX), 2009 International Workshop on*, pages 180–185, 2009.

[23] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47(1-3):7–42, 2002.

[24] P. J. H. Seuntiens, L. M. J. Meesters, and W. A. IJsselsteijn. Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation. *TAP*, 3(2):95–109, 2006.

[25] P. Seuntiëns, L. Meesters, and W. IJsselsteijn. Perceptual attributes of crosstalk in 3D images. *Displays*, 26(4–5):177–183, 2005.

[26] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[27] A. Woods. Understanding crosstalk in stereoscopic displays. In *Three-Dimensional Systems and Applications*, 2010.

[28] A. Woods, T. Docherty, and R. Koch. Image distortions in stereoscopic video systems. In *Stereoscopic Displays and Applications IV*, volume 1915, pages 36–48, 1993.

[29] L. Xing, T. Ebrahimi, and A. Perkis. Subjective evaluation of stereoscopic crosstalk perception. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 77441V–77441V–9, 2010.

[30] L. Xing, J. You, T. Ebrahimi, and A. Perkis. In *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*.

[31] L. Xing, J. You, T. Ebrahimi, and A. Perkis. Estimating quality of experience on stereoscopic images. In *Intelligent Signal Processing and Communication Systems (ISPACS), 2010 International Symposium on*, pages 1–4, 2010.

[32] L. Xing, J. You, T. Ebrahimi, and A. Perkis. A perceptual quality metric for stereoscopic crosstalk perception. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 4033–4036, 2010.

[33] L. Xing, J. You, T. Ebrahimi, and A. Perkis. Factors impacting quality of experience in stereoscopic images. volume 786304, pages 786304–786304–8, 2011.

[34] L. Xing, J. You, T. Ebrahimi, and A. Perkis. Objective metrics for quality of experience in stereoscopic images. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 3105–3108, 2011.

[35] L. Xing, J. You, T. Ebrahimi, and A. Perkis. Assessment of stereoscopic crosstalk perception. *Multimedia, IEEE Transactions on*, 14(2):326–337, 2012.

[36] J. You, G. Jiang, L. Xing, and A. Perkis. Quality of visual experience for 3D presentation: stereoscopic image. In *High-Quality Visual Experience: Creation, Processing and Interactivity of High-resolution and High-dimensional Video Signals*, pages 51–77. Springer-Verlag, 2010.

D Factors Impacting Quality of Experience in Stereoscopic Images

Author

Liyuan Xing
Junyong You
Touradj Ebrahimi
Andrew Perkis

Conference

Stereoscopic Displays and Applications XXII (SDA), 2011.

Abstract

The stereoscopic 3D industry has fallen short of achieving acceptable quality of experience (QoE) because of various technical limitations, such as excessive disparity, accommodation-convergence mismatch. This study investigates the effect of scene content, camera baseline, screen size and viewing location on stereoscopic QoE in a holistic approach. 240 typical test configurations are taken into account, in which a wide range of disparity constructed from the shooting conditions (scene content, camera baseline, sensor resolution/screen size) was selected from datasets, making the constructed disparities locate in different ranges of maximal disparity supported by viewing environment (viewing location). Second, an extensive subjective test is conducted using a single stimulus methodology, in which 15 samples at each viewing location were obtained. Finally, a statistical analysis is performed and the results reveal that scene content, camera baseline, as well as the interactions between screen size, scene content and camera baseline, have significant impact on QoE in stereoscopic images, while other factors, especially viewing location involved, have almost no significant impact. The generated mean opinion scores (MOS) and the statistical results can be used to design stereoscopic quality metrics and validate their performance.

D.1 Introduction

The introduction of three-dimensional television (3DTV) to the home environment is approaching and has been compared to the transition from black-and-white to color TV. The most important incentive for deploying 3DTV is that it can offer richer immersive and realistic experience by introducing binocular depth. Nowadays, 3DTV has adopted stereoscopic 3D imaging technology which is believed to be the most mature and cost effective technique, when compared to integral imaging and holography. Specifically, stereoscopic imaging is based on capturing a pair of two-dimensional (2D) images, and then separately delivering them to respective eyes. Consequently, 3D perception is generated in the human visual system (HVS). However, due to excessive disparity and accommodation-convergence mismatch, stereoscopic imaging technology still fails to provide satisfying viewing experience with all different cases. Therefore, kinds of factors and their impact on quality of experience (QoE) of stereoscopic imaging need to be thoroughly investigated.

It is already known that stereoscopic 3D experience is influenced by a complex combination of different factors. Typically, these factors include the characteristics of 3D content, the configurations of capture and display systems, selections of image processing mechanism, and the perceptual-physiological limitations of HVS. In [6], a comprehensive summery of factors affecting stereoscopic QoE are listed.

Usually subjective tests are conducted to understand how part of aforementioned influence factors impact the overall QoE. Some of these tests [17, 19, 12, 13, 16, 3] focused on how the distortions induced by image processing technology affect QoE. Compression related factors [13, 16, 3] are widely studied. Some efforts have also been devoted to studying the influence of different capture and display configurations on QoE. Factors, such as camera baseline [17, 16, 7, 5], screen size [8] and viewing location [15, 11], are studied independently. While these works deepen our knowledge, a comprehensive understanding on requisite factors for stereoscopic imaging is still missing. Here requisite factors refer to the factors in the simplest stereoscopic system containing capture and display systems only. Although the authors of [18] demonstrated how a rectilinear grid in front of the camera system has been distorted upon display under various settings of most requisite factors, further influence of requisite factors on the perceived quality for human subjects needs to be evaluated quantitatively.

In our study, we use a holistic approach to investigate the influence of some requisite factors, such as scene content, camera baseline, screen size and viewing location on stereo QoE in variety of conditions. Particularly, both the characteristics of test stimuli and limitations of HVS were taken into consideration explicitly when designing the test configurations. Thus, the test configurations were prepared in order to cover a full range of QoE for a certain display. Then extensive subjective tests were conducted and statistical analysis was carried out to study the influence of requisite factors to QoE.

The remainder of this paper is organized as follows. In Section D.2, we present the test conditions. The test methodology is described in Section D.3. The subjective

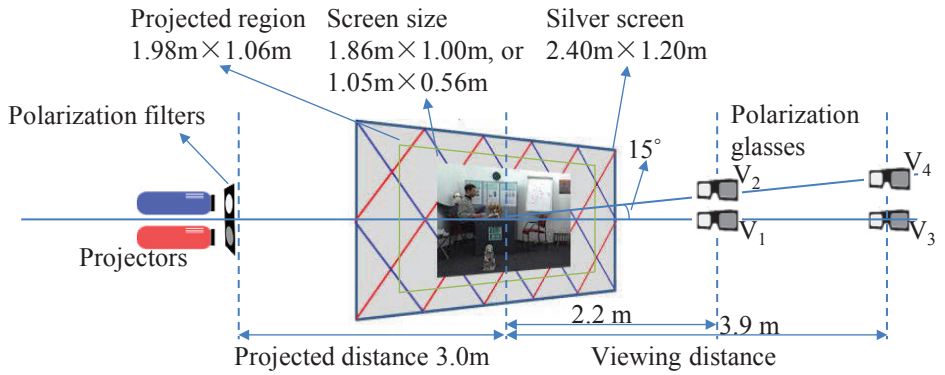


Figure D.1: The display system for subjective tests.

results are analyzed statistically in Section D.4. Finally, concluding remarks are given in Section D.5.

D.2 Test Conditions

Since there is a strong interaction between scene content, camera baseline, screen size and viewing location on QoE, a variety of situations covering a full range of QoE for certain display have been designed. In the following, we will provide the details of the display system that was used to perform the subjective tests and show how the test configurations were designed.

D.2.1 Display System

Polarization technique was used to present 3D images, as illustrated in Figure D.1. Specifically, two Canon XEED SX50 projectors with resolutions of 1360×768 were placed on a Chief ASE-2000 Adjusta-Set Slide Stacker. The stacker can be adjusted with $+/-7°$ swivel, $+/-20°$ tilt and $+/-7°$ leveling range. Two Cavision linear polarizing glass filters with sizes of 4×4" were installed orthogonally in front of the projectors. In this way, two views were projected and superimposed onto the backside of a 2.40m×1.20m silver screen. The projected distance between the projectors and the silver screen was about 3 meters, forming a projected region occupying the central area of the silver screen with the width of 1.98m and height of 1.06m. Images with different resolutions were displayed in an actual size mode in order to simulate different screen sizes. The subjects were asked to view the 3D images with polarized glasses on the opposite side of the silver screen at predefined locations.

Prior to the subjective tests, the display system was adjusted in order to minimize the equipment-introduced crosstalk. In particular, the positions of projectors were adjusted to guarantee that the center points of projectors, projected region and

silver screen, located in the same horizontal line (center horizontal line as shown in Figure D.1) and the line was perpendicular to the silver screen. Moreover, the angles of stackers and the keystones of the projectors were adjusted with the help of projected Hermann grid images. The adjustment of display system was finished when expert viewers could not observe any crosstalk on the projected Hermann grid image.

During the subjective tests, all the doors and windows of the test room were closed and covered by black curtains. In addition, the lights in the room were turned off except for one reading lamp to lighten the answering sheet when entering subjective scores.

D.2.2 Test Design

Binocular disparity is especially invoked in stereoscopic displays and presented to viewers for inducing the perception of stereoscopic depth. However, current stereoscopic displays have convergence-accommodation mismatch that the screen is always accommodated regardless where the convergence point is located. This decorrelation is not a natural function of the HVS and can further cause viewers' discomfort. Therefore, there is a limited comfortable zone for stereoscopic display [14]. Moreover, the authors of [4] pointed out that such comfortable zone varies across stereoscopic displays. They employed equations to calculate the maximal disparity for comfortable viewing of certain display. We believe that stimuli located in different range of comfortable zone can provide different viewing experiences. Therefore, based on the calculated maximal disparity of the display system in our subjective quality experiments, kinds of stimuli with different disparity levels were generated to cover a full range of QoE. 240 test configurations in total (10 scene contents×3 camera baselines×2 screen sizes×4 viewing locations) were designed. The details are given as following. First, the maximal disparity of our display system is calculated using the following equations as in [4] and viewing locations are further selected according to the maximal disparity.

$$Z_f = d - \frac{1}{1/d + dof}, Z_b = \begin{cases} \frac{1}{1/d - dof} - d, if d < 5 \\ \infty, if d \geq 5 \end{cases} \tag{D.1}$$

$$D_f = \frac{Z_f \times e}{(d - Z_f) \times p_w}, D_b = \frac{Z_b \times e}{(Z_b + d) \times p_w} \tag{D.2}$$

where d, e, p_w and dof are the viewing distance, pupillary distance, pixel width of the display and depth of focus (DOF), respectively, Z_f and Z_b denote the foreground and background distances of the comfort viewing zone in meter, while D_f and D_b are the corresponding maximal foreground and background disparity in terms of pixel. When d is smaller than 5, D_f equals to D_b. Especially, e and dof are usually assumed to be 65mm and 0.2diopter, respectively. Moreover, p_w is also fixed in our display system, which is 1.46mm (1.98m/1360). Thus, the viewing distance is the only variable decides the maximal disparity.

We designed two viewing distances 2.2m and 3.9m which resulted in the maximal disparity to be 20 and 35 pixels, respectively. In addition, an ordinary scenario at the home environment is that a couple usually watches TV programs side by side in sofa. Thus, the side effect of 3D perception was designed to be investigated also by adding another two viewing locations with 15 degrees away from the center positions. Therefore, x and y coordinates of the four viewing locations were (0m, 2.2m), (0.6m, 2.2m), (0m, 3.9m), and (1.1m, 3.9m), in which we assumed that the X and Y axis were parallel and perpendicular to the silver screen respectively and the origin was the projection point of the screen center on the ground. Subsequently, the four locations were named as Near Center, Near Side, Far Center, Far Side, respectively.

Second, test stimuli with a wide range of disparities were constructed from different scene contents, camera baselines and image resolutions. These three factors have mutual influence on disparity. Particularly, camera baseline determines the general disparity information of all objects in a scene, while the disparity difference between different objects is dependent on their relative depth structures of the scene content. Moreover, since disparity is usually treated as pixel unit in an image instead of metric unit, the disparity in a unit of pixel is also influenced by the resolution of an imaging sensor.

Five indoor and seven outdoor scenes, as shown in Figure D.2, were selected from the EPFL database [1, 2]. Among these scenes, contents Trees and Grass were selected as the training examples and dummy test stimuli, respectively. All the scene contents were originally captured at the resolution of 1920×1080 and stored at a high quality without any perceptible compression distortions. The main consideration of combining outdoor and indoor scenes is to assess different depth ranges of scene content for various disparity structures. In addition, these scene contents also contain various complex features in textures and colors. The captured images were post processed by spatial and color alignment and then used as the test stimuli. For each scene content, three camera baselines (10cm, 20cm and 30cm) and two resolutions (1280×720 and 720×405) were selected to constitute different disparities, which were located in different range of maximal disparities (20 and 35 pixels) of our display. In particular, three baselines were shot originally, but two resolutions were sampled from 1920×1080 using bicubic interpolation. Two screen sizes corresponding to the two resolutions (1280×720 and 720×405) were simulated because the images were presented in their actual sizes. Since screen size was what perceived directly by viewers, we would rather call the factor resolution as screen size and name the two screen sizes 1.86m×1.00m and 1.05m×0.56m as Large and Small, respectively. If the two screen sizes were combined with the two designed viewing distances, there were four field of views (FOV) which were 15°, 27°(two cases) and 46°. Table D.1 lists the constructed maximum disparity for each scene in various camera baselines and screen sizes, which were measured manually by the authors. It can be seen from the table that scene contents in configurations of (10cm, Large), (10cm, Small) and (20cm, Small) were mostly located in the range of maximal display disparity 35 pixels. This indicates that these stimuli can be viewed without any discomfort in a far viewing distance. Others configurations were

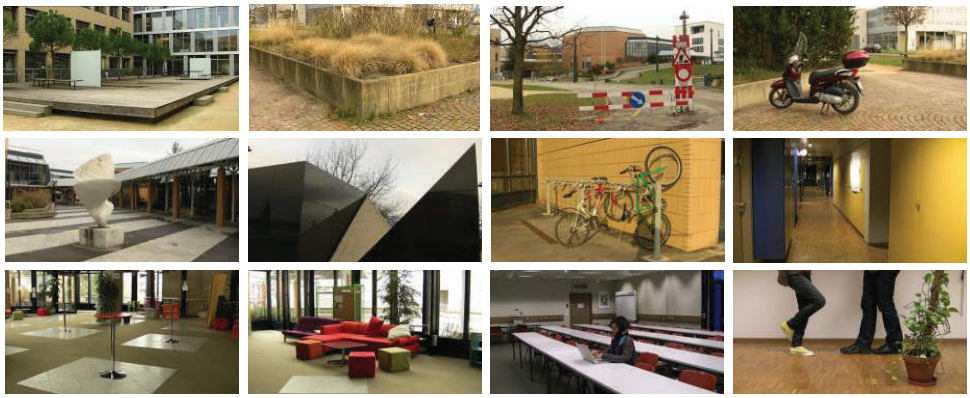


Figure D.2: Visual samples of different scenes selected from the EPFL database. From top left to bottom right: Trees, Grass, Construction, Moped, Sculpture, Closeup, Bikes, Hallway, Tables, Sofa, Notebook, Feet.

Table D.1: Maximum disparity of different scene contents at three camera baselines and two screen sizes.

Title	10cm		20cm		30cm	
	Large	Small	Large	Small	Large	Small
Construction	26	9	36	20	54	31
Moped	29	17	65	37	59	55
Sculpture	19	11	43	24	64	36
Closeup	20	11	46	26	72	40
Bikes	26	14	50	28	77	44
Hallway	50	28	103	58	156	88
Tables	23	13	49	28	78	44
Sofa	23	13	45	25	69	39
Notebook	36	20	77	43	131	74
Feet	35	20	72	40	134	75

beyond 35 pixels but in different extents for the cases of uncomfortable viewing. When a near viewing distance was considered, scene content in configuration of (10cm, Small) is the only one in the comfortable viewing zone.

D.3 Test Methodology

D.3.1 Single Stimulus

Among methodologies for subjective quality assessment of SDTV pictures in ITU-R BT. 500 [10], three widely used methods are double stimulus continuous quality

scale (DSCQS), double stimulus impairment scale (DSIS), and single stimulus (SS). In this study, as several camera baselines were considered for each scene in subjective tests, an original reference 3D image was not available. Therefore, we adopted the SS method. The SS method was also adopted for assessing the quality levels of stereoscopic images with varying camera baselines in [7].

D.3.2 Subjects

Before training sessions, visual perception related characteristics of subjects were collected, including pupillary distance (measured by a ruler), normal or corrected binocular vision (tested by the Snellen chart), color vision (tested by the Ishihara), and stereo vision (tested by the TV-04 and TV-07 in ITU-R BT. 1438 [9]). A total of 30 subjects took part in the tests, consisting of 18 males and 12 females, aged from 23 to 47 years old. All the subjects have binocular visual acuity larger than 1.0, stereo vision smaller than 30 second of arc and no colorblind. The pupillary distance of subjects varied between 62 and 74 millimeters.

D.3.3 Train Session

In this study, two subjects as a group were involved in both training and test sessions. During the training session, the two subjects sit in Far Center and Far Side viewing locations, where the examples of the five categorical adjectival levels for Trees (see Table D.2) were shown in order to benchmark and harmonize the measure scale among subjects. These examples had been selected by expert viewers in a way that each quality level is represented by an example and that the full range of quality levels within the set of test stimuli is covered. When an example was presented, the experimenter verbally explained to the subjects with the corresponding quality. The subjects were told to judge quality by comparing from the positive 3D effect to the negative, in which the positive effect includes the binocular depth, immersivity and reality, while the negative effect consists of discomfort arising from fusion limitation, visible binocular artifacts and geometry distortion. In addition, a detailed definition of each scale was provided to the subjects in a form of written instructions (see Table D.2). The subjects were encouraged to view the representative examples as long as they wished and ask questions if they needed any further clarifications. Answering sheet marked with the corresponding quality level (diagonal line on the continual scale) was shown to the subjects when the examples were displayed in order to make them be familiar with the same procedure which would process in the real test session. Training sessions would continue until the subjects could understand and distinguish the five different quality levels.

D.3.4 Test Session

Since the number of test configurations (240) was too large for a single session, each test session was spitted into two sub-sessions by the experimenter. During each sub-session, 60 test images were randomly displayed to two subjects A and

Table D.2: Explanations of five categorical adjectival levels and training examples of the Trees scene

Levels	**Explanation**	**Examples** (baseline, size)
Excellent	Positive≫Negative: Enjoy the experience very much.	(10cm, Large)
Good	Positive>Negative: Accept viewing such quality happily in daily life.	(20cm, Large)
Fair	Positive≈Negative: Accept viewing such quality but reluctantly.	(30cm, Small)
Poor	Positive<Negative: Refuse to accept viewing such quality.	(40cm, Small)
Bad	Positive≪Negative: Feel headache and do not want to view 3D again.	(50cm, Large)

B. If they started with Far Center (A), Far Side (B) in the first test sub-session, then they would switch to Near Center (B), Near Side (A) in the second test sub-session, or they can start with Near locations first. Additionally, at the beginning of each test sub-session, the subjects were first presented with 3 dummy 3D images of Grass (50cm, Large), (10cm, Small) and (30cm, Large) in order. These dummy images were used to stabilize the subject's judgment only, and the corresponding scores were not used in the analyses. During the whole test period, the subjects were not allowed to ask questions anymore and have any interruption during entire test sessions. Each sub-session lasted about 15 minutes, where each test image was displayed for 10 seconds, with another 5 seconds for entering the score on the answering sheet. A 5 minutes break was arranged to make the subjects relax the eyes and body, but still stay in the test environment. In this way, we obtained 15 samples for each viewing location, because in total 15 groups of subjects attended the test.

D.4 Results Analysis

The subjective scores of the 3D images evaluated by 30 subjects are analyzed in this section. Particularly, we aim to analyze how QoE depends on scene content, camera baseline, image size and viewing location.

D.4.1 Normality Test and Outlier Removal

In order to apply arithmetic mean value as Mean Opinion Score (MOS) and use parametric statistical analysis such as ANalysis Of VAriance (ANOVA), the normality of subjective scores across subjects needs to be satisfied. The β_2 test [4] based on calculating the kurtosis coefficient was adopted for a normality test. We

classified the β_2 test results into three groups, which were normal ($2 \leq \beta_2 \leq 4$), close to normal ($1 \leq \beta_2 < 2$ or $4 < \beta_2 \leq 5$) and abnormal ($\beta_2 < 1$ or $\beta_2 > 5$). If the total ratio of normal and close to normal was more than 75%, we assume the subjective scores are distributed normally. The results showed that the majority (159 of 240) were normal and (48 of 240) were close to normal while others (33 of 240) were abnormal. Therefore, the subjective scores across the subjects followed a normal distribution. The screening of subjects was also performed according to the guidelines described in [4]. Subjects who had produced votes significantly distant from the average scores should be removed. No outlier was indentified.

D.4.2 MOS and Observations

After the outlier removal, MOS and 95% Confidence Interval (CI) were computed and plotted as a function of camera baseline and viewing conditions for all the 10 scene contents separately, as shown in Figure D.3. It can be seen that the confidence intervals are usually very small, which indicates that results given by different subjects are highly correlative and that the training/test sessions were effective. A number of observations can be made based on the results in those plots.

Generally speaking, the MOS values are quite different across scene contents and they decrease as the camera baseline increases. Therefore, both scene content and camera baseline may have impact on the QoE individually. Moreover, when the camera baseline increases, the amount of reduction in MOS for Moped, Hallway, Tables and Notebook scenes is more obvious than that in other scenes. It indicates that the differences among camera baselines are not same for each scene contents, which may indicate that there is an interaction between camera baseline and scene content. Moreover, when considering the factor of screen size, it can be seen that the difference between Large and Small sizes on Construction and Bikes is larger than that on Closeup and Tables. Thus, there might be an interaction between screen size and scene content. For the contents Moped, Bikes, Hallway, Sofa, Notebook and Feet, we can also see that the difference between Large and Small sizes at different camera baselines is relatively large. Thereby, the interaction between screen size and camera baseline or even between screen size, camera baseline and scene content might exist. However, no clear difference has been observed between Near and Far, Center and Side, since different viewing locations stick to each other in all scene contents, camera baselines and screen sizes. Therefore, viewing locations might have no impact of on QoE.

D.4.3 Statistical Analysis

In order to verify the aforementioned observations, ANOVA was adopted to study the impact of independent parameters (scene content, camera baseline, screen size and viewing location) on dependent parameter (stereo QoE). ANOVA is a general technique that can be used to test the equality hypothesis of means among two or more groups. These groups are classified by factors (independent variables whose

settings are controlled and varied by the experimenter) or levels (the intensity settings of a factor). An N-way ANOVA treats N factors and the null hypothesis includes 1) there is no difference in the means of each factor; 2) there is no interaction between n-factors ($2 \leq n \leq N$). The null hypothesis is verified using F-test and can be easily judged through p-value. When p-value is smaller than 0.05, the null hypothesis is rejected, which means there is a significant difference in means. In particular, there is a significant difference between the levels of a factor such that the factor has significant effect; or the difference between the levels of one factor is not same for the levels of other factors such that there is an interaction between different factors.

We used Statistical Package for the Social Sciences (SPSS) statistics 17.0 for the analysis and Figure D.4 shows the results of ANOVA analysis for the entire null hypothesis. In the figure, C, B and LS denote within-subjects factors scene content, camera baseline and screen size, respectively. NF and CS denote two groups of viewing locations, Near versus Far and Center versus Side, which are between-subjects factors. The * denotes interactions between these factors, and the horizontal red line equaling to 0.95 indicates a threshold for a significant difference. It can be seen from the figure that C, B, LS*C, LS*B, C*B and LS*C*B are bigger than the significant difference threshold. It means that scene content and camera baseline have significant impact on QoE individually, as well as there are 2-factors and 3-factors interactions between screen size, scene content and camera baseline in terms of QoE. However, other factors have no significant impact or interaction on QoE. These results are consistent with what we have observed in Subsection D.4.2.

D.5 Conclusions

In this study, we have investigated how four factors scene content, camera baseline, screen size and viewing location impact the overall stereoscopic QoE. Typical test configurations were designed to cover a full range of QoE in terms of disparity coverage between constructed disparity by shooting condition (scene content, camera baseline, sensor resolution/screen size) and maximal disparity supported by viewing environment (viewing location). Then a series of subjective evaluations of QoE were conducted. Both the observations from MOS scatter plots and ANOVA statistical results show that scene content, camera baseline, as well as the interactions between screen size, scene content and camera baseline, have significant impact on QoE of stereoscopic images, while other combinations, especially viewing location involved, have no significant impact.

The ground truth as well as the statistical analysis of results can be further used to compare and design objective stereo quality metrics. We have also developed an objective metric to model QoE, the preliminary results shows a correlation ratio of 87% with the ground truth. Such metrics can be applied to prepare scene content of shot and set camera parameters for photographers, as well as recommend optimal screen sizes and viewing locations.

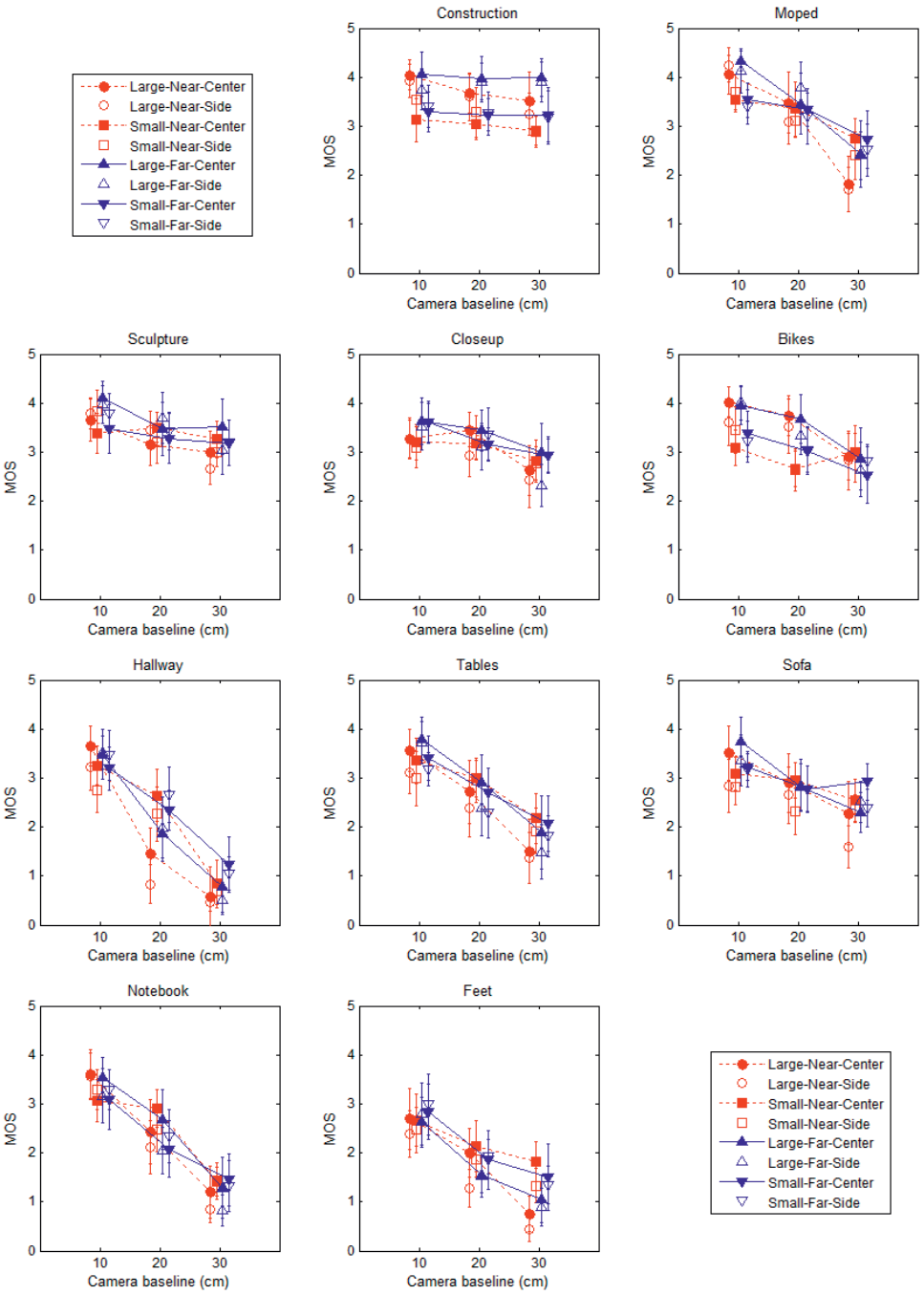


Figure D.3: MOS and CI (significance of 95%) of subjective scores for all test images.

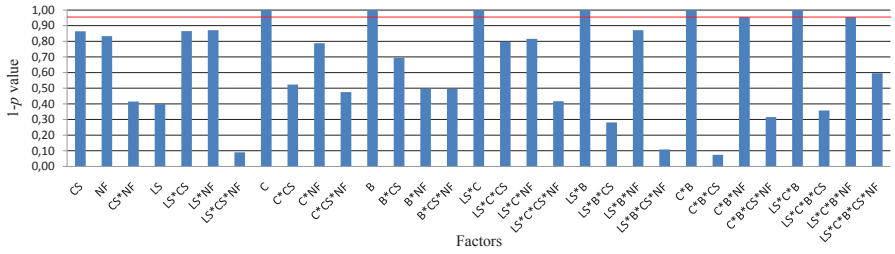


Figure D.4: Significant difference histogram of a 5-way ANOVA.

References

[1] 3D Image Quality Assessment. http://mmspg.epfl.ch/3diqa. 2010.

[2] 3D Video Quality Assessment. http://mmspg.epfl.ch/3dvqa. 2010.

[3] P. Aflaki, M. Hannuksela, J. Hakkinen, P. Lindroos, and M. Gabbouj. Subjective study on compressed asymmetric stereoscopic video. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 4021–4024, 2010.

[4] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet. New requirements of subjective video quality assessment methodologies for 3DTV. In *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2010.

[5] L. Goldmann, F. De Simone, and T. Ebrahimi. A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. In *Three-Dimensional Image Processing (3DIP) and Applications*, volume 7526, 2010. [available] http://mmspg.epfl.ch/3diqa, http://mmspg.epfl.ch/3dvqa.

[6] J. Häkkinen, T. Kawai, J. Takatalo, T. Leisti, J. Radun, A. Hirsaho, and G. Nyman. Measuring stereoscopic image quality experience with interpretation based quality methodology. In *Image Quality and System Performance V*, volume 6808, pages 68081B–68081B–12, 2008.

[7] W. IJsselsteijn, H. de Ridder, and J. Vliegen. Subjective evaluation of stereoscopic images: effects of camera parameters and display duration. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(2):225–233, 2000.

[8] W. IJsselsteijn, H. d. Ridder, J. Freeman, S. E. Avons, and D. Bouwhuis. Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence. *Presence: Teleoper. Virtual Environ.*, 10(3):298–311, 2001.

[9] ITU-R BT.1438. Subjective assessment of stereoscopic television pictures. 2000.

[10] ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television pictures. 2002.

[11] L. Kaufman, J. Kaufman, R. Noble, S. Edlund, S. Bai, and T. King. Perceptual distance and the constancy of size and stereoscopic depth. *Spatial Vision*, 19(5):439–457, 2006.

[12] S. Kishi, S. H. Kim, T. Shibata, T. Kawai, J. Häkkinen, J. Takatalo, and G. Nyman. Scalable 3D image conversion and ergonomic evaluation. In *Stereoscopic Displays and Applications XIX*, volume 6803, pages 68030F–68030F–9, 2008.

[13] K. Klimaszewski, K. Wegner, and M. Domanski. Distortions of synthesized views caused by compression of views and depth maps. In *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009*, pages 1–4, 2009.

[14] B. Mendiburu. *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*. Focal Press/Elsevier, 2009.

[15] S. Patel, H. Bedell, D. Tsang, and M. Ukwade. Relationship between threshold and suprathreshold perception of position and stereoscopic depth. In *J. Opt. Soc. Am. A Opt. Image Sci. Vis.*, volume 26, pages 847–861, 2009.

[16] P. J. H. Seuntiens, L. M. J. Meesters, and W. A. IJsselsteijn. Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation. *TAP*, 3(2):95–109, 2006.

[17] P. Seuntiëns, L. Meesters, and W. IJsselsteijn. Perceptual attributes of crosstalk in 3D images. *Displays*, 26(4–5):177–183, 2005.

[18] A. Woods, T. Docherty, and R. Koch. Image distortions in stereoscopic video systems. In *Stereoscopic Displays and Applications IV*, volume 1915, pages 36–48, 1993.

[19] H. Yamanoue, M. Okui, and F. Okano. Geometrical analysis of puppet-theater and cardboard effects in stereoscopic HDTV images. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16(6):744–752, 2006.

E Objective Metrics for Quality of Experience in Stereoscopic Images

Author

Liyuan Xing
Junyong You
Touradj Ebrahimi
Andrew Perkis

Conference

IEEE International Conference on Image Processing (ICIP), 2011.

Abstract

Stereoscopic quality of experience (QoE) is the result of a complex combination of different influencing factors. Previously we had investigated the effect of factors such as scene content, camera baseline, screen size and viewing position on stereoscopic QoE using subjective tests. In this paper, we propose two objective metrics for predicting stereoscopic QoE using bottom-up and top-down approaches, respectively. Specifically, the bottom-up metric is based on characterizing the significant factors of QoE directly, which are scene content, camera baseline, screen size and crosstalk level. While the top-down metric interprets QoE from its perceptual attributes, including crosstalk perception and perceived depth. These perceptual attributes are modeled by their individual relationship with the significant factors and then combined linearly to build the top-down metric. Both proposed metrics have been validated against our own database and a publicly available database, showing a high correlation (over 86%) with the subjective scores of stereoscopic QoE.

E.1 Introduction

Stereoscopic three-dimensional (3D) media services are becoming increasingly popular and expected to be the next big step forward in the evolution of media after addition of color and sound information. Providing the third dimension (binocular depth) by stereoscopic media is thought to be a fundamental change in the revolution of the image presentations and believed to bring new experience to viewers, not just an enhancement of the quality.

However, stereoscopic 3D services do not always prevail when compared to their two-dimensional (2D) counterparts. Nowadays, stereoscopic imaging is based on simultaneously capturing a pair of 2D images, and then separately delivering them to respective eyes. Consequently, 3D perception is generated in the human visual system (HVS). Although the principle is simple, it is difficult to implement stereoscopic imaging in practice. Various limitations, such as excessive disparity, accommodation-convergence mismatch and binocular artefacts may exist with current techniques and potentially degrade the stereoscopic quality of experience (QoE). It is therefore imperative to quantify the QoE of stereoscopic imaging systems.

Specifically, the degrees of aforementioned limitations exhibited in stereoscopic QoE are the result of a complex combination of different influencing factors. These factors include the characteristics of 3D content, the configurations of capture and display systems, selections of image processing mechanism, and the perceptual-physiological limitations of HVS. Usually subjective testing is adopted to quantitatively evaluate the influence of factors on the overall QoE, since it is the most reliable way of assessing the perceived quality. Our previous subjective tests [14] have qualified how scene content, camera baseline, screen size and viewing position impact stereoscopic QoE.

Nevertheless, subjective testing is slow, expensive, and inconvenient for most applications. To deal with these drawbacks, objective metrics that can automatically predict the perceived quality are desired. Such metrics can be employed to monitor, benchmark and optimize the media system. However, an effective quality metric cannot be proposed without a deep understanding of the perception mechanism.

So far, the development of objective quality models for stereoscopic media is still in its early stage. Researchers first started with exploring whether or not 2D quality metrics can be applied to stereoscopic quality assessment [1, 15]. Subsequently, a few objective metrics that take into account the characteristics of stereoscopic images have been proposed [5, 3, 6]. However, most of the stereoscopic objective metrics are designed to assess quality degradations caused by lossy compression schemes. In fact, a minimum stereoscopic system having acquisition and restitution stages only, already fails to provide satisfying viewing experience. To model the non-compression quality degradations, the authors of [4] proposed a metric linearly combining three measurements, including perceived depth, visual fatigue and temporal consistency. We have proposed a metric for crosstalk perception [13] in a top-down approach, which makes use of the perceptual attributes of crosstalk. In addition, two similar objective metrics [11, 12] for QoE assessment were proposed

Table E.1: Overview of EPFL and Q2S-QoE subjective tests

Database	**EPFL[2]**	**Q2S-QoE[14]**
Test stimuli	$9SC \times 6CB \times 1SS \times 1VL$	$10SC \times 3CB \times 2SS \times 4VL$
SC	Various characteristics in color, texture, depth, but no perceptible compression distortions	
CB [m]	0.1;0.2;0.3;0.4;0.5;0.6	0.1;0.2;0.3
SS [m]	1.02×0.57	1.86×1.00; 1.05×0.56
VP(x,y) [m]	(0.0, 2.2)	(0.0, 2.2); (0.6, 2.2); (0.0, 3.9); (1.1, 3.9)
Post-proc.	Geometric and color adjustment	
Display	Hyundai S465D	Polarized projectors
Glasses	Circular	Linear
Resolution	1920×1080	1280×960; 720×405
Test methods	Single stimulus	
Training	Illustrated examples for five categorical levels	
Test session	Random display with dummy at the beginning	
Participant	14 Male, 6 Female	18 Male, 12 Female

in a bottom-up approach with respect to the EPFL database [2] containing acquisition distortions and our own database Q2S-QoE [14] including both acquisition and restitution distortions, respectively. In this paper, we combine the two bottom-up metrics into a single metric for characterizing all the significant factors contributing to QoE and propose a new top-down metric for QoE by describing perceptual attributes from the perception viewpoint.

The rest of the paper is organized as follows. In the next section, the subjective tests and significant factors of QoE are introduced briefly. Then in Section E.3, we present the bottom-up and top-down objective quality metrics in detail. Both metrics are verified on EPFL and Q2S-QoE databases and the experimental results are reported in Section E.4. Finally, we conclude the paper in Section E.5.

E.2 Subjective Evaluation

As a first step towards a reliable objective quality metric, subjective tests were conducted to obtain the human judgement on QoE. Moreover, significant factors were found to guide the design of objective metrics. In the following, we will provide some details of the tests.

E.2.1 Subjective Tests of QoE

As listed in Table E.1, both the EPFL and the Q2S-QoE subjective tests followed a similar evaluation methodology. The main difference was in the test stimuli presented. In particular, more viewing conditions were designed in Q2S-QoE subjective

tests in order to cover a full range of QoE. The design principle is that we believe that stimuli located in different ranges of comfort zone can provide different viewing experiences. Therefore, the comfort zone of the display system in our subjective tests was calculated [14]. For a particular display system, this comfort zone only varies when the viewing position changes. Thus, viewing positions were selected to create stimuli with different disparity levels located in different ranges of comfort zone supported by viewing conditions.

E.2.2 Factors Contributing to QoE

Based on the analysis on the EPFL database, it was concluded that the influence of the camera distance on the quality of the stereoscopic images is largely scene-dependent [2]. Furthermore, when the Q2S-QoE database was analyzed [14], we demonstrated that scene content, camera baseline, as well as the interactions between screen size, scene content and camera baseline, have significant impact on QoE in stereoscopic images, while other factors, especially viewing position, have virtually no significant impact.

However, crosstalk level of the display system was not observed in these two subjective tests. In [10], we demonstrated that crosstalk level, camera baseline and scene content have significant impacts on crosstalk perception, respectively, and that these three factors have various interactions in terms of impact on crosstalk perception. Since crosstalk is one of the most annoying distortions in the visualization stage of a stereoscopic imaging system, it obviously impacts QoE. Thus, crosstalk level is another significant factor.

E.3 Objective Quality Metrics

We believe that objective metrics can be proposed from two viewpoints, namely bottom-up and top-down. A bottom-up metric describes the significant factors contributing to QoE, while a top-down metric models the perceptual attributes of QoE, to predict the viewing experience. This section will present these processes in detail.

E.3.1 Bottom-up Metric

Based on our understanding of QoE factors mentioned earlier, significant factors to be modelled for predicting QoE are scene content, camera baseline, screen size and crosstalk level, as well as some of their interactions. As we illustrated in [12], the disparity map is impacted mainly by scene content, camera baseline, screen size/image resolution and their interactions. This means that the significant factors of the disparity map are the sum of the six significant factors of QoE plus a single factor, screen size, when crosstalk level is not considered. This similarity between disparity map and QoE regarding significant factors proves the disparity map as

an efficient expression of QoE. Thus, the metric was developed as follows:

$$R_{dis} = SSDMF(R_o, L_o) \tag{E.1}$$

$$V_{bu} = AVG(1 - R_{dis}/255) + w \times S \tag{E.2}$$

where R_o and L_o are the original right and left displayed image pair. R_{dis}, S and w denote the disparity map, screen size and weight of screen size on QoE, respectively. $SSDMF$ and AVG denote the SSDMF method in [7] for disparity estimation and the averaging operation. V_{sdis} is the quality metric, which combines the disparity and screen size.

The metric in [11] for QoE takes into consideration crosstalk level but not screen size. Since the crosstalk level can be reflected by an SSIM map [13], the QoE metric was then modelled by the following equations:

$$\begin{cases} R_c = R_o + p \times L_o \\ L_c = L_o + p \times R_o \end{cases} \tag{E.3}$$

$$L_s = SSIM(L_o, L_c) \tag{E.4}$$

$$V_{cdis} = AVG(L_s \times (1 - R_{dis}/255)) \tag{E.5}$$

where p is the level of crosstalk distortion, and R_c and L_c denote the distorted right and left views after adding crosstalk. L_s denotes the left view SSIM map which is obtained using the SSIM algorithm in [9]. V_{cdis} is the final predicted value, which is the averaged from a weighted disparity map by the SSIM map. The value of p should be set according to the equipment introduced crosstalk level. In our experiments, p is assumed to be 3%.

In this paper, we combine the above two metrics into a single metric to describe all the significant factors of QoE. Furthermore, a special case when disparity is very small is also included. We believe that 3D perception does not exist anymore when the disparity is small enough and QoE should be expressed by a conventional 2D metric instead. Meanwhile, there should be no abrupt change at the transition zone from 3D metric to 2D metric. Moreover, both the transition zone and the final 2D perception are determined by the content. The bottom up metric is thus summarized as follows,

$$C_{bu} = \begin{cases} AVG(L_s \times (1 - R_{dis}/255)) & if AVG(R_{dis}(i,j)) \geq C \\ (1 - C/255) \times AVG(L_s) & if AVG(R_{dis}(i,j)) < C \end{cases} \tag{E.6}$$

$$V_{bu} = C_{bu} + \alpha \times S \tag{E.7}$$

where threshold C varies across different scene contents. When the average of R_{dis} is larger than C, the combination C_{bu} follows as equation E.5, otherwise, it multiplies a 2D metric $AVG(L_s)$ by a constant $1 - C/255$. Moreover, the final

bottom-up metric V_{bu} is the linear combination between C_{bu} and S in the same way as equation E.2. Specifically, the content dependent threshold C is the average of minimal disparity of each content in our database. The values of S were 1.86 and 1.05 for two screen sizes, respectively. An optimal weight 0.07 was obtained for using an optimization method on the Q2S-QoE database.

E.3.2 Top-down Metric

The top-down metric is proposed based on the understanding of perceptual attributes of QoE. It was pointed out in [8] that the overall QoE is a trade-off between perceived image distortion, perceived depth and visual strain. Specifically, crosstalk is one of the most annoying distortions of perceived image distortion and cause visual discomfort. While binocular depth is the main contribution by stereoscopic imaging, we believe that it is also possible to model the QoE from its two main perceptual attributes, namely a negative attribute crosstalk perception and a positive attribute depth perception.

The crosstalk perception was modeled in our previous work [13] in a top-down way. The three perceptual attributes of crosstalk are shadow degree, separation distance and screen deviation of crosstalk. In particular, shadow degree of crosstalk is determined by the contrast of scene content and crosstalk level, while separation distance of crosstalk is decided by camera baseline and relative depth structure of scene content together, namely, disparity. Both attributes can be reflected by SSIM map. Moreover, screen deviation of crosstalk is the distance between reconstructed object and viewer screen, and perception of crosstalk is mostly dominated by the crosstalk which has the maximal screen deviation. Thus, maximal disparity was used for representing the maximal screen deviation. The crosstalk perception metric is therefore the integration between the SSIM map and maximal disparity as follows,

$$V_c = AVG(L_s)/MAX(R_{dis}) \tag{E.8}$$

We believe that the depth perception should reflect both the percentage of pixels whose depth located in the comfort zone of the display system and the depth amplitude of those pixels. In particular, the depth amplitude could be represented by depth plane or pixel disparity. Moreover, since the viewing position which impacts the comfort zone is not a significant factor influencing QoE, only one depth threshold is specified. Thus the depth perception part is defined by the following equation:

$$V_d = SUM(R_{dis}(i,j))/Res \qquad if R_{dis}(i,j) \geq D \tag{E.9}$$

where Res is the resolution of the display image, D denotes the depth threshold, which is 11 pixels in our case. This threshold is obtained using an optimization method on the Q2S-QoE dataset.

Finally, the overall QoE is the linear combination between crosstalk perception V_c and depth perception V_d as shown in the following equation:

$$V_{td} = V_c + \beta \times V_d \tag{E.10}$$

Table E.2: Evaluation results of different metrics

Database	Criterion	PSNR	SSIM	V_{bu}	V_{td}
	RMSE	0.998	0.901	0.421	0.508
EPFL	**Pearson**	0.509	0.503	0.915	0.874
	Spearman	0.514	0.424	0.927	0.885
	RMSE	0.774	0.716	0.394	0.862
Q2S-QoE	**Pearson**	0.424	0.546	0.888	0.862
	Spearman	0.460	0.545	0.848	0.745

where β is a weight. The optimal weight calculated on the Q2S-QoE dataset was 0.08.

E.4 Experimental Results

Both proposed metrics are validated against the EPFL and Q2S-QoE databases. They are compared with traditional 2D metrics PSNR and SSIM, which are calculated between the original left and the crosstalk added left image, as in [12].

To evaluate each metric V, root mean squared error (RMSE), Pearson correlation coefficient, and Spearman rank-order correlation coefficient are selected as criteria. They are calculated between objective values MOSp after a nonlinear regression using equation E.11, suggested by VQEG, and the subjective scores MOS.

$$MOS_p = b_1/(1 + exp(-b_2 \times (r(V) - b_3))) \tag{E.11}$$

Table E.2 gives the evaluation results of objective metrics on both databases. It can be seen that both proposed metrics exhibit similar performance and are significantly better when compared to PSNR and SSIM. This implies that metrics considering 3D characteristics can give better prediction of stereoscopic QoE when compared to purely 2D metrics. Moreover, the performance of Vtd is slightly inferior to that of Vbu in both databases. However, an advantage of Vtd is that it can describe the best viewing experience, which could exist. This happens when all pixels have their depth in comfort zone and the decrease of depth perception compensates the increase of crosstalk perception contributing to QoE. However, Vbu makes QoE always increase with the decreasing disparity. The latter however, could not be verified in this paper because of the lack of available dataset in which the disparity is very small. As the Pearson correlation of the proposed metrics are over 86% on both databases, the proposed metrics are promising for evaluation of QoE in stereoscopic presentations.

E.5 Conclusions

In this paper, we have introduced results from prior QoE subjective tests and briefly identified the main contributing factors in stereoscopic QoE. We have mainly sum-

marized a bottom-up metric and proposed a top-down metric for QoE in details. Both metrics take the characteristics of QoE on stereoscopic images into consideration from different viewpoints. Specifically, the bottom-up metric describes the significant factors and the top-down metric models the perceptual attributes of QoE, respectively. The experimental results on EPFL and Q2S-QoE databases demonstrated the promising performance of the proposed metrics, achieving more than 86% in Pearson correlation with MOS. However, the robustness of both proposed metrics to small disparity cases and the type of displays with different crosstalk levels and screen sizes need to be further validated in future work.

References

[1] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau. Quality assessment of stereoscopic images. *EURASIP Journal on Image and Video Processing*, 2008(1):659024, 2008.

[2] L. Goldmann, F. De Simone, and T. Ebrahimi. A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. In *Three-Dimensional Image Processing (3DIP) and Applications*, volume 7526, 2010. [available] http://mmspg.epfl.ch/3diqa, http://mmspg.epfl.ch/3dvqa.

[3] P. Gorley and N. Holliman. Stereoscopic image quality metrics and compression. In *Stereoscopic Displays and Applications XIX*, volume 6803, pages 680305–680305–12, 2008.

[4] D. Kim, D. Min, J. Oh, S. Jeon, and K. Sohn. Depth map quality metric for three-dimensional video. In *Stereoscopic Displays and Applications XX*, volume 7237, pages 723719–723719–9, 2009.

[5] R. Olsson and M. Sjostrom. A depth dependent quality metric for evaluation of coded integral imaging based 3D-images. In *3DTV Conference, 2007*, pages 1–4, 2007.

[6] Z. Sazzad, S. Yamanaka, Y. Kawayokeita, and Y. Horita. Stereoscopic image quality prediction. In *Quality of Multimedia Experience (QoMEX), 2009 International Workshop on*, pages 180–185, 2009.

[7] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47(1-3):7–42, 2002.

[8] P. Seuntiëns, L. Meesters, and W. IJsselsteijn. Perceptual attributes of crosstalk in 3D images. *Displays*, 26(4–5):177–183, 2005.

[9] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[10] L. Xing, T. Ebrahimi, and A. Perkis. Subjective evaluation of stereoscopic crosstalk perception. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 77441V–77441V–9, 2010.

[11] L. Xing, J. You, T. Ebrahimi, and A. Perkis. In *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*.

[12] L. Xing, J. You, T. Ebrahimi, and A. Perkis. Estimating quality of experience on stereoscopic images. In *Intelligent Signal Processing and Communication Systems (ISPACS), 2010 International Symposium on*, pages 1–4, 2010.

[13] L. Xing, J. You, T. Ebrahimi, and A. Perkis. A perceptual quality metric for stereoscopic crosstalk perception. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 4033–4036, 2010.

[14] L. Xing, J. You, T. Ebrahimi, and A. Perkis. Factors impacting quality of experience in stereoscopic images. volume 786304, pages 786304–786304–8, 2011.

[15] J. You, G. Jiang, L. Xing, and A. Perkis. Quality of visual experience for 3D presentation - stereoscopic image. In M. Mrak, M. Grgic, and M. Kunt, editors, *High-Quality Visual Experience*, Signals and Communication Technology, pages 51–77. Springer Berlin Heidelberg, 2010.

F Towards Certification of 3D Video Quality Assessment

Author

Andrew Perkis, Junyong You, Liyuan Xing
Touradj Ebrahimi, Francesca De Simone, Martin Rerabek
Panos Nasiopoulos, Zicong Mai, Mahaa Pourazad
Kjell Brunnstrom, Kun Wang, Borje Andren

Conference

International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), 2012.

Abstract

Subjective quality assessment is widely used to understand and to study human perception of multimedia quality and as a basis for developing objective metrics to automatically predict the quality of audiovisual presentations. There are several recognized international protocols and procedures for reliable assessment of quality in multimedia systems and services, with emphasis on speech, audio and video modalities. However, the aspect of certification is not yet well understood in this context. This paper discusses various issues regarding certification of multimedia quality assessment. To be concrete, the discussion is illustrated by the procedure implemented to assess 3D video compression technologies within the MPEG effort for the definition of a 3D video coding standard. Selected results from four laboratories, Acreo, EPFL, NTNU and UBC, which participated in the assessment are presented. This case study is used in an early attempt to define a process for certification of subjective test campaigns, based on a cross-validation of the test results across different laboratories, towards the ultimate goal of quality of experience (QoE) certification.

F.1 Introduction

With the rapid growth of three-dimensional (3D) video technology, standardized compression algorithms for 3D video are needed. In 2011, MPEG committee issued a Call for Proposal (CfP) on 3D video coding technology with the objective to "define a data format and associated compression technology to enable the high-quality reconstruction of synthesized views for 3D displays" [5]. Both stereoscopic and auto-stereoscopic multiview display technologies were targeted. Responding to this call, 22 proponents submitted their 3D video coding algorithms for competition. In order to analyze and to compare the performance of the proposed technologies a formal subjective quality evaluation was carried out, and a set of test video sequences, encoded with the proposed technologies, was produced. The European COST Action QUALINET (European Network on Quality of Experience in Multimedia Systems and Services) was invited by MPEG to take part in the evaluation campaign of this test material, referred to as 3DV tests in the rest of this paper.

We believe a critical issue in any subjective evaluation is the establishment of a proper certification mechanism to carry out quality evaluations, such as those performed in 3DV tests. Certification usually refers to the confirmation of certain attributes of an object, organization, person, or a process of production [1]. For example, the ISO 9000 family of standards have been designed and issued by ISO to ensure manufacturers can meet certain requirements for the quality of their management. In video quality assessment scenarios, such as in 3DV tests, a standard certificate mechanism also becomes critical. Naturally, a recognized quality assessment experiment conducted by different laboratories, using identical video content and following similar methodologies and instructions, can serve as an appropriate platform for demonstrating the certification procedure of quality assessment. In such a process, cross-laboratory analysis should be performed to find out whether or not consistent results can be obtained. To this end, four laboratories participating in 3DV test have made an effort to illustrate steps towards certification mechanisms. A cross-laboratory analysis has been performed to estimate the correlation of quality scores obtained by each laboratory and to perform a significance test. These analyses show that laboratories employing different subjects could still produce highly correlated results, as they follow similar guidelines to carry out assessments. This confirms that the participating laboratories fulfill an essential condition towards their certification.

The remainder of the paper is organized as follows. Section F.2 discusses issues and important steps towards a formal certification procedure of video quality assessment. Section F.3 introduces the MPEG 3DV quality test. Test results obtained in the four laboratories and relevant analyses are presented in Section F.4. Finally, concluding remarks and discussions on future work are provided in Section F.5.

F.2 Towards Certification Procedure

As introduced in Section F.1, certification generally refers to the confirmation of certain characteristics of an object, a person, or an organization. This confirmation is often but not always, provided by some form of external review, academic degree, or assessment. In first-party certification, an individual or organization providing a good or service, offers assurance that it meets certain claims. In second-party certification, an association to which the individual or organization belongs, provides such an assurance. Third-party certification involves an independent assessment declaring that specified requirements pertaining to a product, a person, a process or a management system have been met [1].

As an important step towards QoE certification, the European COST Action QUALINET is making an effort to better understand the concept of certification in QoE assessment scenario. The QUALINET Memorandum of Understanding states:"Observing that there are currently no European networks focusing on the concept of QoE, this certification task also aims at bringing a substantial scientific impact on fragmented efforts carried out in this field, by coordinating the research under the catalytic COST umbrella, and at setting up a European network of experts facilitating transfer of technology and know-how to industry, coordination in standardization, and certification of products and services." [9, 8]

The general objective of certification is to assess and to guarantee uniformity of devices, processes or installations, thus allowing improving interoperability and checking quality targets. In this context, a certification process may target either products, such as laboratories and codec, or services such as quality assessment, or even some content.

Within QUALINET, the certification process may follow one or more of the following main approaches:

- Certification entity based: This approach is more centralized, as certification entities should also be certified themselves. Such an approach may, in principle, be more credible and reliable, but requires a well defined process to certify the certification entities.
- Auto-certification: This approach is not centralized. Each organization may certify itself its systems or services, based on a specific and agreed procedure. At the expense of a lighter process, the reliability and credibility of such an approach may be less certain.
- Peer-certification: This approach is not centralized neither, but requires other entities (not necessarily certified themselves) to provide confirmation or opinions about the degree of fulfillment of identified requirements by a system or service. The use of social networks and open peer review mechanisms, if properly implemented, can contribute to increase the reliability and credibility of such an approach.

In addition to the above, certification process may also involve the following elements and entities:

- Certification applicant: an institution making a certification request (to a certification entity) for a product, service or content.
- Certification entity: an institution or individual, which has either undergone a process to become an authority, or simply acts as an open peer reviewer.
- Certificate or label: a diploma or label that provides a proof to applicant, endorsing the product or service provided fulfills a well-defined set of requirements, along with other information such as the underlying conditions, including duration of the certificate.

Depending on the types of certifications to be addressed, adequate certification methodologies can be defined involving the following main steps:

- Certification request: the applicant should present to the certification entity a request, indicating the type of certification requested and providing all the information and elements necessary for this task. This information and elements will be defined in details in a procedure designed for each type of certification, e.g., laboratory facilities certification or content certification.
- Certification assessment: the certification entity will perform the appropriate steps defined, certifying or not the relevant product, service or content.
- Certificate or label: in the case of positive outcome, the applicant will be provided with a proof stating that it may use a label for its products, services or contents, along with additional information and conditions such as the duration of the certificate.

Currently, several accreditation companies professionally perform ISO/IEC 17025 certifications by using specified procedures [2] and forms [4, 3]. If such a procedure is considered, the existing ISO/IEC 17025 standard needs to be taken into account. Since the main goal of certifications consists of guaranteeing that certain standards are met, the issue of providing a legal liability or privacy protection mechanisms might also need to be taken into consideration. More importantly, quality assessment results across different laboratories will play a kernel role in a certification procedure, e.g., certified laboratories should produce correlative results when conducting similar quality evaluations. The next sections of this paper concentrate on this last issue as a first and key step towards certification in quality assessment, with emphasis on a use case in 3D video quality evaluation.

F.3 3DV Tests - Background, Methodology and Laboratory Set Up

F.3.1 Test Material

The 3DV CfP defines some “classes” of test sequences, i.e. sets of spatio-temporal resolutions, and some target coding bit rates. Among them, Class A, with frame

size of 1920×1088 pixels and a frame rate of 25 frames per second, and Class B, with frame size of 1024×768 pixels and a frame rate of 30 frames per second, along with four target coding bit rates, were used for the evaluation of the proponent technologies. For both stereoscopic and auto-stereoscopic codec comparisons, the test materials included four different contents in Class A (Poznan Street, Poznan Hall2, Undo Dancer, GT Fly) and four different contents in Class B (Kendo, Balloons, Lovebird1, and Newspaper). All test materials were progressively scanned and used 4:2:0 color sampling with 8 bits precision per pixel.

The video data evaluated in the subjective tests were generated from a dense set of synthesized views provided by proponents and fed uncompressed into 3D monitors thanks to a specially designed video server configuration. Particularly, two test scenarios, namely, a 2-view input configuration, to be evaluated on stereoscopic display, and a 3-view input configuration, to be evaluated on both auto-stereoscopic as well as stereoscopic display, were considered. The depth data and camera parameters for view synthesis and rendering were also provided. Readers can refer to the 3DV CfP for more details [5].

F.3.2 Proponents

By responding to the CfP, 22 proponents submitted their codec descriptions, and encoded and decoded test sequences at requested target bit rates. Two anchors, i.e., H.264/AVC and HEVC, were also included in the set of coding technologies under assessment.

F.3.3 Laboratories, Hardware, Software, and Instrumentation Set Up

The 3DV tests involved 12 evaluation laboratories from around the world. Each laboratory was assigned a certain number of test sessions, either stereoscopic, auto-stereoscopic, or both, based on the availability of hardware and other facilities. All laboratories used the exact same evaluation methodology, described below, including the same monitors (a 46" Hyundai S465D polarized stereoscopic monitor and a 52" Dimenco BDL5231V auto-stereoscopic monitor, with native resolutions of 1920x1080 pixels), the same implementation of Graphical User Interface (GUI), and similar test room configuration.

The hardware and software environments used in all laboratories were designed and tested to ensure meeting well specified requirements by conducting dry runs before actual evaluations took place.

Eighteen naive viewers evaluated the quality of each test sequence. Since a maximum of 3 (5) subjects could be seated in front of a stereoscopic (auto-stereoscopic) monitor, without deteriorating the perception of the 3D rendering, several subjects could be grouped to attend a same test session. Hence, the test room set up, common in all the laboratories, included 3 to 5 subjects seating in a row, perpendicular to the center of the monitor, for the auto-stereoscopic and the stereoscopic

viewings, respectively. The viewers were seated at roughly 3.5 meters from the auto-stereoscopic monitor, as required in [5], and at roughly 2.3 meters from the stereoscopic monitor, as suggested in the ITU-R BT.710 recommendation for HDTV [7]. The laboratory setup was controlled in order to ensure the reproducibility of results by avoiding as much as possible, involuntary influence of external factors. The test rooms were equipped with a controlled lighting system with a 6500K color temperature and an ambient luminance at 15% of maximum screen luminance. Each laboratory reported the details of the calibration settings used for each monitor, as well as the gender percentages and average age of their sample of viewers, and the exact number of subjects per session.

F.3.4 Test Data Rendering

In order to render correctly the test materials on the stereoscopic and auto-stereoscopic monitors, the following processing was performed on the raw video files received from proponents. For the auto-stereoscopic display, 28 yuv files, each containing a different view for the same video sequence, were interleaved and merged into a single avi file, using an interleaving software tool provided by Dimenco. For the stereoscopic viewing, two pre-selected yuv video sequences, corresponding to the left and right views, were cropped and horizontally shifted in order to obtain a pre-defined depth for each content (different shift parameters were set for different content) and finally interlaced (right view on top) and padded to produce a full HD resolution video.

F.3.5 Evaluation Methodology: Stimulus Presentation and Rating Scale

The Double Stimulus Impairment Scale (DSIS) evaluation methodology was selected to perform the tests. Subjects were presented with pairs of video sequences (i.e., stimuli), where the first was always an unimpaired, reference, video (stimulus A) and the second, the same content processed (stimulus B). Subjects were asked to rate the quality of each stimulus B, keeping in mind that of stimulus A. A dedicated GUI was developed for the test campaign: before each video sequence, a grey screen with the letter “A” (“B”) was shown for two seconds, informing subjects that the reference (test) stimulus would be shown. After the presentation of each pair of sequences, a grey screen with the message “Vote” was shown for five seconds. The test subjects were asked to enter their quality score for the stimulus B in paper scoring sheets during these five seconds.

An 11-grade numerical categorical scale was used [6]. The rating scale ranged from 0 to 10, with 10 indicating the highest quality, i.e., the test sequence is indistinguishable from the reference, and 0 indicating the lowest quality.

F.3.6 Screening

All subjects taking part in the evaluations underwent a screening to examine their visual acuity, using the Snellen chart, and color vision, using the Ishihara chart. Their stereo vision was also tested using the Randot test.

F.3.7 Training

Before each test session, written instructions and a short explanation by a test operator were provided to the subjects. Also, a training session was run to show the GUI, the rating sheets, and examples of processed video sequences. The training video sequences were produced using two different contents (Pantomime and Champagne) and with coding conditions similar to those used to produce the actual test materials.

It is important to stress that the same training instructions were provided to subjects in all the laboratories. Particularly, during training, specific scores were given to the training sequences. These scores had been agreed upon across the evaluation laboratories in order to ensure close correlation and consistency of results.

F.3.8 Test Sessions

A basic test session of DSIS methodology including 24 test pairs, three dummy stimuli pairs, and one reference versus reference pair, was designed. Thus, the test materials resulted in a total of: 16 sessions for each of the two classes of autostereoscopic data, 16 sessions for each of the two classes of 2-view stereoscopic data, 16 sessions for the Class A 3-view stereoscopic data, and 32 sessions for the Class B 3-view stereoscopic data. In each session, the stimulus pairs were presented in random orders, but never with the same video content in consecutive pairs.

F.4 Selected Test Results and Analysis

Some test sessions were performed by more than one laboratory in order to analyze inter-laboratory cross-correlations. In this section, we report the results of the test sessions performed by four laboratories, namely NTNU, Acreo, EPFL, and UBC, including some crossvalidation analysis for the common sessions.

Different overlapping data within each laboratory groups were used. These included:

- Class A, 2-view stereo, 4 sessions (EPFL - UBC)
- Class A, auto stereo, 4 sessions (EPFL - UBC)
- Class B, 2-view stereo, 8 sessions (NTNU - Acreo)

Thus, cross-laboratory results analysis could be performed between EPFL and UBC, and between NTNU and Acreo. Figure F.1 shows the pairwise scatter plots

and correlation coefficients on the overlapping data. It can be observed that the subjective quality results uniformly span over the entire range of quality levels from 0 to 10, which can be considered as an indication of appropriate experiment design and their implementation. More importantly, there exists a high correlation between different laboratories. The Pearson linear coefficient measures the distribution of the points around the linear trend, while the Spearman coefficient measures the monotonicity of the quality scores between different laboratories, that is, how well an arbitrary monotonic function describes the relationship between two sets of data. The results show that the data from NTNU and Acreo, as well as EPFL and UBC, are highly correlated.

Additionally, an ANOVA analysis, where two laboratories were considered as between group variables and the different Processed Video Sequence (PVS) as within group variables, was performed on the raw data, yielding a significant main effect of the "laboratory" variable on the results of the two pairs of laboratories (for instance, on the NTNU-Acreo data: $F(1, 34) = 5.6$, $p = 0.02 < 0.05$). One could also observe an expected significant effect of the PVS (for instance on NTNUAcreo data: $F(223, 7582) = 112.6$, $p = 0.00 < 0.05$), as well as, a significant interaction between the PVS and laboratories, (For instance on NTNU-Acreo data: $F(223, 7582) = 2.2$, $p = 0.00 < 0.05$). Considering the data from NTNU-Acreo, a linear transformation:

$$y = 0.9375 \cdot x + 0.7423 \tag{F.1}$$

(Figure F.2) will make the significant main effect of laboratories disappear ($F(1, 34) = 0.00$, $p = 1.0 > 0.05$), but the significant main effect of PVS ($F(223, 7582) = 111.7$, $p = 0.00 < 0.05$) and interaction between PVS and laboratories remain ($F(223, 7582) = 2.2$, $p = 0.00 < 0.05$). In addition to the above observations, a Student t-test was applied to each pair of PVS from the different laboratories, identifying the significant different PVS, shown in Figure F.3 for the NTNU-Acreo comparison. It can be observed that the significant different PVSs are spread quite evenly over the entire quality range. However, when compared to NTNU, the subjects at Acreo gave significantly higher scores for the mid range qualities, and clearly lower scores at the lower end of the scale. Finally, Figure F.4 compares the score difference between UBC and EPFL at different bit rate levels. Note that in this figure the video bit rate increases from "Rate 1" to "Rate 5". As it is observed, for both the stereo and auto-stereoscopic cases, the score differences are higher at lower bit rates (i.e., low video quality), indicating that subjects had difficulties to precisely quantify low quality content.

This additional comparative analysis indicates that although correlations between laboratories are high and exhibit good correspondence, more complex differences exist in the voting patterns that cannot be modeled by simple transformations, like for instance the linear transformation in Figure F.2.

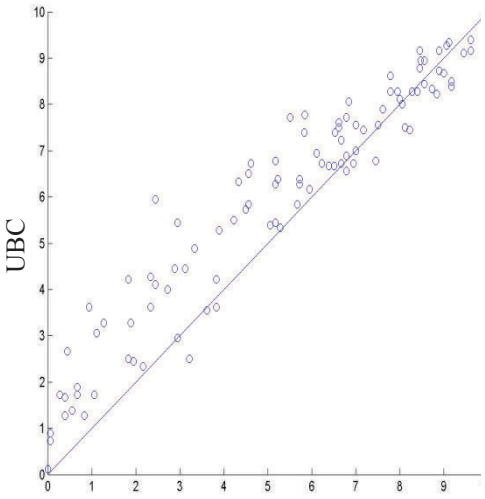
F.5 Conclusions and Further Work

This paper presented cross validation analysis from a recent MPEG 3DV quality test campaign conducted with the help of a European COST Action QUALINET. The test results, obtained from four laboratories in Europe and North America, participating in this test campaign, have been presented and analyzed. Various analyses demonstrated that different laboratories can produce similar quality assessment results when they follow appropriately selected evaluation procedures. The quality test across different laboratories with the identical video contents provides an appropriate first step for laboratory certification purpose. An effort is under way by COST Action QUALINET towards a better understanding of certification mechanism of QoE in multimedia services and systems. This could lead to the definition of a roadmap, which could hopefully help in the implementation of appropriate certification mechanisms in QoE for multimedia applications.

Acknowledgment

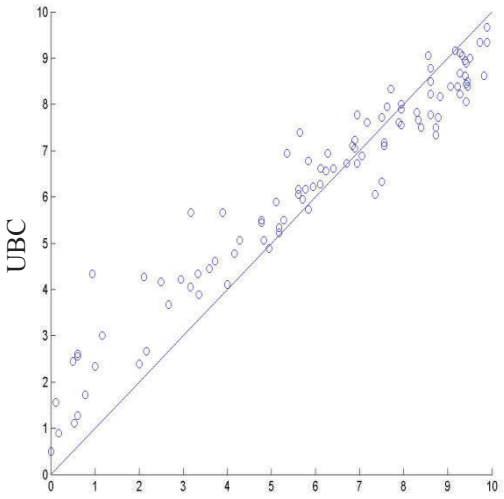
The authors would like to acknowledge the efforts by all those involved in the MPEG 3DV tests, including the MPEG test coordinator, all laboratories which participated in the evaluations, and proponents responding to MPEG 3DV CfP. This work was performed in the framework of the COST Action IC1003, QUALINET.

Pear. 0.95821 Spear. 0.96085



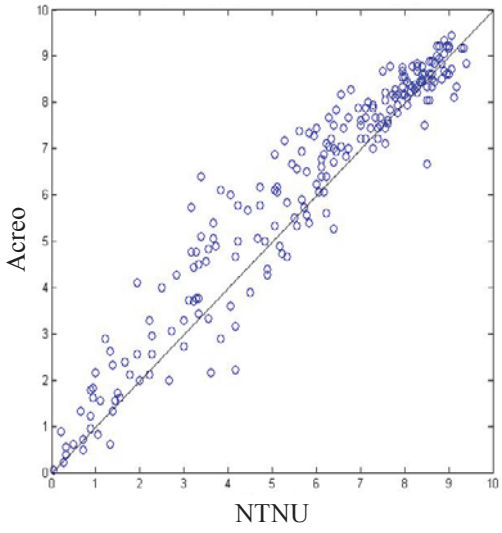
(a) Scatter plot between EPFL and UBC with respect to stereo quality test

Pear. 0.96777 Spear. 0.96333



(b) Scatter plot between EPFL and UBC with respect to auto-stereo quality test

Pear. 0.96158 Spear. 0.95737



(c) Scatter plot between NTNU and Acreo with respect to stereo quality test

Figure F.1: Scatter plots of 3DV quality test among four laboratories.

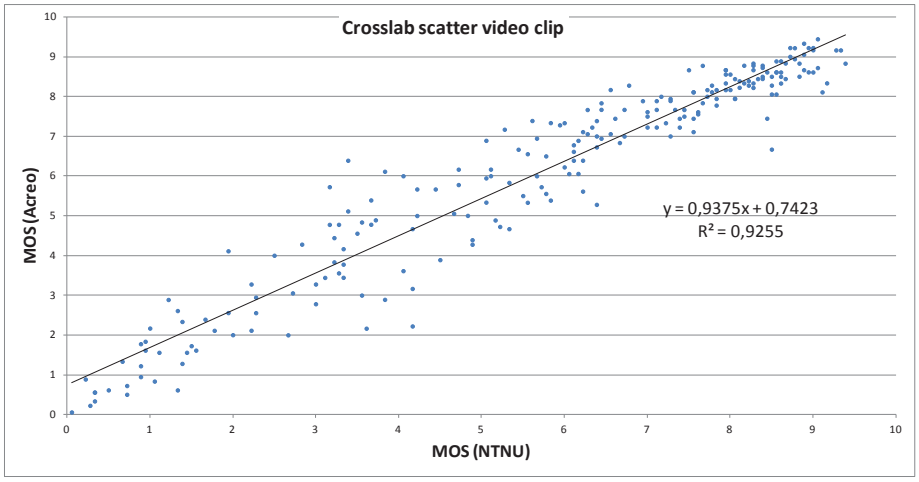


Figure F.2: Scatter plot between NTNU and Acreo with estimated regression line.

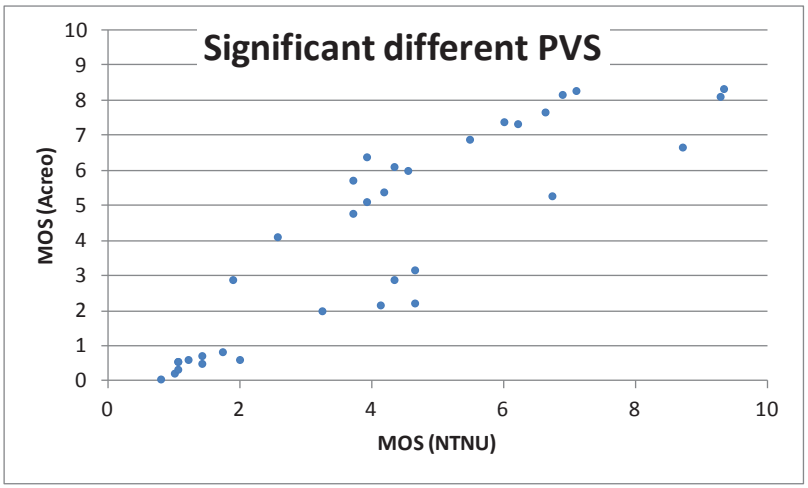
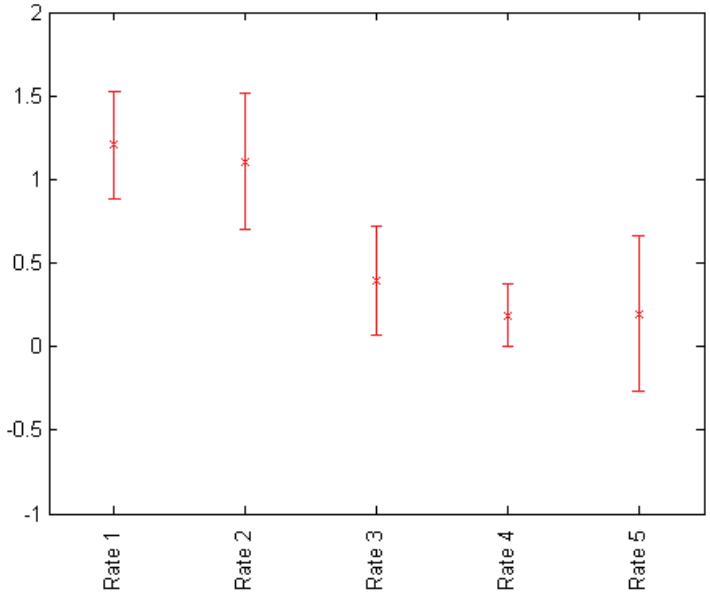
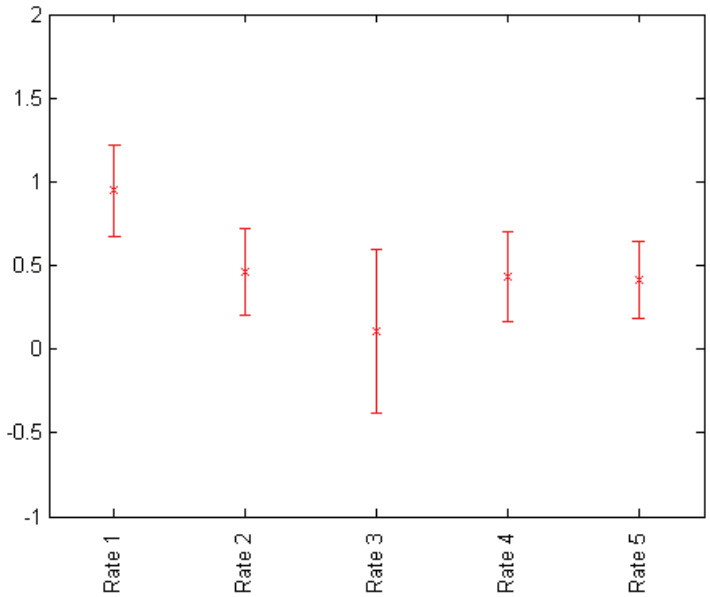


Figure F.3: Scatter plot between NTNU and Acreo of significantly different PVSs.



(a) Score difference between EPFL and UBC at different bit rates with respect to stereo test



(b) Score difference between EPFL and UBC at different bit rates with respect to autostereo test

Figure F.4: Score differences between EPFL and UBC at different bit rates.

References

[1] Certification. http://en.wikipedia.org/wiki/certification.

[2] T. A. A. for Laboratory Accreditation (A2LA). General requirements: Accreditation of ISO/IEC 17025 laboratories. 2011.

[3] ISO 17025 quality forms. http://www.17025.com/ quality_records.html. 2011.

[4] ISO/IEC 17025. 2005 working document. In *Perry Johnson Laboratory Accreditation, Inc.*, 2007.

[5] ISO/IEC JTC1/SC29/WG11 MPEG2011/N12036. Call for proposal on 3D video coding technology. In *Video and Requirement Group*, 2011.

[6] ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television pictures. 2002.

[7] ITU-R BT.710-4. Subjective assessment methods for image quality in high-definition television. 1998.

[8] F. Pereira and S. Buchinger. First thoughts on QUALINET certification. In *QUALINET*, 2011.

[9] Qualinet. Memorandum of understanding. 2010.