Visual comfort of 3-D TV : models and measurements

DOI:
10.6100/IR723145

Document status and date:
Published: 01/01/2012

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Visual Comfort of 3-D TV

Models and Measurements

Marc T.M. Lambooij

This PhD project was a collaborative project between the Human-Technology Interaction Group of the Eindhoven University of Technology and Visual Perception Group of Philips Research, the Netherlands.

Visual Comfort of 3-D TV - Models and Measurements
By Marcus Theodorus Maria Lambooij
Eindhoven: Eindhoven University of Technology, 2012. - Proefschrift. –

A catalogue record is available from the Eindhoven University of Technology Library.

ISBN: 978-90-386-3074-8
NUR 778
Keywords: Visual perception / Three-dimensional displays / Asthenopia

Cover design: Proefschriftmaken.nl || Printyourthesis.com
Printed by: Proefschriftmaken.nl || Printyourthesis.com
Published by: Uitgeverij BOXPress, Oisterwijk

Visual Comfort of 3-D TV

Models and Measurements

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op woensdag 18 januari 2012 om 16.00 uur

door

Marcus Theodorus Maria Lambooij

geboren te Eindhoven

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. I.E.J. Heynderickx
en
prof.dr. D.G. Bouwhuis

Copromotor:
dr.W.A. IJsselsteijn

"When I gained stereopsis, I felt like I was immersed in a medium more substantial than air, a medium on which tree branches, flower blossoms, and pine needles floated. I wondered if this sense of the air was what Monet spoke about in the following quote:

I want the unobtainable. Other artists paint a bridge, a house, a boat, and that's the end. They are finished. I want to paint the air which surrounds the bridge, the house, the boat, the beauty of the air in which these objects are located, and that is nothing short of impossible.

-Claude Monet

Or perhaps Eric Woznysmith, a strabismic, echoed Monet's thoughts when he described what it was like for him to see with stereopsis. Eric had studied drawing and learned that artists pay attention not just to the objects they will draw but also to "negative space", that is, the space, or the air, to the sides, in front of, and behind objects. When he gained stereovision, he told me he could see one hundred times more negative space."

A part from Susan Barry's Fixing my gaze (Barry, 2009)

-CONTENTS-

-List of Abbreviations-

AC	accommodation convergence
AC/A	accommodation convergence per accommodation change
ANOVA	analysis of variance
BI	base in
BIS	binocular status
BO	base out
CA/C	vergence accommodation change per vergence change
CBD	camera base distance
COA	continuous assessment
COAS	part of continuous assessment that corresponds to the single assessment
CISS	Convergence Insufficiency Symptom Survey
CRT	cathode ray tube
D	diopter
DEQ	Dry Eye Questionnaire
DOF	depth of field
EEG	electro encephalography
EndDisp	final maximum screen disparity
fMRI	functional magnetic resonance imaging
GBIS	good binocular status
HD	High Definition
HMD	head mounted display
IniDisp	initial maximum screen disparity
IPD	interpupillary distance
ISAF	Inter Stimulus Adaptation Field
JPEG	joint photographic experts group
LCD	liquid crystal display
LSE	least squares estimation
MAVOVA	Multivariate analysis of variance
MBIS	moderate binocular status
MEG	magneto encephalography
MEM	monocular estimation method
MPEG	motion picture experts group
PCA	principal component analysis
PD	prism diopter
RA	retrospective assessment
RGB	red green blue
SA	single assessment
SD	screen disparity
SSCQE	Single Stimulus Continuous Quality Evaluation
SSQE	Single Stimulus Quality Evaluation
STD	standard deviation
VDT	video display terminal
VFQ	Visual Function Questionnaire
WRRT	Wilkins Rate of Reading test

-PROLOGUE-

The most remarkable story I came across during my PhD is by far the story of Susan Barry, also known as Stereo Sue. The first 40 years of her life the world appeared to her as 'a child's drawing'; she had been cross-eyed since birth and perceived the world with one eye at a time while her eyes were rapidly and unconsciously alternating. Operations during childhood failed to provide her stereoscopic vision and as the years followed Susan was unaware of what she was missing. She remained ignorant until the age of 20, when a neurophysiology lecture astonished her that there is a new way of vision that she doesn't have. This claim triggered her to investigate her own binocular vision: she failed all stereo tests. She questioned if she would ever gain stereovision. Her doctor replied that

'*Stereopsis is just a little fine tuning for the visual system. You don't need stereo vision, because you don't have stereo vision*' (Barry, 2009; p. xiv)

where the logic of the latter remark I still cannot place. As a consequence, it took her 25 years before she visited a developmental optometrist. At this point she herself became an excellent brain scientist. The consequent optometric vision therapy changed her perception of the world: it provided her with a first a sense of stereoscopic depth.

"As I looked up to adjust the rear-view mirror, the mirror popped out at me, floating in front of the windshield. I was transfixed. Throughout the day, my stereovision would emerge- intermittently, fleetingly, unexpectedly- bringing me moments of absolute wonder and delight" (Barry, 2009; p. 94).

Barry quoted Frederick Brock in saying "It must be repeated here that, before stereopsis is actually experienced by the patient, there is nothing one can do or say which will adequately explain to him the actual sensation experienced" (Barry, 2009; p. 103).

"When I gained stereopsis, borders and edges around objects appeared much sharper and crisper than ever before. This effect was almost as dramatic as my new sense of space. An engineer would describe the world before my vision therapy as "low pass filtered" (Barry, 2009; p. 112). Stephanie Willen Brown, a woman who experienced something similar put this very well: "This clarity is everything, everywhere…. There are edges to everything!" (Barry, 2009; p. 113)

"Ordinary things looked extraordinary. Light fixtures floated and water faucets stuck way out into space. But it was "also a bit confusing"….. …. It is a bit like I am in a fun house or high on drugs. I keep staring at things…. The world really does look different."
(In The New Yorker, by Oliver Sacks, 2006; p. 96)

Susan Barry a.k.a. the 'post-chirurgical middle-aged alternating esotroop' (Barry, 2009)

-Chapter 1-

Introduction

The Miracle of the Age!!! A LION in your lap! A LOVER in your arms!

Tagline of Bwana Devil, the first feature length 3-D cinema movie (1952)

1.1. Rationale of this thesis

Three-dimensional (3-D) cinema seems to be widely accepted by the public; 3-D movie releases are rapidly succeeding each other, and their ticket sales exceed those of their 2-D counterparts (Mendiburu, 2009). Hollywood's embracing of 3-D movies as well as the development of fast LCD panels (needed for 3-D television (TV) systems) has spurred consumer electronics companies to start marketing 3-D displays for home use, bringing 3-D movies and games in the comfort of the living room. 3-D programs have already been broadcast in Japan in June 2008, and Sky Broadcast will launch a 3-D channel in the near future.

By introducing 3-D TV and its desktop-counterpart for gaming and internet applications on the public consumer market, viewers will be provided with a whole new experience. This next major step is not just a quality enhancement, but a fundamental change in the character of the image (IJsselsteijn, 2004). The difference between 3-D TV and its predecessor is the introduction of binocular disparity. The human eyes are horizontally separated, and therefore, have their own perspective of the world. They receive a slightly different retinal image, from which the brain extracts relative stereoscopic depth information. A 3-D TV set implements this concept by providing a different view to each eye, resulting in content that is rendered in depth and projected both in front of and behind the display.

The first stereoscopic display originated in 1838, and was location-multipled, in which the left and right views are separately generated and redirected to the appropriate eye through separate channels (Wheatstone, 1838). Nowadays, there are various imaging technologies to realize binocular disparity on TV and generally four distinguishing features are used to characterize stereoscopic TV technologies: 1) whether the technology is stereoscopic, where the viewer wears glasses to direct the left and right images to the appropriate eye, or autostereoscopic, where the technique to separate both views is integrated in the display, 2) the method applied to separate the left- and right-eye views; 3) whether it supports motion parallax, i.e., the apparent displacement of an object caused by a change in viewing position, and 4) the number of viewers that can watch a stereoscopic sequence simultaneously.

Stereoscopic technologies can be anaglyph or colour-multiplexing (left and right views are filtered with near-complementary colours), polarization-multiplexing (left and right views are separated with polarized light) and temporal multiplexing (left and right views are occluded alternately in rapid succession using shutter glasses in sync with the left-right alternating image information on the screen). These glasses-based technologies are relatively cheap and easy to construct, yet the anaglyphs suffer from poor colour rendering, the polarization filters reduce the brightness and the shutter glasses induce flicker.

Autostereoscopic display technologies create viewing zones by use of a parallax barrier (an occlusion mask with slits placed over a display that permits specific parts of the image to be visible for each eye in a viewing zone), a lenticular sheet (a array of cylindrical lenses placed over a display that directs light from alternate pixel columns to a viewing zone) (Dodgson, 2005), or head-tracking combined with steerable viewing zones (Surman, Sexton, Bates, Lee, Craven and Yow, 2003). In case of multi-view autostereoscopic display, a discrete set of perspective views of a scene (more than two) is created and distributed across the viewing field in viewing zones. This provides a certain viewing freedom and introduces motion parallax or so-called look-around capabilities, yet this is accompanied by a great loss of spatial resolution. More advanced autostereoscopic technologies are volumetric displays (Favalora, 2005), where the projection is done inside a three-dimensional physical space, rather than on a planar display, and holographic displays, in which a wavefront is reconstructed that is identical to the original scene (Sexton and Surman, 1999). Hence, light is generated at the exact depth where the image information is situated and not via a projection of the light on the retina, as in the other systems. Since the image is transparent, however, this technology is not suitable for 3-D movies.

Central to these technologies should be the viewer's experiences which will determine the success or failure of proposed innovative imaging technology. The recent introduction of 3-D TV to the general audience necessitates a large scale and longitudinal research effort on the potential adverse effects of 3-D viewing on a wide variety of viewers, including people with known visual problems, and children with developing visual systems. Two decades ago a case study reported changes within the binocular visual system of a four-year old child as a result of viewing a stereoscopic movie that required surgical correction (Tsukuda and Murai, 1988). A discussion arose that was similar to the discussion concerning visual discomfort from video display terminals ten years earlier (Scheiman, 1996) and similar concerns accompanied the introduction of helmet mounted displays (HMD) two decades ago. Mon-Williams, Wann and Rushton (1993) reported ocular changes resulting from short-term use of HMD's:

> *The possibility of producing short term ocular symptoms is unsatisfactory, but of far greater concern is the scenario where a virtual reality user gets into a motor vehicle following immersion in a virtual world and attempts to drive with unstable binocular vision and a decrease in visual acuity. Another worrying situation is that of a child or adult with already unstable binocular fusion using a poorly configured virtual reality display and suffering permanent breakdown of binocular function with the possibility of resultant strabismus and diplopia.*

This potential visual safety scenarios, caused worries and disturbance in the media (Wann and Mon-Williams, 1997) and did not contribute to the acceptance of HMDs. Consequently, HMDs lacked a proper market introduction since they suffered from "scaremongering" in the popular press. It was only years later that research revealed that stereoscopic viewing with HMD caused no harmful effects either in adults (Peli, 1998) or in children (Kozulin, Ames and McBrien, 2009). Hence, research findings that hint at potential visual health problems should be presented with caution because they nourish the negative impact of social opinion on 3-D TV. Since some studies reported on visual safety problems that specifically are of concern for children (Tsukuda and Murai, 1988), The Japan Broadcasting Corporation recently decided to suspend all stereo television transmissions as long as there is uncertainty about any long-term negative consequences for the development of children's visual system at large (van Nes, 2009). According to van Nes (2009) this was a drastic measure, but one that certainly seems to be justifiable on ethical grounds.

For a proper introduction into the home consumer market it is thus essential that such innovative display technologies are evaluated in terms of their benefits and drawbacks from a consumer point of view. The compromise between the added value of stereoscopic depth and potential visual discomfort will influence the perception of the image, yet it is unknown how these factors affect the overall percept, and how to best assess this. This thesis describes how to evaluate the added value of stereoscopic depth as well as the potentially accompanying visual discomfort.

1.2. 3-D evaluation

Most perceptual evaluations of 3-D TV systems are performed with evaluation metrics based on 2-D image quality models. Assessment of perceived image quality is one of the standard methods to subjectively evaluate the performance of a 2-D imaging system. The relationship between the perceived image quality and the technology parameters of the imaging system is of main interest for display manufacturers in general. However, it is often very time consuming and inefficient to assess this relationship directly. To model this relationship Engeldrum has developed the Image Quality Circle (Engeldrum, 1999), depicted in Figure 1. In this model, perceived image quality is regarded as a multi-dimensional psychological construct that reflects several image attributes such as sharpness, contrast, colour, and various artifacts. The Image Quality Circle links the perceptual impact of these attributes and artifacts to technology parameters such as pixel pitch, thickness of colour filters, etc. via the physical characteristics of the luminance output.

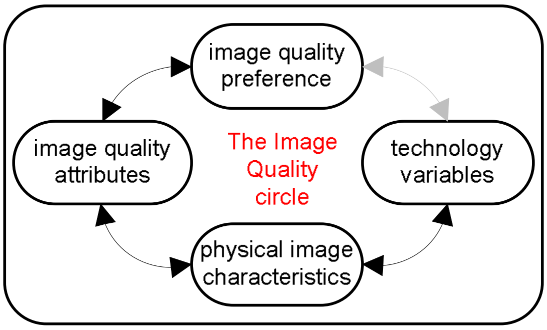


Figure 1. Engeldrum's Image Quality circle that describes the relationship between the perceived image quality and the technology parameters of the imaging system.

Since similar technology variables are important in 3-D TV, the framework of the Image Quality Circle seems also useful for the evaluation of such innovative imaging systems. Perceived image quality proved a reliable construct to evaluate 2-D content (Engeldrum, 2004), yet recent enrichments such as 3-D appear to go beyond the concept of image quality (Benoit, Callet, Campisi and Cousseau, 2008; Häkkinen, Kawai, Takatalo, Leisti, et al., 2008; IJsselsteijn, Bouwhuis, Freeman and de Ridder, 2002; Meesters, IJsselsteijn and Seuntiëns, 2004; Seuntiëns, Meesters and IJsselsteijn, 2006; Seuntiëns, Heynderickx and IJsselsteijn, 2008; Tam, Stelmach and Corriveau, 1998). That is, the sensation of stereoscopic depth, which is clearly present in stereoscopic image material, is not accounted for by the model, and image quality cannot be the only adequate description of relevant image properties in user perception. Tam et al. (1998) evaluated the psychovisual impact of stereoscopic images on viewers in terms of perceived image quality, perceived sharpness and perceived depth. The images were degraded with MPEG-compression (blocking artifacts). They revealed a high correlation between perceived image quality and perceived sharpness, and a low correlation between perceived image quality and perceived depth. They concluded that stereoscopic depth in natural scenes does not improve image quality compared with the same 2-D-images. Seuntiëns et al. (2006) found that perceived image quality followed variations in JPEG-compression, but remained unaffected by various levels of stereoscopic depth.

These results suggests that the term image quality is not sensitive to changes in stereoscopic depth in the image, which makes the Image Quality Circle incomplete to evaluate the added value of stereoscopic depth. However, to improve existing 3-D display systems, or to develop new ones, information is also required about how the depth quality preference is affected by specific technology variables of the display system. Technology variables can be display related, e.g., type of imaging system, lenticular thickness, material of polarisation filters and pixel size. Technology variables can also be signal related, e.g., formats and settings in 2-D-to-3-D conversion algorithms. Establishing the relationship between perceived depth and technology parameters is often very time consuming and inefficient, since many technology variables are dependent on each other. To overcome this problem, the relationship between perceived depth and technology parameters can be modelled along the dark arrows in Figure 2, which describe a similar model as the Image Quality Circle.

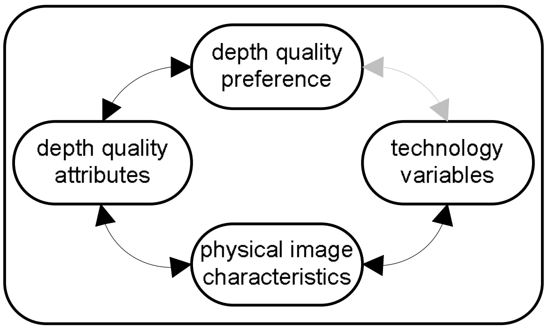


Figure 2. A model that describes the relationship between the depth quality preference and the technology parameters of the imaging system.

In Figure 2, perceived depth is regarded as a multi-dimensional psychological construct that reflects several attributes such as depth sharpness and depth contrast. Perceived depth also reflects depth artifacts such as frame violation, i.e., the conflict between occlusion of the display frame and screen disparity information (Ware, Gobrecht and Paton, 1998), crosstalk, i.e., the mixing of the left- and right-eye images that is perceived as blur or ghosting (van Berkel and Clarke, 1997), and unnatural viewing conditions, e.g., suppressed depth range or a limited depth of field) (IJsselsteijn, 2004; Meesters et al., 2004; Peli, 1999). Each of these depth attributes is related to one or more physical characteristics of the image, such as crosstalk, screen disparity, spectral luminance distribution and reflectance of polarisation filters. The relationship between physical characteristics of the image and the technology parameters is usually specified by technicians for a specific display system. This thesis describes the relationship between specific technology variables and perceived depth along the dark arrows in Figure 2, and more importantly, how these technology parameters affect the balance between perceived depth and perceived image quality.

When evaluating 3-D display systems, an important aspect is visual discomfort. It is well-known that technical choices or compromises made throughout the entire chain of stereoscopic image generation, transmission, rendering and display, can affect the overall visual comfort experienced by the viewer (Patterson, 2009). The technology parameters that impact visual comfort are similar to those that impact perceived image quality and perceived depth. This relationship between the technology variables and visual comfort therefore, can be obtained along the dark arrows in Figure 3.

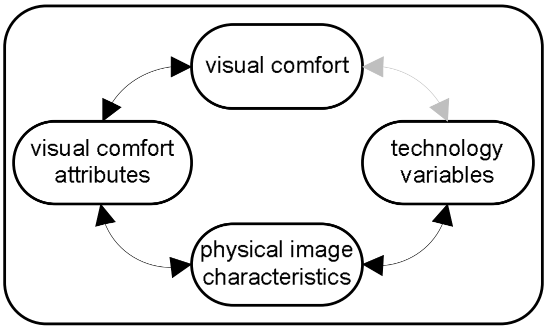


Figure 3. A model that describes the relationship between visual comfort and the technology parameters of the imaging system.

This thesis describes the various causes and aspects of visual comfort from technological and perceptual points of view in order to model visual comfort along the dark arrows in Figure 3. In summary, 3-D TV goes beyond optimizing image quality by displaying stereoscopic depth. In most cases this can only be achieved at the expense of spatial and/or temporal resolution (i.e., at the expense of important image quality aspects), whereas in addition some people can experience visual discomfort when watching 3-D content. Hence, the total visual experience of a 3-D display is expected to be a balance between image quality, any added value of having stereoscopic depth and the possible annoyance of visual discomfort. This thesis is aimed at describing this total viewing experience by combining the Image Quality Circle with the models for perceived depth and visual comfort as depicted in Figure 4.

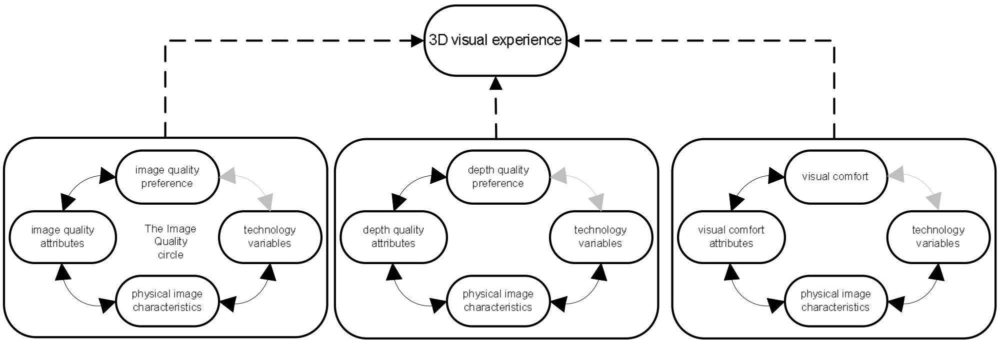


Figure 4. 3-D Visual Experience Model, which describes the overall 3-D visual experience as a combination of image quality, stereoscopic depth and visual comfort.

1.3. Overview

The aim of this thesis is to understand, measure and eventually, model and predict the added value of stereoscopic depth on our experience and the accompanying visual discomfort for a 3-D TV. This aim can be divided into five sub aims:

1. Define the balance between image quality and perceived depth in general, and for certain specific stereoscopic display technologies.

2. Describe the concept "visual discomfort" in order to understand the underlying perceptual mechanisms as well as its operationalization in display evaluation (and identify potential areas that could benefit from further study).
3. Determine the effect of relevant image characteristics on visual comfort.
4. Establish a valid and reliable protocol for the measurement of visual discomfort via objective and subjective indicators.
5. Model the overall visual experience reflecting image quality, perceived depth and visual comfort aspects.

Chapter 2 describes a 3-D Quality Model that reflects the balance between image quality and perceived depth. In three experiments, higher level evaluation metrics (naturalness and viewing experience) are proposed that are sensitive to both image quality and stereoscopic depth. In the 3-D Quality Model such higher level evaluation metrics are expressed as a weighted sum of image quality and perceived depth.

Chapter 3 presents an extensive overview of state-of-the-art viewing discomfort. Classical factors, such as conflicts between accommodation and convergence and excessive binocular parallax as well as some additional causes that might have become more relevant nowadays with the evolution in 3-D systems, are critically reviewed from a human perception perspective and a technological perspective. Experimental settings and potential evaluation methods necessary to qualify or quantify the degree of visual comfort in an unambiguous manner are also discussed.

In Chapter 4 visual discomfort associated with stereoscopic displays is related to video characteristics, which can induce visual discomfort such as motion and changes in screen disparity. In Experiment 4 the 3-D movie comprised relatively simplistic 3-D content. In Experiment 5, stereoscopic content with higher spatial and temporal complexity was used.

In Chapter 5 a measurement protocol for the evaluation of the visual discomfort associated with stereoscopic displays is established in three experiments. In Experiment 6 different clinical optometric evaluation methods are compared. The results showed that amongst people with normal vision, some are more susceptible to visual discomfort based on poorer binocular functioning. This finding is further verified in Experiment 7, which resulted in a test to categorize people based on their binocular functioning. In Experiment 8 the outcome of this test is related to thresholds in screen disparity for comfort.

Chapter 6 presents a summary and discussion of the most important findings. The quality, depth and comfort circles introduced in Chapter 1 will be revisited and integrated into the overall 3-D visual experience model. In addition, the practical applicability of our main findings is discussed for consumers, display manufacturers, movie producers, program makers, and eyecare practitioners.

-Chapter 2-

Beyond 2-D quality

"Stereoscopic viewing was indeed fashionable. As if by magic the world was available for all to see, as entertainment, as education, in startling realism in the comfort of the home."

Portrayal of the enthusiasm around 1855 (Sammons, 1934; p. 9)

ABSTRACT

Perceived image quality is a standard evaluation metric for 2-D imaging systems. When applied to stereoscopic 3-D imaging systems, however, it does not incorporate any added value of stereoscopic depth. Higher level evaluation metrics (naturalness and viewing experience) are proposed that are sensitive to both image quality and stereoscopic depth. A 3-D Quality Model is constructed in which such higher level evaluation metrics are expressed as a weighted sum of image quality and perceived depth. This model is validated by means of three experiments, in which stereoscopic depth (camera base distances and screen disparity) and image quality (white Gaussian noise and Gaussian blur) are varied. The resulting stimuli are evaluated in terms of naturalness, viewing experience, image quality and depth percept.

Analysis revealed that viewing experience and naturalness incorporated variations in image quality to a similar extent, yet the added value of stereoscopic depth is incorporated significantly more by naturalness. This result classifies naturalness as the most appropriate evaluation metric to evaluate 3-D quality of stereoscopic stills. The 3-D Quality Model based on naturalness as evaluation metric is validly applicable to stereoscopic stills and its score is determined approximately 75% by image quality and approximately 25% by the added value of stereoscopic depth.

This chapter is based on:

Lambooij, M., IJsselsteijn, W., Bouwhuis, D., and Heynderick, I. (2010). Evaluation of stereoscopic stills: Beyond 2-D quality. *IEEE Transactions on Broadcasting*. 57: 432-444.

2.1. Introduction

In theory a 3-D display simply adds depth to a conventional 2-D display by means of stereoscopy. In practice, 3-D displays suffer from quality degradation relative to an equivalent 2-D display because the introduction of stereoscopic 3-D usually comes at a cost. Two-dimensional quality can be compromised due to loss of temporal or spatial resolution. On the other hand, there is an enrichment in spatial rendering as a consequence of the appearance of stereoscopic depth. Understanding the balance between these aspects is particularly relevant for the optimization of (auto)stereoscopic displays from the consumer's point of view. Hence, it is essential that stereoscopic images are not just evaluated in terms of perceived image quality, but rather in terms of evaluation criteria that reflect the full extent of the user experience. In the recent past, it has been suggested that the concepts presence, naturalness and viewing experience can be suitable candidates that more adequately describe the added value of 3-D in case of stereoscopic TV (IJsselsteijn et al., 2002; Meesters et al., 2004; Seuntiëns et al., 2006; Seuntiëns et al., 2008). The current chapter aims to take this work a step further by formulating a 3-D Quality Model that incorporates aspects of image quality based on the Image Quality Circle of Engeldrum (Engeldrum, 1999) as well as the added value of stereoscopic depth.

2.2. Background

Since image quality does not include the perception of stereoscopic depth as stated in Chapter 1, other evaluation metrics are needed in order to fully describe and measure the entire user experience. Freeman and Avons (2000) used focus groups to explore viewers' reactions and sensations to 3-D TV. Their study revealed that non-expert viewers described sensations of 'presence' or of 'being there' while viewing stereoscopic image material, which led them to relate the concept of presence to involvement, realism and naturalness. These results were in line with other research that showed correlations between presence and depth, and presence and naturalness (IJsselsteijn, de Ridder, Hamberg, Bouwhuis and Freeman, 1998a). Based among others on these results, Lessiter, Freeman, Keogh and Davidoff (2001) developed a reliable and cross-media applicable presence questionnaire. Their questionnaire reflects four underlying dimensions: a sense of physical space, engagement, naturalness and negative effects.

Naturalness, defined as a realistic and truthful reproduction of reality, was originally introduced to determine the perceived quality of colour reproductions, because images of high quality were assumed to be perceived as 'natural' (Yendrikhovskij, 1998). Despite small shifts between measures of perceived image quality and naturalness in both optimal chromaticity for monoscopic images (de Ridder, Blommaert and Fedorovskaya, 1995; de Ridder, 1996) and optimal depth for stereoscopic images (IJsselsteijn, de Ridder and Hamberg, 1998b), their evaluations showed high similarities. Recently, Seuntiëns et al. (2008) applied naturalness as an evaluation concept for both 2-D and 3-D stills, that were degraded in image quality by adding Gaussian white noise. This revealed that naturalness reflected variations in image quality as well as the added value of stereoscopic depth. These results are depicted in Figure 5. Hence, naturalness is believed to have a depth component as well as a quality component.

Since it is generally expected that stereoscopic displays enhance viewers' viewing experience, Seuntiëns et al. (2008) also applied this variable as an evaluation concept. The concept of viewing experience is defined as the users' perceptual and cognitive experience of the entire application, which to our knowledge has not been used as an evaluation concept before. The results, depicted in Figure 6, were quite similar to those of naturalness. Viewing experience also followed the variations in image quality, yet incorporated the added value of stereoscopic depth though to a somewhat lesser extent than naturalness. More specifically, for both naturalness and viewing experience the slopes of the lines representing 2-D and 3-D content are similar, yet the offset between them is larger for naturalness. In a more recent study interviews were used to explore experiences produced by stereoscopic content (Häkkinen et al., 2008). The results confirmed that viewers' experiences could be related to three underlying experience dimensions: presence, life-like vs. artificial and depth impression. These dimensions show high resemblance with the concepts naturalness, viewing experience and presence.

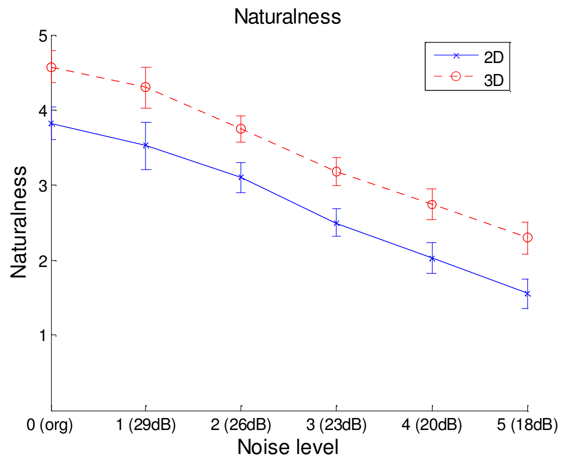


Figure 5. Naturalness ratings with their 95% confidence intervals as a function of the level of noise (expressed in dB) in a 2-D and 3-D display (copied from Seuntiëns et al., 2008).

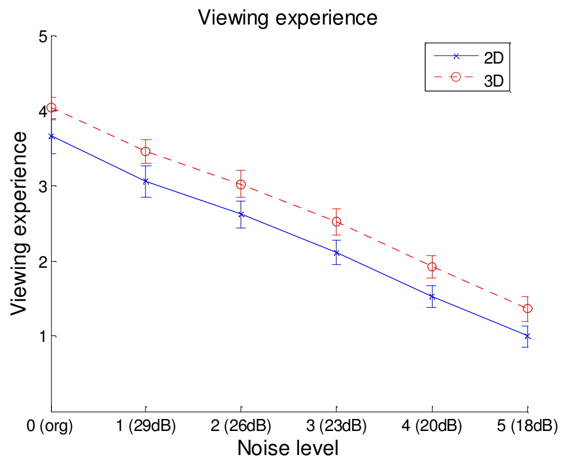


Figure 6. Viewing experience ratings with their 95% confidence intervals as a function of the level of noise (expressed in dB) in a 2-D and 3-D display (copied from Seuntiëns et al., 2008).

2.3. Towards a 3-D Quality Model

Perceived image quality does not incorporate the added value of depth in stereoscopic images. In the absence of 3-D artifacts and with a proper left and right image separation, image quality and perceived depth can be assessed independently (Seuntiëns et al., 2006; Seuntiëns et al., 2008; Tam et al., 1998). This implies that varying one should not influence the perception of the other. In order to evaluate 3-D quality, higher level concepts such as presence, naturalness and viewing experience can be applied that are expected to reflect both image quality and depth aspects. Based on these considerations, the 3-D Quality Model, as depicted in Figure 7, is constructed. It can be represented in mathematical terms by Equation 1, in which a higher level evaluation metric (E.M.) is factored in terms of perceived image quality (α·IQ) and perceived depth (β·D).

$$E.M. = \alpha \cdot IQ + \beta \cdot D \qquad (1)$$

with *E.M.* = evaluation metric
IQ = perceived image quality
D = perceived depth
α, β = weights of image quality and depth respectively

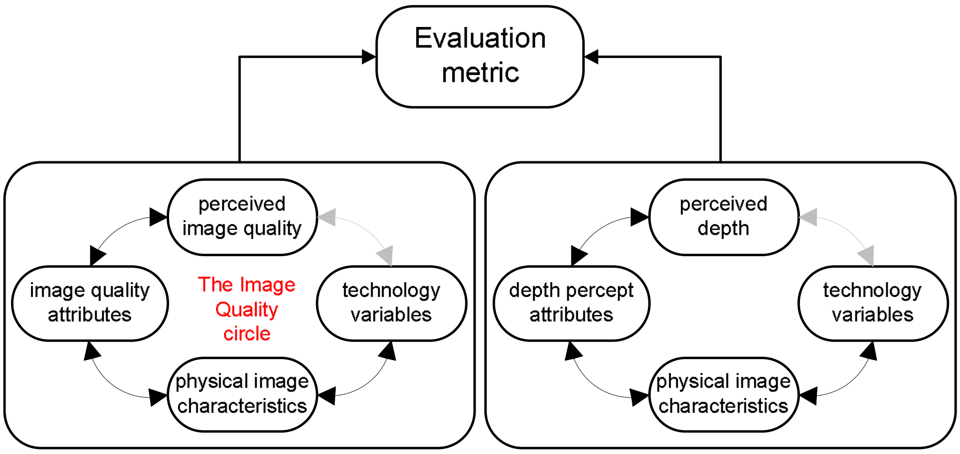


Figure 7. The 3-D Quality Model. A higher level evaluation metric to assess 3-D quality is factored in a weighted sum of perceived image quality, represented by Engeldrum's Image Quality Circle (Engeldrum, 1999), and perceived depth.

The next step is to validate the model and to determine the weights α and β. In order to test the model in clear experimental conditions without any side-effects such as 2-D and 3-D artifacts, the use of a stereoscopic imaging system that has perfect image separation, e.g., a Wheatstone based stereo viewer, is a prerequisite. For our current purposes, however, it is also important to investigate 3-D displays that are most likely to be used in consumers' applications, such as a multi-view autostereoscopic display. For the majority of such displays stereoscopic depth can only be rendered at the costs of image quality and 3-D artifacts (e.g. crosstalk) (Berkel and Clarke, 1997). Additionally, real-world content is expected to be converted from 2-D to 3-D, especially with the upcoming digital TV in future applications (Redert, Berretty, Varekamp, van Geest and Bruijns, 2007), which is a process that is not entirely artifact free. Such 2-D-to-3-D conversion algorithms generate a 2.5D representation of video material as depicted in Figure 8: in its basic form a representation of a video as a conventional video stream enhanced with per-pixel depth information (RGB + depth).

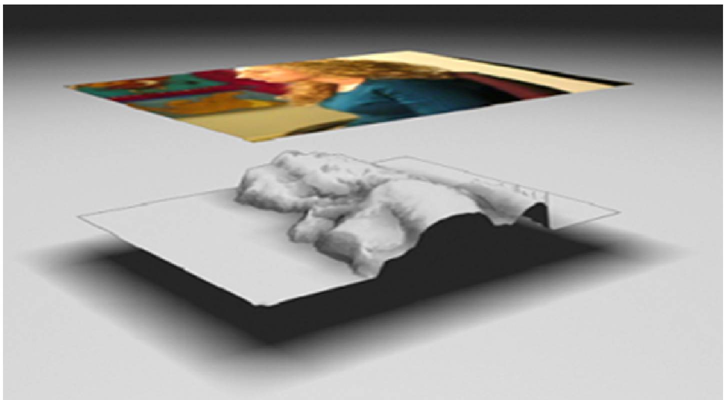


Figure 8. Illustration of a conventional video stream layer and a 3-D depth information layer.

This 2.5D format is highly compatible with conventional video streams and since the amount of 2-D content is practically infinite, so will be the amount of 3-D material. It is a quasi depth ordering process that relies on assumptions, estimations and heuristic approximations of a scene's depth structure. It can result in 2-D and 3-D artifacts that can affect the viewing experience negatively (see Chapter 3). Including such suboptimal means of rendering and displaying 3-D images provides a closer approximation of the likely real-life viewing situation, thereby increasing the ecological validity of our approach. Hence, modelling the actual experience of 2-D-to-3-D converted content on a multi-view autostereoscopic display is also highly relevant for this purpose.

Research revealed that assessment of perceived depth mainly reflects the variation of stereoscopic depth, i.e., added value, and that the perception of accompanying 2-D and 3-D artifacts is mainly incorporated into the image quality assessment (Strohmeier and Tech, 2010; Seuntiëns, Meesters, IJsselsteijn, 2005). Strohmeier, Jumisko-Pyykkö, and Kunze (2010) combined perceptual evaluation and qualitative attribute elicitation to get a better understanding of 3-D quality. Though 3-D content increased the depth impression, it also decreased the overall satisfaction as a result of, among others, blur and unstable quality. This result indicates that the added value induced by the depth perception in stereoscopic presentation is only valid when the level of visible artifacts is low. Consequently, the variation in stereoscopic depth can also affect the perceived image quality; i.e., accompanying artifacts such as a reduction in sharpness resulting from crosstalk or spatial artifacts resulting from 2-D-to-3-D conversion affect image quality negatively. Vice versa, variation in image quality can also affect perceived depth, i.e., reduction of sharpness and contrast reduces the ability to distinguish objects and as such affects the depth percept.

It is important to emphasize that any 2-D and 3-D artifacts introduced by generating, rendering and displaying 3-D content in this sense do not constrain the model. Of course, the artifacts will affect the perceptual impact of the stimuli, which can have an effect on both perceived image quality and perceived depth. In other words, with respect to the 3-D

Quality model it can be stated that certain attributes impact on both the perceived image quality and perceived depth. This can be accounted for (see paragraph 2.4.8. for more details), which still allows the perceptual impacts of depth and image quality to be modelled independently. For our purposes, it is less relevant how image quality or perceived depth are affected by a specific 2-D or 3-D variation. More precisely, the 3-D Quality Model is aimed at describing how image quality and perceived depth are affected with respect to each other, i.e., in a *relative* sense.

Hence, the main objective of this chapter is to model the behaviour of image quality and perceived depth relative to each other, i.e., to determine the weights α and β. This is accomplished via three main experiments. Stimuli that varied in stereoscopic depth and image quality were assessed in terms of perceived image quality and perceived depth, as well as in terms of two higher level evaluation metrics, namely naturalness and viewing experience. Presence was not applied as an evaluation metric, because a pilot study revealed that in line with previous research (IJsselsteijn, de Ridder, Avons and Bouwhuis, 2001), presence is more appropriate for evaluating moving images than stills. In Experiment 1 the model is validated with a stereoscopic imaging system with perfect image separation: a Wheatstone based stereo viewer. In Experiment 2 a more realistic TV application is used: a 42” Philips nine-view auto-stereoscopic lenticular display. In Experiment 3 the coverage of imaging technology is further increased by modelling the actual experience of 2-D-to-3-D converted content on a 20” Philips nine-view autostereoscopic lenticular display.

2.4. Experiment 1

This experiment aims to provide a first validation of the 3-D quality model with clear experimental conditions without any unwanted side-effects such as 2-D and 3-D artifacts.

2.4.1. Experimental set-up

Forty-four participants were divided into four groups that each assessed two separate sets of stimuli in terms of one of four evaluation metrics: image quality, perceived depth, naturalness and viewing experience. Stereoscopic depth was varied with three camera base distances (CBD) and image quality was varied with four levels of white Gaussian noise (set 1) or Gaussian blur (set 2). All stimuli were presented twice. Hence, the experiment was a 2 x 2 x 2 x 3 x 4 (repetition x set x image x CBD x noise/blur) mixed-subject design resulting in 96 conditions per evaluation metric. The stimuli were randomised per evaluation metric and the evaluation metrics were assigned randomly to the participants in four different sessions. The viewing distance was 0.40 meters.

Participants

Twelve females and thirty-two males participated. Most of the participants were aged between 18 and 30 years, and two male participants were 32 and 58 years old. All participants had a normal or corrected to normal visual acuity of > 1 (as tested with the Landolt-C test) and a stereo acuity of < 30 seconds of arc (as tested with the Randot stereo test).

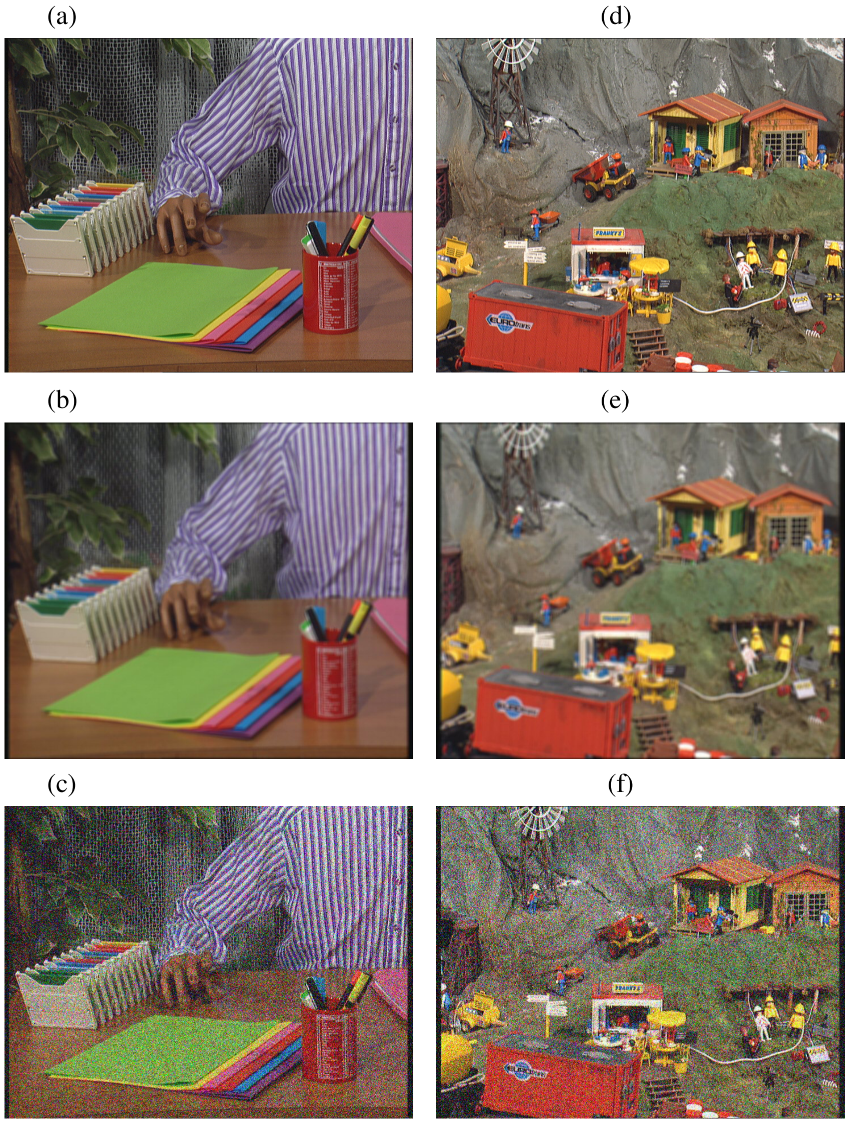


Figure 9. The original images used in Experiment 1, Bureau and Playmobiles, are shown in panel (a) and (d) respectively, and the degraded images with blur in panel (b) and (e) and with noise in panel (c) and (f).

Stimuli

The image material consisted of two still images, Bureau and Playmobiles, that varied in CBD and noise or blur. The original images as well as the degraded images are depicted in Figure 9. Both original images have been used by Seuntiëns et al. (2005) as well.

Each image had a resolution of 720 by 576 pixels. For the variation in stereoscopic depth a professional stereoscopic studio camera in a toed-in configuration was used (i.e., an arrangement with rotation of the cameras so the camera axes verge at a single point) with a convergence distance of the cameras of 1.30 m, and three CBD's namely 0 mm (i.e. monoscopic), 40 mm and 80 mm.

For the variation in image quality white Gaussian noise and Gaussian blur were used. Four levels of white Gaussian noise with variances (σ^2) of 0, 0.00125, 0.005 and 0.01 and four levels of Gaussian blur with standard deviations (σ) of 0, 1, 1.5 and 5 (both expressed in terms of pixel units) were added to the original images using Matlab®. Figure 9 depicts the degraded images with the highest levels of degration.

Equipment

To display the stimuli a Screenscope (mirror stereoscope) was used to direct the left- and right-eye image of a side-by-side stereo pair displayed on a single monitor to the appropriate eye. The Screenscope was attached to the computer screen as shown in Figure 10A. This system is location multiplexed, thus containing zero crosstalk. A high resolution monitor with a size of 17 inch, a resolution of 1600 by 1200 pixels and a viewing distance of 0.40 meters was used as specified in Figure 10B. The principle of the Screenscope viewer (Figure 10B) is based on the Wheatstone stereoscope. The only difference between these systems is that a Wheatstone stereoscope uses two mirrors and the Screenscope uses four mirrors. In a traditional Wheatstone set-up, the stereograms must be produced as mirror images on the monitor, which is not the case with the Screenscope due to the extra set of mirrors.

A B

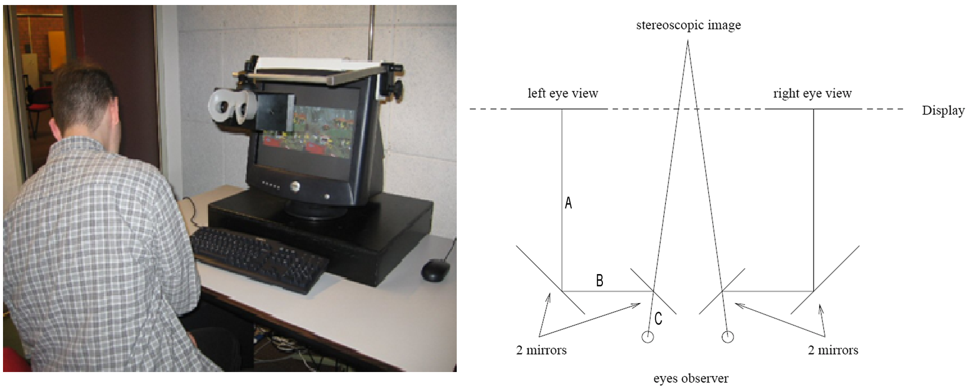


Figure 10. Picture (A) and principle of Screenscope (B), the stereoscopic system used in Experiment 1 to direct the left- and right-eye image of a side-by-side displayed stereo pair to the appropriate eye. The viewing distance (A+B+C) was 0.40 meters.

Procedure

Participants were seated at a viewing distance of 0.40 meters and received a brief instruction concerning the course of the training and the experiment. Participants were asked to rate the evaluation criteria on a 5-point scale labelled with the adjective terms [bad]-[poor]-[fair]-[good]-[excellent] according to the ITU recommendations for methodology for the subjective assessment of the quality of television pictures (ITU-R, 2002). Participants were asked to use the full range of the scale (Jones and McManus, 1986). Any questions concerning the procedure of the experiment were answered.

In order to familiarize participants with the assessment method and the stimuli, a brief training session was held in which six stimuli were presented including the extremes of the CBD and noise or blur levels used in the experiment. Both in the training session and in the actual experiment the stimuli remained on the screen until participants completed their rating in terms of one of the four criteria. Between subsequent stimuli an ISAF screen (Inter Stimulus Adaptation Field) was displayed for three seconds. On average, the experiment took around 35 minutes. For the assessment task Table 1 presents a description of the evaluation metrics that was provided to the participants. Based on this description participants could differentiate between the evaluation metrics, without suppressing their own interpretation or revealing underlying objectives of the experiment. The training session allowed the participants to anchor the evaluation metrics in that they could construct their own internal range. This appeared to reduce variation between participants in the current experiment as well as in previous research (Tam et al., 1998; Seuntiëns et al., 2008; de Ridder et al., 1995; de Ridder, 1996; IJsselsteijn et al., 1998b).

TABLE 1
EXPLANATION OF THE EVALUATION METRICS FOR THE PARTICIPANTS

Evaluation metric		Explanation for participants
image quality	→	excellence of the image
naturalness	→	realistic or truthful reproduction of reality
depth percept	→	amount of depth
viewing experience	→	total experience related to the display

Statistical analyses

The scale labelled with the adjective terms was transformed to a numerical one in such a way that the adjective [bad] corresponded to a rating of 1 and the adjective [excellent] to a rating of 5. To verify whether the ordinal categorical scale was a parametric one, i.e., with perceived equal distances between the adjectives, Thurstone's law of categorical judgement (Engeldrum, 2000) was applied to the data. The raw data were transformed with the software program ThurcatD (Boschman, 2000) to a Thurstone scale. The resulting data indicated that the perceived intervals between the adjectives were perceived as equal for all four evaluation criteria (maximum likelyhood indicates no difference at $p < 0.05$ level), which allowed us to use the raw data for further ANOVA analysis. Per evaluation metric an ANOVA was performed with level of noise or blur, CBD and image content as independent variables and score of the assessment as dependent variable and all two-way interactions were included. For the calculation of the effect size of the main effects and their interactions, the partial eta squared method (η^2) was applied

2.4.2. Results

Figure 11 depicts the average assessment scores averaged over content and participants with their error bars (representing the 95% confidence intervals) per evaluation metric as a function of noise level (increasing along the x-axes) and CBD (as parameter).

Figure 11 shows that naturalness ($F(3,252) = 177.73$, $p < .001$, $\eta^2 = .68$), viewing experience ($F(3,252) = 257.44$, $p < .001$, $\eta^2 = .75$) and image quality ($F(3,252) = 278.31$, $p < .001$, $\eta^2 = .77$) all are similarly affected by the introduced noise. They reveal similar slopes as a function of noise level, whereas noise has less effect on perceived depth ($F(3,252) = 12.04$, $p < .001$, $\eta^2 = .13$). Another aspect that is noteworthy is that naturalness ($F(2,252) = 22.50$, $p < .001$, $\eta^2 = .15$) and viewing experience ($F(2,252) = 16.57$, $p < .001$,

$\eta^2 = .12$) both show lower scores for the CBD of 0 meter than for the two larger CBDs, which is also true for image quality, though to a lesser extent ($F(2,252) = 3.87$, $p < .05$, $\eta^2 = .03$). The depth percept scores show the largest differences between the CBD ($F(2,252) = 231.78$, $p < .001$, $\eta^2 = .65$).

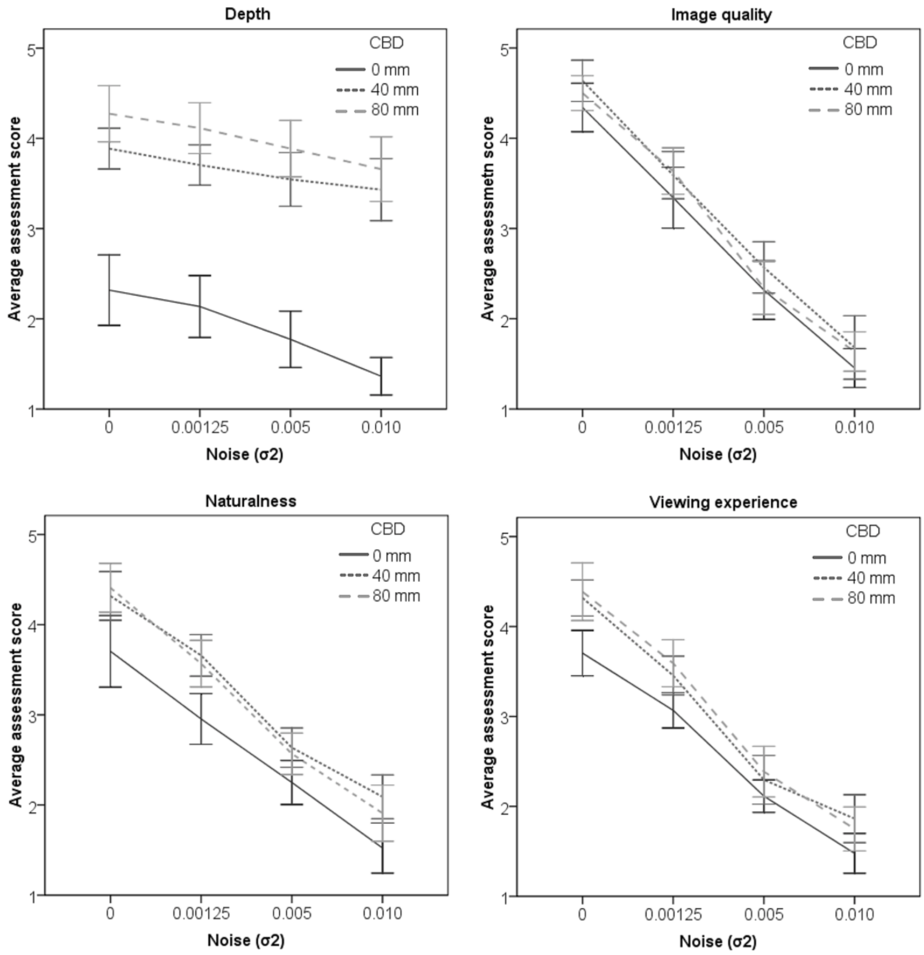


Figure 11. Mean assessment scores with their 95% confidence intervals of the evaluation metrics perceived depth, image quality, naturalness and viewing experience. The x-axes represent the variation in noise, the y-axes represent the averaged scores and the different lines represent the different CBD.

Figure 12 depicts the average assessment scores averaged over content and participants with their error bars (representing the 95% confidence intervals) per evaluation metric as a function of blur level (increasing along the x-axes) and CBD (as parameter).

Figure 12 shows that naturalness ($F(3,252) = 205.66$, $p < .001$, $\eta^2 = .71$), viewing experience ($F(3,252) = 387.71$, $p < .001$, $\eta^2 = .82$) and image quality ($F(3,252) = 198.52$, $p < .001$, $\eta^2 = .70$) all are similarly affected by the introduced blur. They reveal similar slopes as a function of blur level, whereas blur has less effect on perceived depth ($F(3,252) = 46.37$, $p < .001$, $\eta^2 = .35$). Naturalness shows lower scores for the CBD of 0 meter than for the two larger CBD ($F(2,252) = 11.87$, $p < .001$, $\eta^2 = .09$), but CBD only approaches significance for viewing experience ($F(2,252) = 2.91$, $p = .056$, $\eta^2 = .03$). Image quality is not affected by the CBD levels ($F(2,252) = 0.37$, $p = .69$, $\eta^2 = .00$), whereas the depth percept scores show the largest differences between the CBDs ($F(2,252) = 155.61$, $p < .001$, $\eta^2 = .55$).

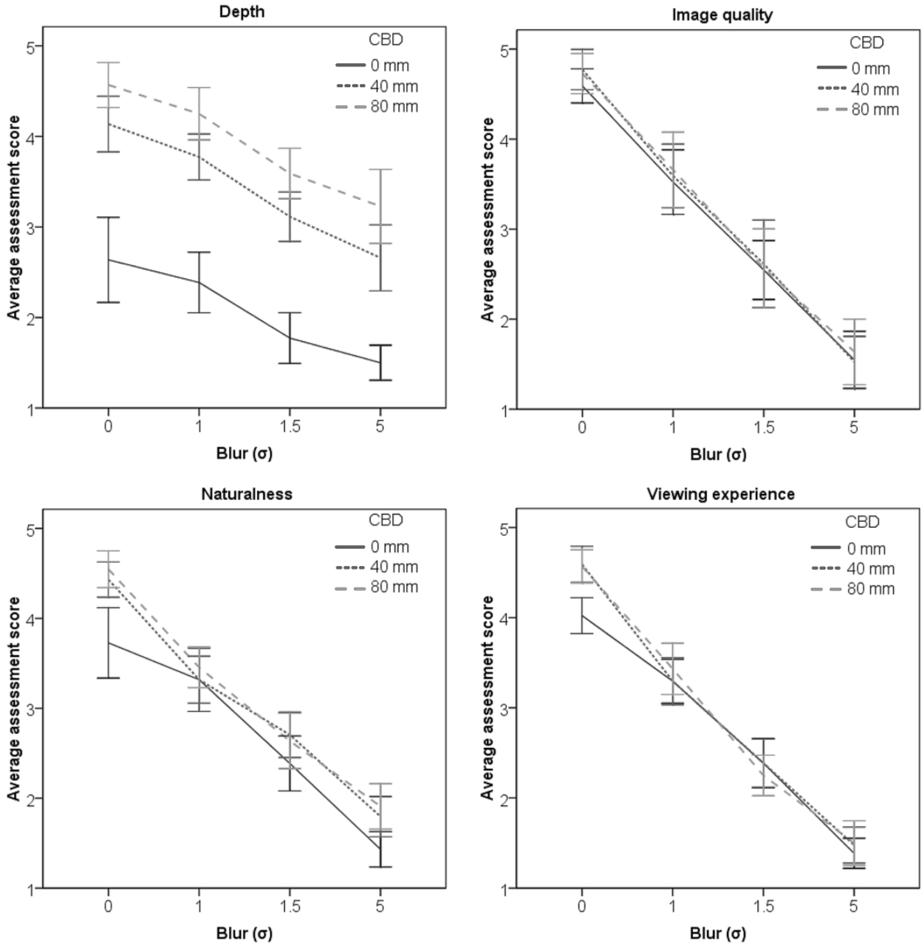


Figure 12. Mean assessment scores with their 95% confidence intervals of the evaluation metrics perceived depth, image quality, naturalness and viewing experience. The x-axes represent the variation in blur, the y-axes represent the averaged scores and the different lines represent the different CBD.

2.4.3. Quantification of the 3-D Quality Model

The 3-D Quality Model that is visualized in Figure 7 and mathematically represented in Equation 1, can now be applied to the data. More specifically, this assessment procedure allows naturalness and viewing experience to be factored as a weighted sum of perceived image quality and perceived depth in a regression analysis. Prior to this analysis two aspects require clarification.

The first aspect is that by asking the participants to use the full range of the assessment scale, the assessment scores of the four evaluation criteria cannot be compared directly, since a 'good'-score in image quality does not necessarily correspond to a 'good'-score in perceived depth. In other words, if a different disparity range was applied, the 'good'-score in perceived depth would correspond to a different disparity. Hence, the values of α and β in equation 1 do not only reflect the weighting coefficients of image quality and perceived depth, but also the weight of the relative ranges of noise/blur and CBD. Ideally, the difference in perceptual range for noise/blur and CBD should be reflected only by the assessment scores, resulting in values for α and β that are independent of scaling effects.

To correct for a difference in perceptual impact of the ranges, in a pilot study a direct comparison between the range of noise/blur and range of CBD was performed. Ten participants had to indicate which attribute, noise/blur or stereoscopic depth, was more present, i.e., the "attention-grabber". This procedure allowed the perceptual impact of the levels of noise/blur to be compared to the perceptual impact of CBD. Table 2 presents the results of this comparison in percentage of blur and noise as the "attention-grabber" for all 12 combinations of level of blur/noise and CBD. These 12 values thus represent a varying contribution of blur/noise and CBD, the relative sizes of which can be estimated by a linear model. The simplest model is that noise/blur and CBD cause an internal representation with a strength commensurate with their level, and that these strengths are additive. To keep the model tractable, it is assumed that the internal representation strength increases with a fixed step for each additional level of noise/blur and CBD. This means that the perceptual impact of the ranges is simply the ratio of the step sizes caused by noise/blur and CBD.

TABLE 2

COMPARISON OF THE PERCEPTUAL IMPACT OF THE LEVELS OF BLUR AND NOISE WITH THE PERCEPTUAL IMPACT OF CBD TO DETERMINE THE RELATIVE IMPACT OF THE RANGES OF BLUR/NOISE AND CBD.

		blur level			
		0	1	2	3
	2-D	20%*	60%	95%	100%
CBD level	3-Dhalf	5%	5%	65%	100%
	3-Dfull	0%	20%	50%	85%

		noise level			
		0	1	2	3
	2-D	15%*	30%	90%	100%
CBD level	3-Dhalf	0%	15%	30%	80%
	3-Dfull	0%	0%	35%	75%

*The percentages reflect the relative perceptual impact of blur as 'attention-grabber'.

In estimating the step sizes Thurstone's scaling model (Torgerson, 1967) was adopted, according to which the internal representation strengths can be obtained by taking the normal deviates of the observed percentages. The estimation methods proposed by Thurstone and Torgerson, however, could not directly be employed as there were two percentages of 100 and one of 0, which both relate to infinite normal deviates. Therefore a linear least squares estimation method (LSE) had been applied on the remaining nine percentages in order to establish the step sizes. Once these had been obtained, it was possible to construct a predicted data matrix that filled the three missing cells, enabling comparison of all 12 observed percentages with the predicted ones. From the LSE analysis it appeared that the step size for blur was 1.16 and that for CBD 0.73. The ratio of these step sizes is 1.6. The fit of the model was based on the correspondence between all 12 observed and predicted percentages and showed a proportion of explained variance of 0.953. This means that the estimated step sizes are sufficiently robust for assessing the relative perceptual impact. A similar procedure was performed to compare the perceptual impact of the range of noise and range of CBD, and revealed that the perceptual impact of the full range of CBD was equivalent to the perceptual impact of 1.49 times the full range of noise with 89% of the data explained. To account for these differences, the depth scores were multiplied by 1.60 in the blur dataset and by 1.49 in the noise data set before performing the regression analysis.

The second aspect that requires clarification is that the results presented in Figure 11, and confirmed by the statistical analyzes, show that both image quality and the depth percept are affected by variations in noise and CBD. The technical parameters 'noise' and 'CBD' themselves do not interact ($p > .177$). A stated before, though perceived image quality and perceived depth can be evaluated independently, in their relationship to the physical image characteristics they are not independent. Hence, instead of incorporating the perceived image quality and depth percept scores in the regression, the regression analysis is performed at the level of the technical parameters noise and CBD. More specifically, the scores of perceived image quality and perceived depth were both rewritten into separate components, one related to the impact of noise/blur and one related to the impact of CBD. Since these two parameters did not interact, we obtained two new independent contributions. This approach is explained in more detail hereafter. We start by explicitly incorporating the effect of both noise and CBD on the scores of perceived image quality and perceived depth as denoted in Equation 2 (a similar approach was performed for the effects of blur and CBD on the scores).

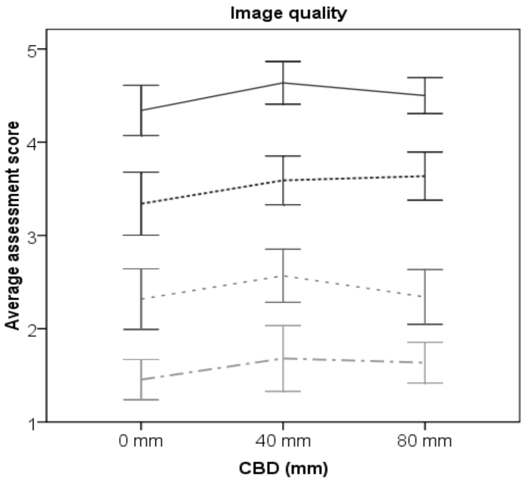
$$E.M.\left(IQ(n,d);D(d,n)\right) = \alpha \cdot IQ(n,d) + \beta \cdot D(d,n) \quad (2)$$

with n = level of noise
d = level of CBD

To determine the impact of noise and CBD on image quality, the image quality scores have to be analyzed as a function of these two parameters, as shown in Equation 3.

$$IQ(n,d) = a(n_{1,2,3,4}) + b \cdot f(d_{1,2,3}) \quad (3)$$

Since noise and CBD are independent as stated before, the contribution of noise to the image quality scores can be written as an offset $a(n)$ depending on the noise level (n_1, n_2, n_3, n_4) and the variation of the image quality scores as a function of CBD is represented by the functional behaviour $b \cdot f(d)$.



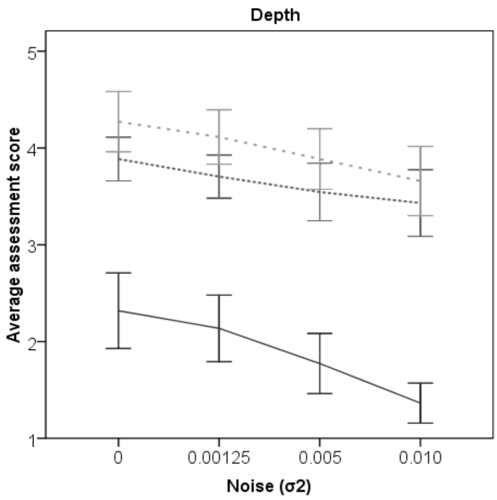
$$n_1 : IQ = 0.0795 \cdot d + 4.333$$

$$n_2 : IQ = 0.1477 \cdot d + 3.227$$

$$n_3 : IQ = 0.0114 \cdot d + 2.386$$

$$n_4 : IQ = 0.0090 \cdot d + 1.409$$

Figure 13. Mean perceived image quality scores (y-axis) per noise level (different lines) as a function of CBD (x-axis), with on the right side a linear fit per noise level represented mathematically.



$$d_1 : D = -0.3102 \cdot n + 6.75$$

$$d_2 : D = -0.2284 \cdot n + 6.03$$

$$d_3 : D = -0.4841 \cdot n + 4.06$$

Figure 14. Mean perceived depth scores (y-axis) per CBD (different lines) as a function of the level of noise (x-axis), with on the right side a linear fit per CBD level represented mathematically.

By fitting different functions for *f(d)* in Equation 3 it became clear that the linear one was the most appropriate ($R^2 > .93$). More complex functions (e.g., exponential, logarithmic, polynomial) made the model unwieldy without significantly explaining more variance of the data. Figure 13 depicts the linear solution of Equation 3 per noise level graphically as well as mathematically. A similar approach is used to describe the impact of noise and CBD on perceived depth in Equation 4.

$$D(d,n) = a(d_{1,2,3}) + b \cdot f(n_{1,2,3,4}) \tag{4}$$

Also here the analysis yielded the linear solution as being most appropriate and the result is depicted both graphically and mathematically per CBD level in Figure 14.

Substituting the expressions of Equation 3 and Equation 4 into Equation 2 yields – in general terms – Equation 5.

$$E.M.(IQ(n,d);D(d,n)) = \alpha \cdot (a_{IQ}(n) + b_D \cdot d) + \beta \cdot (a_D(d) + b_{IQ} \cdot n) \tag{5}$$

Rearranging the parameters into a noise dependent and a CBD dependent component results in Equation 6, and more generally in Equation 7.

$$E.M.(IQ(n);D(d)) = \alpha \cdot \left(a_{IQ}(n) + \frac{\beta \cdot b_{IQ}}{\alpha} \cdot n \right) + \beta \cdot \left(a_D(d) + \frac{\alpha \cdot b_D}{\beta} \cdot d \right) = \tag{6}$$

$$E.M.(IQ(n);D(d)) = \alpha \cdot IQ'(n) + \beta \cdot D'(d) \tag{7}$$

Performing a regression analysis on these new components using the data of Figure 13 and Figure 14 yields the results for naturalness and viewing experience outlined in Table 3.

TABLE 3
PREDICTED WEIGHT COEFFICIENTS EXPERIMENT 1.

		IQ'	D'	R^2
noise	naturalness	0.66	0.18	0.95
	viewing experience	0.74	0.19	0.96
blur	naturalness	0.60	0.22	0.97
	viewing experience	0.86	0.12	0.98

Figure 15 visualizes the observed vs. predicted naturalness and viewing experience scores. For the predicted scores the weights were normalized to a sum of one for both naturalness and viewing experience as outlined in Table 4.

TABLE 4
NORMALISED PREDICTED WEIGHT COEFFICIENTS EXPERIMENT 1.

		IQ'	D'
noise	naturalness	0.78	0.22
	viewing experience	0.79	0.21
blur	naturalness	0.73	0.27
	viewing experience	0.88	0.12

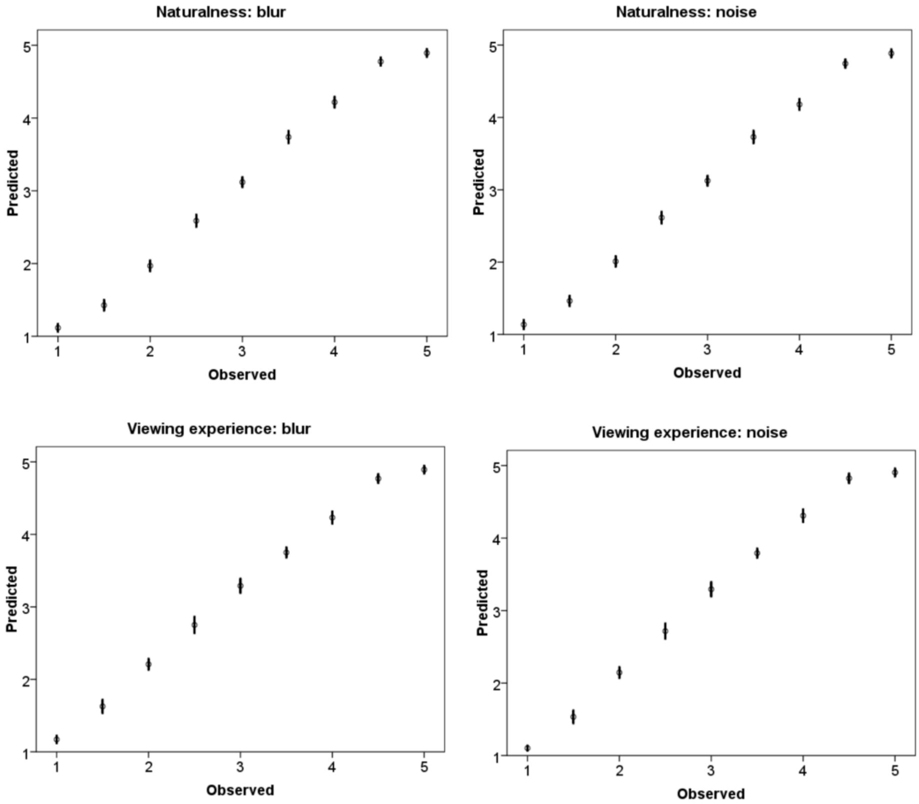


Figure 15. The predicted vs. observed naturalness and viewing experience scores of Experiment 1.

2.4.4. Discussion

The results confirm that both naturalness and viewing experience follow variations in image quality introduced by different levels of noise and blur, as well as variations in stereoscopic depth introduced by different levels of CBD. This is in line with previous research (Seuntiëns et al., 2008).

Provided that the concepts of perceived depth and image quality can be assessed independently and thus be analyzed as such, the 3-D Quality Model describes how the evaluation criteria naturalness and viewing experience are affected in a relative sense by variations in stereoscopic depth and image quality. The results obtained here for a stereoscopic imaging system with perfect image separation, reveal, however, that perceived depth and image quality are both affected by the level of noise/blur and CBD. In other words, the perceived depth is not only affected by the change in CBD, but also by the introduced blur and noise. Usually, blur and noise are unwanted properties of a display system and can be regarded as image quality artifacts. In addition, both artifacts have the perceptual effect that distinctions between objects in the image become less apparent. Since this is also true for distinctions between objects in the 3-D direction, both artifacts degrade the perceived depth. Even more, blur is also a monoscopic depth cue; it directly stimulates accommodation (for more details see Chapter 3) and can negatively affect depth perception. The perceived image quality scores are only significantly affected by a change in CBD for the noise data set. The effect is small and at this point it remains difficult to determine the cause.

For the model to be applied to this data set, the effect of CBD and noise/blur had to be incorporated into both the depth percept and the image quality assessment. Hence, the approach chosen was to rewrite the scores of perceived image quality and perceived depth into separate components, one related to the impact of noise/blur and one related to the impact of CBD. Since these two parameters did not interact, we obtained two new independent contributions for the regression analysis. The linear solution was good enough ($R^2 > .93$) to model the relationship of perceived image quality and depth to the physical image characteristics. In addition, for our current purposes it is less relevant to determine how image quality or perceived depth are precisely affected by a specific 2-D or 3-D variation in order to increase the R^2's. We are more interested in constructing a model that accounts for the dependency of perceived image quality and depth in their relationship to the physical image characteristics, without losing its comprehensibility. This approach provided a good approximation of the original assessment scores for naturalness and viewing experience. The modelled assessments confirmed that 1) naturalness (*Nat.*) and viewing experience (*V.E.*) both incorporated variations of noise and blur to a similar degree, and 2) naturalness incorporated variations of CBD to a higher degree than viewing experience. This approach resulted in the following averaged model weights: $Nat.=0.75 \cdot IQ'(n)+0.25 \cdot D'(d)$ and $V.E.=0.83 \cdot IQ'(n)+0.17 \cdot D'(d)$.

2.5. Experiment 2

This experiment aims to validate the 3-D Quality model and as such confirm the findings of Experiment 1, though based on a more real-life viewing environment by using an autostereoscopic display.

2.5.1. Experimental set-up

Nineteen participants assessed four original images that were varied in image quality (four levels of Gaussian blur) and stereoscopic depth (three screen disparities (SD)) in terms of three evaluation metrics: image quality, perceived depth and naturalness. As a consequence of too little time, the stimuli were not evaluated in terms of viewing experience. Hence, the experiment was a 4 x 3 x 4 (blur x CBD x image) within-subject design resulting in 48 conditions per evaluation metric. The three evaluation metrics were run in separate sessions with at least four days between sessions per participant. The stimuli were randomised per session and the order of the sessions was randomised over the participants. The viewing distance was 3 meters.

Participants

Two females and seventeen males participated. Their age ranged from 22 to 46 years. All participants had a normal or corrected to normal visual acuity of > 1 (as tested with the Landolt-C test) and a stereo acuity of < 30 seconds of arc (as tested with the Randot stereo test).

Stimuli

The image material used consisted of four still images, Playmobiles, Puzzle, Nature and Balloon, that varied in SD and level of blur. The original images are depicted in Figure 16. The Playmobiles image has been used by Seuntiëns et al. (2005) and in Experiment 1.

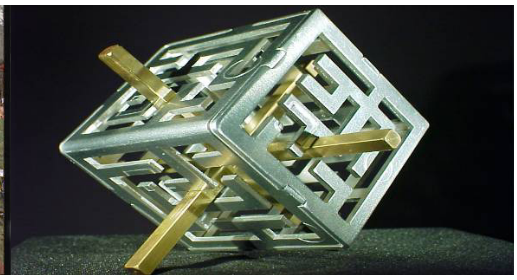
The input format for the Philips 3-D display is an RGB image (original) + corresponding Z-image (depth map), with a resolution of 940 x 540 each (Redert et al., 2007). Three SD levels were created by varying the gain factor of the Z-image. This resulted in a 2-D level (zero screen disparity) and two stereoscopic levels with a maximum screen disparity that corresponded to 1.6 mm and 3.2 mm pixel shift between subsequent views.

Four levels of blur, similar as those in Experiment 1, were introduced to each original image to vary the image quality, i.e., by using a Gaussian filter in Matlab® with standard deviations (σ) of 0, 1, 1.5 and 5 on the RGB image.

Equipment

The stimuli were displayed on a 42" Philips autostereoscopic nine-view lenticular LCD with a resolution of 960 x 544 and a viewing cone (i.e. sum of all nine views) of 21 degrees (van Berkel and Clarke, 1997). Figure 17 depicts the principle of such a autostereoscopic display for a four view version as an example. It consists of a lenticular sheet (i.e., a sheet of cylindrical lenses) that is placed on top of an LCD in such a way that if the correct image information is put on the pixels underneath the lenses, nine different views can be transmitted in nine different directions.

Playmobiles



Puzzle

Nature

Balloon

Figure 16. The original images used in Experiment 2.

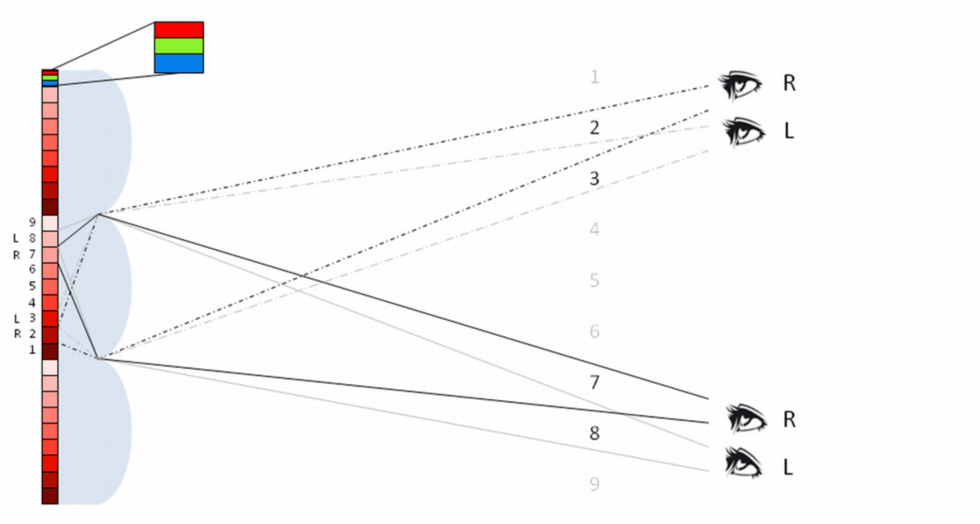


Figure 17. Principles of an autostereoscopic lenticular LCD display.

At the right viewing distance, a viewer perceives a different view in each eye, and as such perceives 3-D. The lenticular is designed such that a viewing distance of three meters provides the best depth perception. Autostereoscopic displays are still regarded as a promising 3-D technology since it provides 'unrestricted' 3-D viewing (Wang, Tu, Chen, Zhang, Teunissen & Heynderickx, 2010). The choice of using an autostereoscopic display was based on several advantages. Firstly, it allows viewers to view the stereo image without the use of glasses. Secondly, it supports motion parallax, allowing viewers to look around objects by moving their head. Next, multiple viewers can be accommodated at the same time due to the use of nine views that are repeated in multiple viewing cones. And finally, some disadvantages present in current stereoscopic displays are absent, e.g., reduction in brightness as a result of the use of polarization filters or parallax barriers, colour break-up as a consequence of using time-sequential fields, or perceived flicker due to using shutter glasses. In our experiment, participants were tested individually and were asked to use a chinrest. As a consequence the second and third advantage of having an autostereoscopic display were not relevant for this particular experiment. A disadvantage of the design choice for this particular display is a loss of spatial resolution due to the generation of the nine views. Another disadvantage is that crosstalk has been intentionally employed in the display design in order to avoid a picket-fence effect (banding) and to minimize image flipping (the discrete transitions between neighbouring views) (van Berkel and Clarke, 1997). As a consequence, the display suffers from a level of blur that is noticeable, and can become annoying as screen disparity increases. Even more, since the input format consists of a single view with accompanying depth map, not all image information needed for the outer views of the display is available in the input signal, which can result in 3-D artifacts.

Procedure and Statistical analyses

The procedure and statistical methods were similar to those of Experiment 1.

2.5.2. Results

Figure 18 depicts the average rating scores averaged over content en participants with their error bars (representing the 95% confidence intervals) per evaluation metric as a function of blur level (increasing along the x-axes) and SD (as parameter).

Figure 18 shows that naturalness ($F(3,900) = 225.52$, $p < .001$, $\eta^2 = .43$) and image quality ($F(3,900) = 326.11$, $p < .001$, $\eta^2 = .55$) are similarly affected by the introduced blur. Image quality and naturalness reveal similar slopes as a function of blur level, whereas blur has less effect on perceived depth ($F(3,900) = 39.38$, $p < .001$, $\eta^2 = .11$). Naturalness receives lower scores for the SD of 0 than for the two larger SD values ($F(2,900) = 8.21$, $p < .001$, $\eta^2 = .12$), whereas the opposite is true for image quality, though to a lesser extent ($F(2,900) = 15.09$, $p < .001$, $\eta^2 = .04$). The depth percept scores show the largest differences between the SD values ($F(2,900) = 473.27$, $p < .001$, $\eta^2 = .57$).

2.5.3. Quantification of the 3-D Quality Model

A similar procedure as described in detail in paragraph 2.4.8. was followed and yields the results for naturalness outlined in Table 5. Figure 19 visualizes the observed vs. predicted naturalness scores. For the predicted scores the weights were normalized to a sum of one as outlined in Table 5.

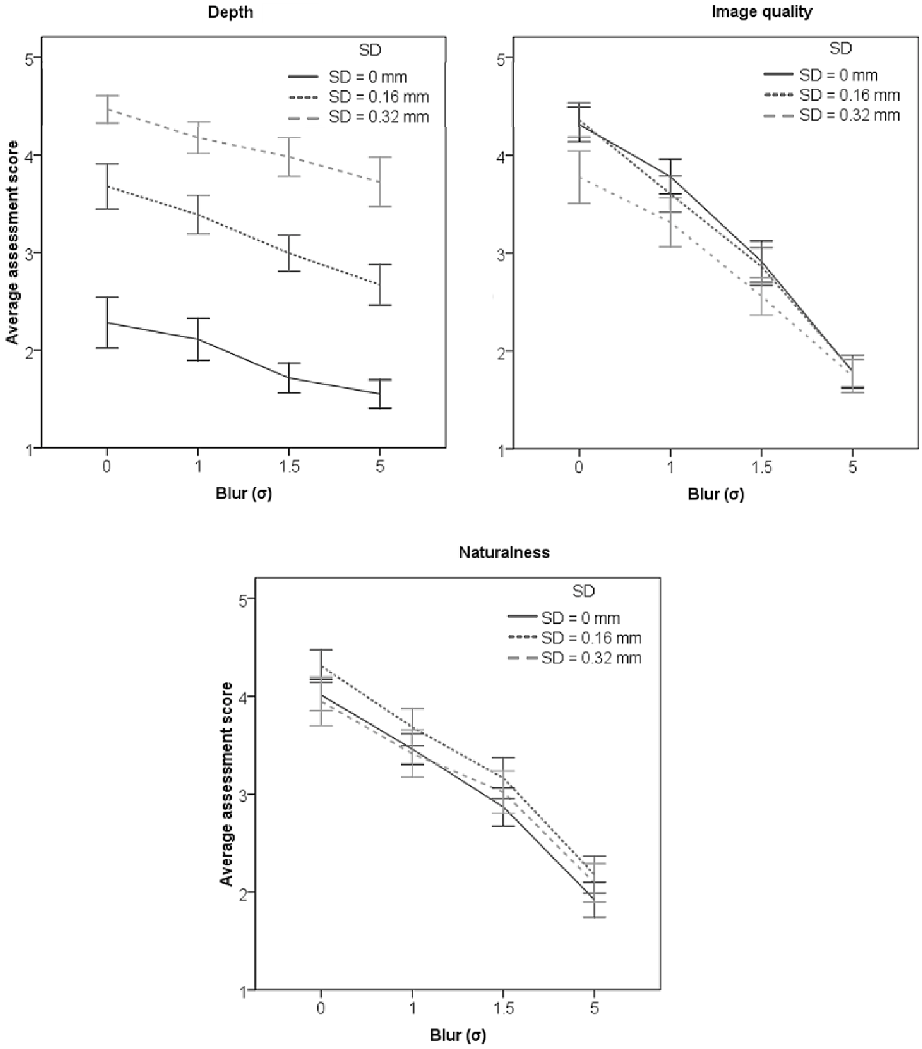


Figure 18. Mean assessment scores with their 95% confidence intervals of the evaluation metrics perceived depth, image quality and naturalness. The x-axes represent the variation in blur, the y-axes represent the averaged scores and the different lines represent the different SD.

TABLE 5
PREDICTED WEIGHT COEFFICIENTS EXPERIMENT 2.

	IQ'	D'	R^2
Naturalness (regression)	0.63	0.19	0.96
Naturalness (normalised)	0.76	0.24	

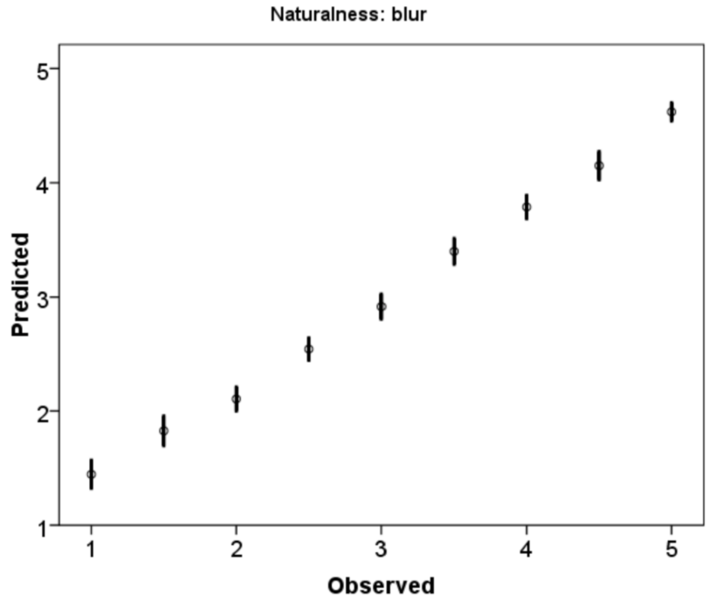


Figure 19. The predicted vs. observed naturalness scores of Experiment 2.

2.5.4. Discussion

Our results confirm that for lenticular screens naturalness follows variations in image quality induced by different levels of blur, as well as variations in stereoscopic depth induced by different levels of SD. This is in line with previous research (Seuntiëns et al., 2008) and with the results of Experiment 1.

The results also reveal that perceived depth and image quality are both affected by blur and SD. The effect size of the interaction effect is even larger than in Experiment 1 and was a consequence of some technical characteristics of the 3-D display. Increasing blur affects both perceived image quality and perceived depth, which was already discussed in the paragraph 2.4.9. Increasing the SD is accompanied in this particular display by an increase in the perception of crosstalk. The crosstalk is perceived as unnatural blur and reduces the image quality. This problem was already noted by Seuntiëns et al., (2005).

A similar approach as chosen Experiment 1 was used to analyze the results with respect to the 3-D quality model. The result provides a good approximation to the original scores for the naturalness assessments: $Nat.=0.76\cdot IQ'(n)+0.24\cdot D'(d)$.

2.6. Experiment 3

This experiment aims to validate the 3-D Quality model and as such confirm the findings of Experiments 1 and 2, by increasing the coverage of imaging technology even more, incorporating 2-D-to-3-D converted content and displaying it on an nine-view autostereoscopic display.

2.6.1. Experimental set-up

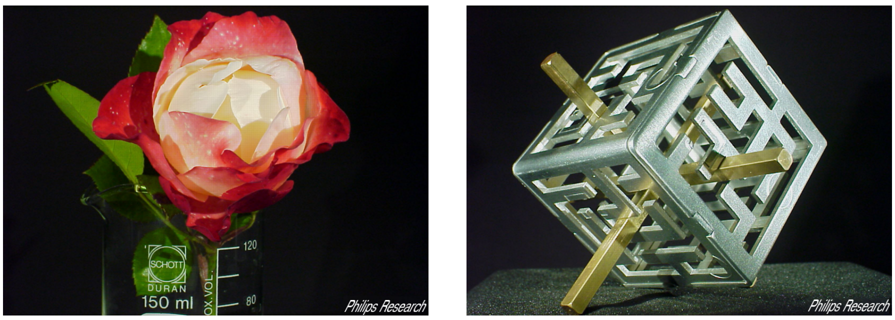
Four evaluation metrics, i.e., image quality, perceived depth, naturalness and viewing experience, were assessed by fourteen participants with good (stereoscopic) vision in different sessions. Three levels of white Gaussian noise and three camera base distances (CBD) were used to vary image quality and depth respectively of two images. Hence, the experiment was a 2 x 2 x 3 x 3 (repetition x image x noise x CBD) within-subject design resulting in 36 conditions per session. The stimuli were randomised per session and the order of the sessions was randomised over the participants. The viewing distance was 0.40 meters.

Participants

Two female and eighteen males participated. Their ages ranged from 24 to 32 years. All participants had a normal or corrected to normal visual acuity of > 1 (as tested with the Landolt-C test) and a stereo acuity of < 30 seconds of arc (as tested with the Randot stereo test).

Stimuli

Two original images, rose and puzzle as depicted in Figure 20, were used. These images had been captured with a nine-camera set-up yielding accurate image information for each of the nine views and were used by Seuntiëns et al. (2008) as well. The original image puzzle has been used in Experiment 2 as well.



rose puzzle

Figure 20. The original images used in the Experiment 3.

For the introduction of stereoscopic depth three CBDs between neighbouring views, i.e., 0, 0.01 and 0.025 meters, were implemented. For the CBD of 0 meters, i.e., the 2-D level, the original images was displayed nine times to ensure a similar resolution as for the other CBDs. The CBD of 0.01 meter was obtained with a 2-D-to-3-D conversion algorithm (Redert et al., 2007). The CBD of 0.025 meters, however, was captured with a nine-view camera set-up. The latter was used by Seuntiëns et al. (2008) as well and was incorporated here to compare results.

For the variation in image quality three levels of white Gaussian noise with variances (σ^2) of 0, 0.005 and 0.025 (expressed in pixel units) were added to the original images with Matlab®.

Equipment

The stimuli were displayed on a 20" Philips autostereoscopic nine-view lenticular LCD (van Berkel and Clarke, 1997). Each view had a width of 3.4 degrees and the optics were optimized for a viewing distance of 0.40 meters. The native resolution of the display was 1600x1200 pixels. A more detailed description was already given for Experiment 2.

Procedure and Statistical analyses

The procedure and statistical methods were similar to those of Experiment 1.

2.6.2. Results

Figure 21 depicts the average assessment scores with their error bars (representing the 95% confidence intervals) per evaluation metric as a function of noise level (increasing along the x-axes) and CBD (as parameter).

Figure 21 shows that naturalness ($F(2,242) = 150.04$, $p < .001$, $\eta^2 = .56$), viewing experience ($F(2,242) = 169.20$, $p < .001$, $\eta^2 = .59$) and image quality ($F(2,242) = 185.42$, $p < .001$, $\eta^2 = .61$) all are similarly affected by the introduced noise. Viewing experience, naturalness and image quality show similar slopes as a function of noise level, whereas noise has less effect on perceived depth ($F(2,242) = 14.83$, $p < .001$, $\eta^2 = .11$). Image quality shows higher scores for the CBD of 0 meter than for the two larger CBD ($F(3,252) = 12.10$, $p < .001$, $\eta^2 = .09$). The opposite is true for naturalness ($F(3,252) = 4.72$, $p < .01$, $\eta^2 = .14$), whereas viewing experience does not significantly distinguish between the CBDs ($F(3,252) = .73$, $p = .476$, $\eta^2 = .01$). The depth percept scores show the largest differences between the CBDs ($F(3,252) = 68.14$, $p < .001$, $\eta^2 = .46$).

2.6.3. Quantification of the 3-D Quality Model

A similar procedure as described in detail in paragraph 2.4.8. was followed and yields the regression results for naturalness and viewing experience outlined in Figure 21. Figure 22 visualizes the observed vs. predicted naturalness and viewing experience scores. For the predicted scores the weights were normalized to a sum of one for both naturalness and viewing experience as outlined in Table 6.

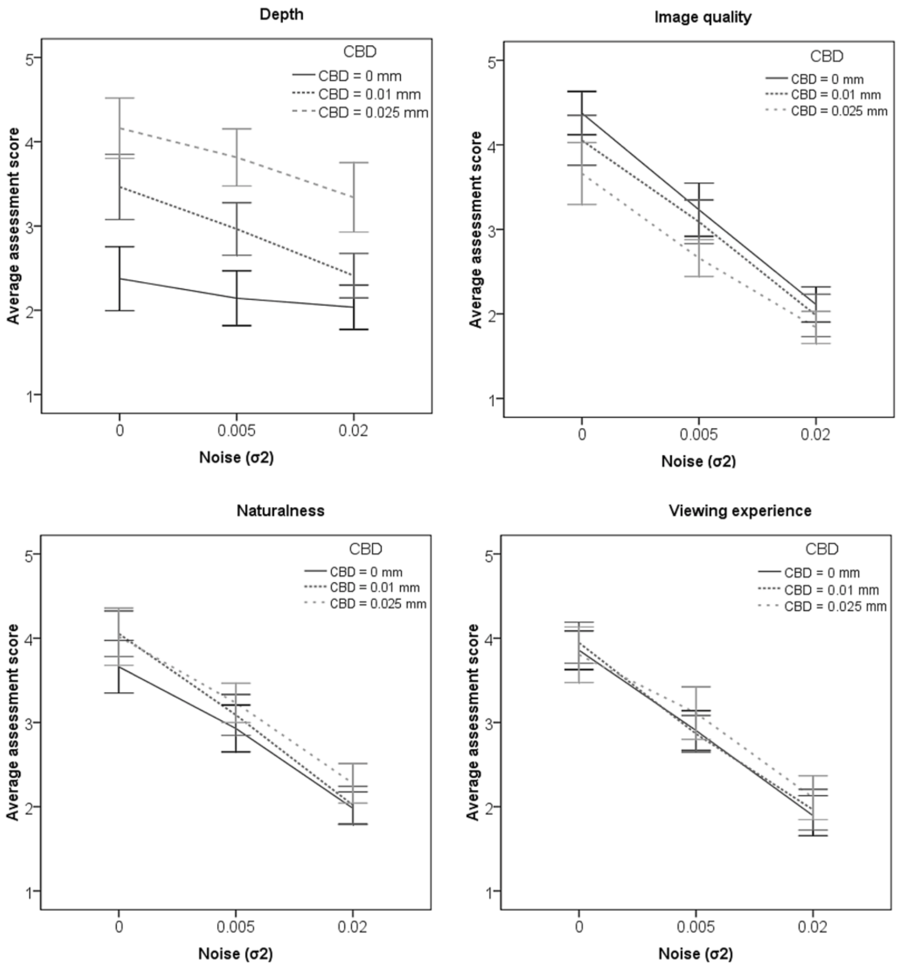


Figure 21. Mean assessment scores with their 95% confidence intervals of the evaluation metrics perceived depth, image quality, naturalness and viewing experience. The x-axes represent the variation in noise, the y-axes represent the averaged scores and the different lines represent the different CBD.

TABLE 6
PREDICTED WEIGHT COEFFICIENTS EXPERIMENT 3.

		IQ'	D'	R^2
naturalness	regression	0.64	0.27	0.97
	normalized	0.73	0.19	
viewing experience	regression	0.70	0.30	0.95
	normalized	0.79	0.21	

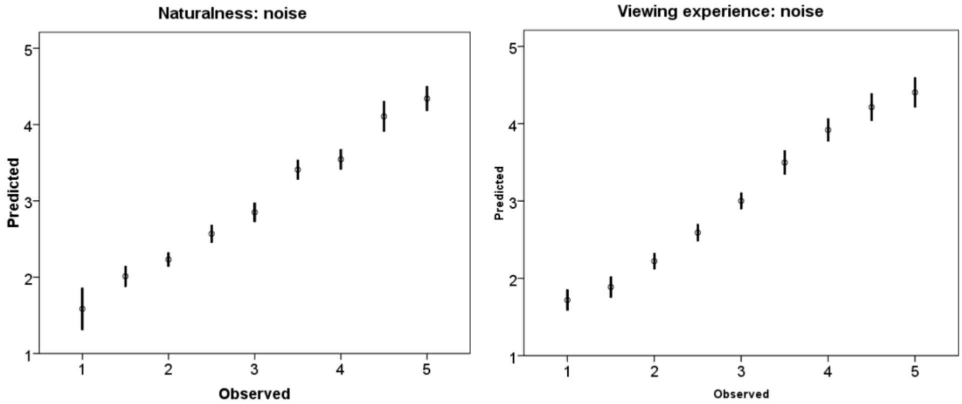


Figure 22. The predicted vs. observed naturalness and viewing experience scores of Experiment 3.

2.6.4. Discussion

The results confirm that naturalness and viewing experience follow variations in image quality induced by different levels of noise, as well as variations in stereoscopic depth induced by different levels of CBD. This is in line with previous research (Seuntiëns et al., 2008) and with the results of Experiments 1 and 2.

The results also reveal that perceived depth and image quality are both affected by noise and CBD. How the change in CBD affects perceived image quality is partly explained in Chapter 2.5.8., i.e., via perceived blur due to an increased perception of crosstalk. In addition, 2-D-to-3-D conversion has a negative effect on the perceived image quality, due to accompanying spatial artifacts that become apparent when screen disparity is increased. Perceived depth is not only affected by the change in CBD, but also by noise. Since noise was added to the single interleaved image of the nine views, it became correlated. As a consequence, it was perceived as a 'noisy window' displayed at a single depth layer, and so affected perceived depth negatively.

A similar approach as chosen for Experiments 1 and 2 was used to analyze the results with respect to the 3-D quality model. These relations provide a good approximation to the original assessment scores. Doing so yields the following relations:

$Nat.=0.70 \cdot IQ'(n)+0.30 \cdot D'(d)$ and $V.E.=0.79 \cdot IQ'(n)+0.21 \cdot D'(d)$.

2.7. Discussion

The results obtained in all three experiments reveal that perceived depth and image quality are affected by physical image characteristics related to display and signal properties. In line with previous research (Seuntiëns et al., 2005; Strohmeier and Tech, 2010), the assessment of perceived depth mainly reflects variations in screen disparity, while the assessment of image quality is based on a perceived quality degradation due to 2-D and 3-D artifacts. Adding blur, noise or screen disparity to image content can affect both perceived image quality and perceived depth. Evidently, adding blur or noise affects image quality, but it also reduces the amount of perceived depth. Adding screen disparity increases the amount of perceived depth, but it can reduce the perceived image quality by magnifying the perception of artifacts. As a consequence, in their relationship to the physical image characteristics, perceived depth and perceived image quality are not independent. However, at the perceived level they are independent, as it is possible to change one without affecting the other (Seuntiëns et al., 2005).

The three experiments confirm that both naturalness and viewing experience follow variations in image quality, as well as variations in perceived depth. The 3-D Quality Model describes how naturalness and viewing experience are affected in a relative sense by variations in perceived depth and image quality. Table 7 provides an overview of all weight coefficients.

TABLE 7
WEIGHT COEFFICIENTS OF ALL THREE EXPERIMENTS AND THE AVERAGED WEIGHT COEFFICIENTS.

	naturalness		viewing experience	
Experiment	image quality	perceived depth	image quality	perceived depth
Experiment 1 (noise)	0,78	0,22	0,79	0,21
Experiment 1 (blur)	0,73	0,27	0,88	0,12
Experiment 2 (blur)	0,76	0,24	-	-
Experiment 3 (noise)	0,70	0,30	0,79	0,21
Average	0,74	0,26	0,82	0,18

The weight coefficient in Table 7 show that the differences in the weights between the experiments are small for both naturalness and viewing experience, even though the differences in experimental design were large (image quality and screen disparity variations, viewing distance, content and imaging system). In this sense, it is demonstrated that both naturalness and viewing experience can consistently be modelled as a weighted sum of perceived image quality and perceived depth. More specifically, both naturalness and viewing experience seem appropriate as higher concept evaluation metrics that reflect image quality as well as the added value of stereoscopic depth of stills. This can be of interest from a display manufacturer and a video processing point of view, i.e., the specific balance between image quality related and perceived depth related technological variables can be used to improve the overall 3-D quality in terms of naturalness or viewing experience (You, Xing, Perkins and Wang, 2010; Benoit, Callet, Campisi and Cousseau, 2008; Sazzad, Yamanaka and Horita, 2010). And of equal importance, the 3-D Quality Model is applicable to different imaging contexts, i.e., ranging from a 'laboratory' setting to a 'realistic consumer's home environment'.

Which evaluation metric is most appropriate to evaluate 3-D depends on the criteria one uses. In our opinion two criteria are of importance: (1) the evaluation metric that weighs the added value of stereoscopic depth most in addition to image quality, and (2), the evaluation metric that differentiates best between different levels of image quality and screen disparity. Since they both reflect variations in image quality to a similar degree, mainly based on the first criterion, it seems that naturalness is more appropriate to evaluate the impression of stereoscopic stills than viewing experience.

For future research the 3-D Quality Model can be applied to moving sequences, for which presence already has proven to be a very appropriate evaluation metric (IJsselsteijn, 2001), or to different display applications such as computer games, mobile phone use or advertising (de Boer, Verleur, Heuvelman & Heynderickx, 2010). It might be interesting to incorporate other attributes in the model, e.g., the impact of visual discomfort. Previous research for example revealed a model where “the subjective quality of stereoscopic images can be described as a function of perceived depth attenuated by subjective eye-strain” (IJsselsteijn, 1998b).

2.8. Conclusion

We propose a 3-D Quality Model as an extension of Engeldrum's Image Quality Circle, since many of our results confirm that the added value of stereoscopic depth is not captured in the Image Quality Circle. As a consequence, higher level evaluation metrics are required, that reflect both image quality and perceived depth. To this end, the 3-D Quality Model describes a higher-level evaluation metric as a weighted sum of perceived image quality and perceived depth. Two evaluation metrics, namely naturalness and viewing experience, were investigated. The results showed that perceived image quality and perceived depth are not independent in their relationship to physical image characteristics, but are at the perceptual level. Variations in image quality are reflected by viewing experience and naturalness to a similar extent, yet the added value of stereoscopic depth in stills is more incorporated in naturalness. The 3-D Quality Model based on naturalness as an evaluation metric is validly applicable to stereoscopic stills, and its score is determined for approximately 74% by image quality and 26% by perceived depth. The model contributes to a more effective design circle for 3-D-TV and its technological parameters with the aim to optimize customers’ preferences.

-CHAPTER 3-

Visual fatigue and visual discomfort

"The utter ease and simplicity and the thoroughly satisfactory results of even elementary stereo movies are so great that the hesitation of the amateur to try it is incredible. Everyone seems to be awaiting some very mysterious, very complex, very magical (and costly) method for doing this simple thing. I only wish I could think of some way to convince you that this hope is all nonsense. You already have available everything you need to make perfectly beautiful stereo movies with very little expenditure and with very good assurance of success in the first roll of film you expose."

Herbert C. McKay (1953; p. 271)

ABSTRACT

Visual discomfort has been the subject of considerable research in relation to stereoscopic and autostereoscopic displays. The current chapter clarifies the importance of various causes and aspects of visual discomfort. When disparity values do not surpass a limit of one degree, which still provides sufficient range to allow satisfactory depth perception in stereoscopic television, classical determinants such as excessive binocular parallax and accommodation-vergence conflict appear to be of minor importance (see section 6.4). Visual discomfort, however, can still occur within the one degree limit, and we believe the following factors to be the most pertinent in contributing: (1) temporally changing demand of accommodation-vergence linkage, e.g., by fast motion in depth, (2) 3-D artifacts resulting from insufficient depth information in the incoming data signal yielding spatial and temporal inconsistencies, and (3) unnatural blur. In order to characterize and understand visual discomfort, multiple types of measurements, both objective and subjective, are required.

This chapter is based on:
Lambooij, M., IJsselsteijn, W. A., Fortuin, M., and Heynderickx, I. (2009). Visual Discomfort and Visual Fatigue of Stereoscopic Displays: a Review. *Journal of Imaging Technology and Science.* 53:1-14.

3.1. Introduction

The introduction of three-dimensional television (3-D TV) on the public consumer market, much like its desktop-counterpart in the gaming and internet industry, is believed to be just a matter of time and has been compared to the transition from black-and-white to colour TV. For it to be successful, strain-free viewing must be guaranteed, and hence, both image quality and visual comfort must at least be comparable to conventional TV standards (Meesters et al., 2004). Since this promise has not yet been fulfilled, extensive research to understand the factors underlying visual discomfort is needed. An overview of the current status of that research is provided in this chapter. Literature in this area mention conflicts between accommodation and vergence, excessive binocular parallax, and dichoptic errors as major problems potentially leading to visual discomfort. These factors are reviewed in this chapter as well as some additional causes that have become more relevant recently with successive innovations in 3-D imaging systems. Additionally, some experimental set-ups necessary to quantify the degree of visual discomfort in an unambiguous manner are discussed. Finally, a variety of measurement methods are addressed, which can roughly be divided into subjective measures (e.g., questionnaires and functional assessments) and objective measures, indicating the physiological state (e.g., optometric methods and brain activity measurements).

3.2. Human perception of depth

3.2.1. Binocular depth perception

Because our eyes are horizontally separated, each eye has its own perspective of the world, and thus both eyes receive slightly different images. Stereopsis is the perception of depth that is constructed based on the difference between these two retinal images. The brain fuses the left and right image and from retinal disparity, i.e., the distance between corresponding points in these images, and it extracts relative depth information. Even without the benefit of stereopsis, depth can be perceived. This is based on monocular cues, such as perspective, interposition or texture gradients. For an overview of the relative importance of different depth cues at various distances see Cutting and Vishton (1995).

Points that are fixated on by both eyes are projected onto corresponding parts of the retina. For any degree of vergence, the horopter is the surface in space that contains all points whose images stimulate corresponding retinal points, i.e., that all have zero retinal disparity. Points that do not fall on the horopter have retinal disparity. Points located in front of the horopter have a negative or crossed retinal disparity and points located behind the horopter have a positive or uncrossed retinal disparity. Panum's fusional area describes the small region around the horopter where sensory fusion takes place, i.e., the neural process of merging the two retinal images into a single stereoscopic image. Panum’s fusional area allows some imprecision in eye movements without the introduction of diplopia, whereas points lying outside Panum’s fusional space can be perceived as double. The receptive fields are relatively small at the fovea (central fusion) and relatively large in the periphery (peripheral fusion). Hence, the limits of Panum's fusional area are not constant over the retina, but expand at increasing eccentricity from the fovea. At the fovea, sensory fusion is limited to a retinal disparity of one-tenth of a degree, at an eccentricity of 6º to a retinal disparity of one-thirds of a degree (Howard, 2002; Patterson and Martin,

1992) and at 12° degrees of eccentricity to a retinal disparity of two-third of a degree (Patterson and Martin, 1992).

3.2.2. Ocular near triad

Accommodation, vergence and pupillary dynamics, i.e., the ocular near triad, continuously interact to control the functioning of the eyes (Takeda, Hashimoto, Hiruma and Fukui, 1999). To obtain clear, binocular single vision, our eyes are accommodated and converged by an amount that depends on the distance between us and the object of interest. Vergence is defined as movement of our eyes in opposite directions to locate the area of interest on the fovea and accommodation as alteration of the lens to obtain and maintain the object of interest focused on the fovea. The interaction between accommodation and vergence is accompanied by changes in pupil diameter. The pupil constricts with near vergence/accommodation to compensate for a narrow depth of field and increased spherical aberration, and dilates with far vergence/accommodation to reduce diffraction and increase retinal illumination (Howard, 2002). The pupillary dynamics are governed by the autonomic nervous system and reflect mental activity. As such, they can indicate visual discomfort (Ukai and Kato, 2002; Ukai and Howarth, 2008). As part of the ocular near triad changes in pupil diameter can affect accommodation and vergence.

3.2.3. Depth of focus

Our eyes can tolerate small amounts of retinal defocus without adjusting accommodation to perceive a sharp image. The depth of focus (DOF) describes the amount of retinal defocus in which accommodation does not change while objects are perceived clearly (Howarth, 1996; Yano, Emoto and Mitsuhashi, 2004). DOF can be defined as "the variation in image distance of a lens or optical system which can be tolerated without incurring an objectionable lack of sharpness in focus" (Wang and Ciuffreda, 2006). Hence, each single eye has a DOF; it does not depend on stereoscopic vision, but it simply defines the zone in which vision is sharpest and deviations in either direction gradually decrease image quality by the introduction of blur (Smith and Atchinson, 1997). For a review that covers the DOF see Wang and Ciuffreda (2006). They illustrate that the range of DOF is influenced by many factors, of which some are related to target attributes, e.g., contrast, luminance and spatial frequency, and some to eye/brain attributes, e.g., pupil size and age. The DOF ranges from 0.04 to 3.50 diopter, with typical values of approximately 0.2 to 0.5 diopter.

3.2.4. The accommodation-vergence model

Vergence and accommodation are generally modelled as two dual parallel feed-back control systems that interact via cross-links as depicted in Figure 23 (Schor and Kotulak, 1986; Hung, 2001; Rushton and Riddell, 1999; Eadie, Gray, Carlin and Mon-Williams, 2000; Ciuffreda, 2002). Accommodation is primarily retinal blur-driven and vergence primarily retinal disparity-driven and both systems respond to proximity information, i.e., apparent target nearness, such as 'pictorial' depth cues and motion-in-depth cues (Ciuffreda, 2002). Each system includes a tonic component, i.e., an adaptive component, which accounts for slower adaptations to altered viewing situations. Both systems interact via reflexive cross-link interactions. The gains of the cross-link 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).

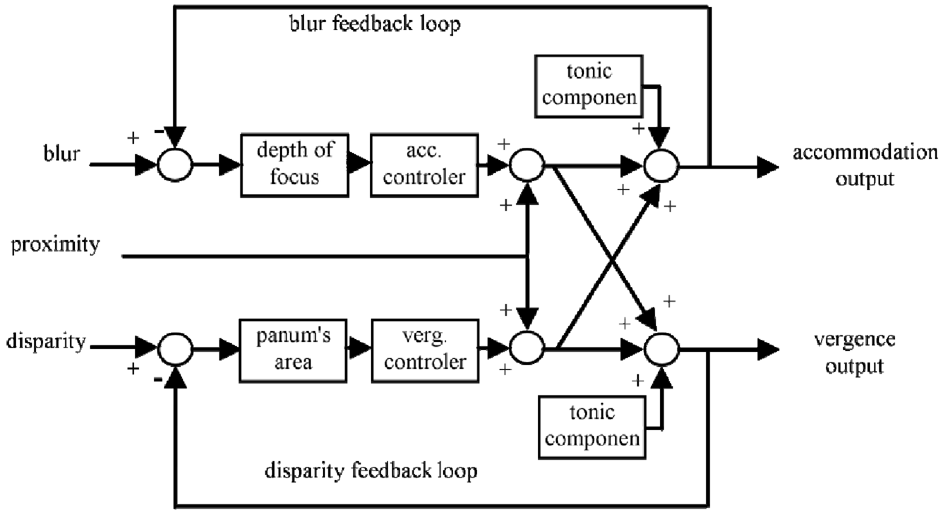


Figure 23. Accommodation and vergence modelled as two dual parallel feed-back control systems that interact via cross-links (Schor and Kotulak, 1986; Hung, 2001; Rushton and Riddell, 1999; Eadie et al., 2000). Accommodation depends on defocus, proximity, tonic adaptation and vergence-accommodation. Vergence depends on retinal disparity, proximity, tonic adaptation and accommodative-vergence. Both systems also provide negative feedback to the input stimuli to obtain stable states. The accommodation-vergence system can be explained as follows. Under natural viewing conditions, accommodation and vergence interact to provide comfortable and clear, binocular, single vision. Small degrees of retinal defocus within the DOF do not drive the accommodation system, and small retinal disparities are fused by sensory fusion and do not drive motoric fusion, i.e., vergence movements. As an object approaches, changes in blur that exceed DOF drive the accommodation controller and changes in retinal disparity that exceed Panum's fusional area drive the vergence controller. The summed output of the controller, the proximal component, the tonic component and the cross-link, describes the overall system's response and provides negative feedback to the input stimuli to obtain a stable state.

3.2.5. Depth cue integration

To provide an accurate, consistent and useful percept of the physical environment, the visual system reduces ambiguity by combining different depth cues. It remains an ongoing debate which strategy the brain uses to extract 3-D depth from optical information in two 2-D retinal images (Howard and Rogers, 2002). A single unified theory about cue integration has not yet been established. Recent research conceptualized depth cue integration as a problem of statistical inference, i.e., the maximum-likelihood estimation of cue combination based on the reliability of the cues (Hillis, Ernst, Banks and Landy, 2002; Hillis, Watt, Landy and Banks, 2004). In stereoscopic displays conflicting cues can be introduced, and it can be even more interesting and important to investigate how the visual system resolves such conflicts. For example, it has been reported that perceived depth decreased when ordinal configural information (i.e., familiarity and convexity) and retinal disparity were 'inconsistent' (Burge, Peterson and Palmer, 2005). Yet, the impact on visual comfort was not addressed.

3.2.6. Individual differences

People differ in human visual system characteristics, which directly determine their ability to perceive stereoscopic depth. One of those characteristics is the interpupillary distance (IPD). People with a small IPD perceive more stereoscopic depth for a fixed set of objects at a fixed viewing distance than people with a large IPD. As such, for a fixed screen disparity, i.e., the distance between two corresponding pixels in two separate views on a stereoscopic display, people with a smaller IPD reach fusional limits more rapidly. Extensive research on the IPD of humans of different gender, race and age showed that the IPD of the vast majority of adults falls within the range of 50 to 70 mm, with a mean and median of approximately 63 mm. To include extremes and children a range of 40 to 80 mm is recommended (Dodgson, 2004).

Visual disorders in early childhood, even if only temporary, can result in stereo blindness. A distinction needs to be drawn between binocular anomalies that typically prevent stereopsis (principally, strabismus and amblyopia, i.e., squint and lazy eye respectively) and non-strabismic binocular anomalies that permit stereopsis but predispose the patient to visual discomfort (asthenopia). The prevalence of strabismus is about 2% (Williams, Northstone, Howard, Harvey, Harrad, Sparrow, 2008) and of amblyopia about 3% (Kanodidou, 2011). For the effect of non-strabismic binocular anomalies see chapter 5. Richards (1970) performed a survey among 150 participants and found that 4% were unable to perceive a hidden Julesz figure in a random-dot stereogram, and 10% had great difficulty detecting its distance relative to the background (Richards, 1970). Visual abilities also vary with age as a result of changes in the structure of the eye. Accommodative ability decreases with age up to about 55 years of age (Ostrin and Glasser, 2004). Conversely, the visual system of children still has a high degree of plasticity, because it is not fully developed until the age of seven to nine (Rushton and Riddell, 1999; Peli, 1999). Moreover, as a result of their small IPD, the impact of too much screen disparity, which can differ between individuals, is larger for children than for adults. Research also revealed that once some visual disorders are established during childhood, such as myopia that is often related to near work, the degree of the disorder typically increases (Goss and Huifang, 1994). This is the main reason why some researchers advise against stereoscopic viewing displays by children, stating that even though little evidence exists that viewing stereoscopic content causes permanent damage to the visual system, there is also no evidence that contradicts this argument.

3.3. Visual fatigue and visual discomfort

Over the last decades, safety and health problems related to video display terminals (VDTs) in general, and specifically stereoscopic displays have been extensively studied. Particularly for stereoscopic displays visual discomfort is mentioned in the literature as one of the important health problems. Hence, for the realization of a comfortable viewing experience on a stereoscopic display, an all-inclusive study of visual discomfort is required.

In the literature, visual discomfort is used interchangeably with visual fatigue. A distinction, however, should be made. In this thesis, visual fatigue refers to a decrease in performance of the human vision system, which can be objectively measured, whereas visual discomfort is its subjective counterpart. This relationship is generally assumed, but to our knowledge never systematically verified. In this thesis the distinction between visual fatigue and visual discomfort will be consistently maintained. When formulated in this way, perceived visual discomfort determined via subjective measurements, is expected to provide an indication of the objectively measurable visual fatigue.

The all-embracing diagnostic term for visual complaints is asthenopia and literally means "weak view". Asthenopia can be concentrated around the eyes, or can be diffuse as a general headache or occur in the neck and shoulders. Much research has been conducted in the past concerning asthenopia, though it seems that the current topics of research such as conceptualizing, measuring and preventing asthenopia to a large extent replicate the pioneering work in the early 1900's (Watten, 1994). Its conceptualization remains ambiguous; different definitions are used across different fields, but no absolute definition exists. In most cases asthenopia is conceptualized as a combination of underlying determinants and symptoms, or by a substitution such as eyestrain (Sheedy, Hayes and Engle, 2003; Murata, Uetake, Otsuka and Takasawa, 2001). Although visual fatigue is nearly synonymous with eyestrain (Sheedy et al., 2003), for clarification in this research a distinction is made. Eyestrain is defined as "the symptoms experienced in the conscious striving of the visual apparatus to clarify vision by ineffectual adjustments" (Panel on Impact of Video Viewing on Vision Workers, 1983). It refers to a specific aspect of the visual system, i.e., continuously resolving ineffectual adjustments. Visual fatigue refers to any visual dysfunction resulting from the use of one's eyes. As such, visual fatigue includes such continuous ineffective adjustments, as well as conflicting or problematic, functional adapted states of the visual system. Hence, visual fatigue is defined as physiological strain or stress resulting from exertion of the visual system.

The determinants of asthenopia are very diverse, and therefore, are still a source of ongoing research. In the area of VDT asthenopia can be caused or induced by anomalies of vision such as heterophoria, vergence insufficiency or accommodative dysfunction. Additionally, it can be related to display problems such as compromised quality of the viewed image, flickering stimuli, suboptimal gaze angles or viewing distance (Sheedy et al., 2003; Blehm, Vishnu, Khattak, Mitra and Yee, 2005). Research concentrated on stereoscopic displays has revealed causes of asthenopia such as (1) anomalies of binocular vision, (2) dichoptic errors, such as geometrical distortions between the left and right image (e.g. keystone distortion, depth-plane curvature, crosstalk and binocular rivalry), (3) conflict between vergence eye movement and accommodation, and (4) excessive binocular parallax (Yano et al., 2004; Emoto, Niida and Okana, 2005; IJsselsteijn, Seuntiëns and Meesters, 2005; Speranza, Tam, Renaud and Hur, 2006; Woods, Docherty and Koch, 1995; Hoffman, Girshick, Akeley and Banks, 2008).

Directly related to the extensive list of determinants is the amount and diversity of symptoms of asthenopia. To give a clear overview, the various symptoms (Sheedy et al., 2003; Murata et al., 2001; Blehm et al., 2005; Emoto et al., 2005; Cooper, Burns, Cotter, Daum, Griffin and Scheiman, 2001) are grouped according to a specific classification provided by Sheedy et al. (2003). They applied a factor analysis to different symptoms and revealed two latent factors, internal and external factors, that can be differentiated by sensation type, sensation location and induced condition. The internal factors include ache, strain and headache, and denote symptoms located behind the eyes. The external factors include burning, tearing, irritation and dryness, and denote symptoms located in front of the eyes.

Consequently, the multiple determinants and symptoms result in numerous and widespread indicators to measure the degree of asthenopia. An essential issue in the determination of asthenopia is that sensations or symptoms can refer to different stimulated anatomical locations. A single underlying factor, e.g., vergence insufficiency, can stimulate anatomical locations such as medial ocular muscles, accommodation of the ciliary body and the tear-gland. Stimulation of each of these will probably results in a different sensation, yet all are due to the same primary underlying determinant (Sheedy et al., 2003). Hence, the concept visual fatigue cannot be evaluated with only one objective indicator. In addition, many of the ocular changes representing visual fatigue can also be regarded as healthy characteristics of our biological system adapting to altered visual environments. The occurrence of visual discomfort also needs to be verified. Only physiological changes that are accompanied by negative psychological effects in function or comfort should be critically examined for their magnitude and subjective impact. Though our visual system adapts and can prevent psychological effects to from occurring in the short-term, their impact can increase in strength after prolonged viewing of stereoscopic content. Hence, the effects of a prolonged period of viewing are also of interest here and should be critically examined for their magnitude and subjective impact. Therefore, multiple types of measurements, both objective as well as subjective, need to be combined in order to determine the degree of visual fatigue and visual discomfort in a sensitive, accurate, reliable and valid way for both short- and long-term viewing. Paragraph 3.5. Measurement methods provides an extensive description of these different measurement methods.

3.4. Determinants: an empirical description

From 1952 to 1954, stereoscopic films were at the height of their popularity, with Hollywood producing more than 65 stereoscopic feature films. However, viewers' interest rapidly declined after this initial success. Part of the reason for this was increased competition from other immersive cinema formats. Undeniably, however, some of the problems with 3-D cinema appeared to be associated with problems of visual discomfort (IJsselsteijn, 2004). In the next section we describe factors that are thought to cause visual discomfort in stereoscopic displays nowadays. These factors are discussed from an empirical point of view in which a distinction between objective visual fatigue and subjective visual comfort is applied.

3.4.1. Excessive screen disparity

As discussed previously, sensory fusion limits can be remarkably small. Without vergence movements and for brief stimulus durations, fusion limits as small as 27 min of arc for crossed and 24 min of arc for uncrossed retinal disparity are found (Yeh and Silverstein, 1990). Many factors affect the limits of fusion, including eye movements, stimulus properties, temporal modulation of retinal disparity information, exposure duration, amount of illuminance and individual differences. The limits of fusion decrease with smaller, detailed and stationary objects and increase with larger, moving objects and the addition of peripheral objects to the fixation object (Howard and Rogers, 2002; Jones and Stephens, 1989; Patterson and Martin, 1992; Schor, Wood and Ogawa, 1984; Westheimer, 1994; Yeh and Silverstein, 1990). With longer stimulus durations and vergence eye movements retinal disparities as large as 4.93 degrees for crossed and 1.57 degrees for uncrossed disparity can be brought into fusion range without diplopia (Yeh and Silverstein, 1990).

However, the classical notion of Panum's fusional area has only limited applicability in establishing absolute limits for screen disparities in stereoscopic displays. A distinction between absolute and relative screen disparity is useful in this sense. The absolute screen disparity refers to a disparity-offset of the whole retinal image of one eye relative to the other (i.e., motoric fusion), whereas the relative screen disparity refers to the disparity differences between objects within the retinal images (i.e., sensory fusion). The absolute screen disparity can be large and can be overcome by appropriate vergence movements, yet clear, single binocular vision can only be perceived as long as the relative screen disparities remain within the fusion range.

3.4.2. Accommodation and vergence mismatch

The mismatch between accommodation and vergence arises due to an intrinsic conflict between the accommodative stimulus that remains fixed on the screen where the image is displayed most sharply, and the vergence stimulus that can fluctuate in depth depending on the degree and sign of screen disparity. Since accommodation and vergence are reflexively coupled mechanisms, their artificial de-coupling when viewing stereoscopic displays has often been theorized as a significant factor underlying the occurrence of visual discomfort (Emoto et al., 2005; Hoffman et al., 2008; Wann, Rushton and Mon-Williams, 1995; Okada, Ukai, Wolffsohn, Gilmartin, Iijima and Bando, 2006). Eadie et al., (2000) revealed that stereoscopic stimuli can initiate changes in the cross-link interaction between vergence and accommodation, i.e., altered AC/A and CA/C ratios, as well as in the tonic components. These changes can have negative consequences for clear and single binocular vision, because changes in the optical alignment of the eyes affect binocular fusion limits and depth perception (Semmlow and Heerema, 1979; Ciuffreda, 2002; Suryakumar and Bobier, 2004). Such alterations can last minutes or even hours, because re-adaptation to the real world is needed (Howard, 2002). Although it is argued that this process of decoupling accommodation and vergence induces visual fatigue, research reveals contradictory results in that accommodation does not remain focused on the screen, but shifts towards the reconstituted object (Ukai and Howarth, 2008; Inoue and Ohzu, 1997). It remains unclear, however, whether the shift of accommodation from the display plane was elicited by vergence-driven accommodation, or that it was a natural under-accommodation that occurs in most people during near work (Goss and Huifang, 1994). Hence, suspicions arise whether a conflict between accommodation and vergence occurs at all as a result of this mismatch, and how it is related to the DOF of the eye (Howarth, 1996). Figure 23 provides clarification. If screen disparity is increased, the retinal disparity of the reconstituted object surpasses Panum's fusional area. Vergence movements relocate the retinal disparity within Panum's fusional area and as such, increase fusion limits (i.e. motoric fusion). As a consequence, accommodation shifts away from the display under the influence of vergence-driven accommodation. As long as the accommodation shift remains within the DOF, accommodation is able to focus the reconstituted object sharply on the retina (Hiruma and T. Fukuda, 1993). If screen disparity is increased up to an amount at which the resulting retinal defocus cannot be accounted for by the DOF, negative accommodation feedback directs accommodation and vergence (via accommodative-vergence) towards the display, thus away from the reconstituted object. As such, the accommodation response conflicts with the vergence response. The accommodation-vergence system is able to cope to some degree with such a conflict, i.e., stereoscopic images are perceived sharply and fusion is preserved, but operates under stress and viewers experience visual discomfort. Especially in case of prolonged viewing, the visual discomfort can increase. The ranges of accommodation and vergence that can be achieved without any excessive errors in either direction are referred to as "the zone of clear single binocular vision" (Howard, 2002). If the conflict between the accommodation and vergence increases even more, three errors can occur: loss of accommodation resulting in a blurred image, loss of fusion resulting in double vision, or both.

3.4.3. Zone of comfortable viewing

The limits of the accommodative output under natural viewing conditions, i.e., range of DOF, concur with the range of fusion (Semmlow and Heerema, 1979; Pastoor, 1993; Wopking, 1995; Nagata, 1996). Objects at increasing distance from the fixation point are perceived as more blurred. As a consequence of this blur, diplopia is postponed, because the limits of fusion increase as a result of the decreased spatial frequency. In principle, if both visual systems complement each other in this manner, it is expected that their limits should match and together define a zone of comfortable viewing.

An accepted limit for DOF in optical power for a 3 mm pupil diameter (common under normal daylight conditions) and the eyes focused at infinity is one-third of a diopter (Hung, 2001; Hoffman et al., 2008). With respect to the revisited Panum's fusion area, i.e., under natural viewing conditions, retinal disparities beyond one degree (a conservative application of the 60 to 70 arcmin recommendation (Speranza, 2006; Wopking, 1995)) are assumed to cause visual discomfort (Iwasaki, Kubota and Tawara, 2009). This one degree is calculated from the characteristics of DOF (Wopking, 1995). It now serves as a rule-of-thumb, but it is acknowledged here as a limit for a zone of comfortable viewing, despite the fact that lower recommendations have also been reported (Woods et al., 1995; Jones, Lee, Holliman and Ezra, 2001; Reichelt, Häussler, Fütterer and Leister, 2010). This one degree limits the screen disparity, and as such imposes restrictions on the generation of 3-D content. In the case of 3-D TV, one popular acknowledged format for stereoscopic content is defined as a red-green-blue (RGB) image with one or more corresponding depth maps (Barenbrug, 2006; Hewage, Worral, Doga, Kodikara Arachchi and Kondoz, 2007; Redert, Berretty, Varekamp, van Geest and Bruijns, 2007). The first, and most important depth map is a grey-scale image, in which the grey value per pixel indicates the relative depth of each corresponding RGB pixel (secondary depth maps can contain additional depth information, e.g., occlusion information). The amount of resulting screen disparity can be set and altered by varying offset and gain factors, when rendering the left and right views calculated from the depth maps on the display.

For the vergence system a zone of comfort proposed by Percival could be considered as an alternative for the one-degree limit. It is defined as the middle third of the amount of binocular vergence with almost no change in accommodation, i.e., the middle third of "the zone of clear, single binocular vision" (Sheard, 1934). These zones, derived from Morgan's normal population norms (Peli, 1998 and 1999), are depicted in Figure 24, including the viewing zone covered by the one-degree disparity limit. The limits of "the zone of clear, single binocular vision" or motoric fusion limits are generally established by increasing prism load and measuring blur and break points, i.e., the prism loads at which blurred vision or diplopia is perceived respectively, in both convergent and divergent directions (Scheiman and Wick, 1994). Note that Figure 24 depicts two Percival areas of comfort due to a lack of consensus in the method of determining Percival's area within the display research area. In some research the break points are used as motoric fusion limits (Emoto et al., 2005), whereas in other research blur points are used (Peli, 1998). Percival himself stated the use of the blur points as limits of "the zone of clear, single binocular vision" (cited by Sheard, 1934).

Previous research already related Percival's area of comfort to stereoscopic viewing (Emoto et al., 2005; Hoffman et al. 2008), yet a few aspects reveal that Percival's area of comfort cannot be simply applied to stereoscopic displays. The first aspect is that Percival's area of comfort is determined by the use of prisms and obtaining stereoscopic content through the use of prisms perceptually differs from the use of stereoscopic displays (Yano et al., 2004; Peli, 1999). Prism loads relate to motoric fusion and change the whole visual field, i.e., absolute disparity, in contrast to only the screen disparity of certain objects in a stereoscopic image, i.e., relative disparity. As stated before, for 3-D TV applications with the RGB plus depth format, the screen disparity is calculated from depth maps containing relative depth information. A second aspect is that the size of Percival's area depends on the viewing distance as depicted in Figure 24. In this figure, the viewing distance or display plane is represented by Donders' line, i.e., the line that represents the perfect amount of vergence required for each level of accommodation for single binocular vision. When Percival's zone of comfort for large viewing distances is based on breakpoints according to Morgan's normal population norms (Peli, 1998), it does not include Donders' line. Using Morgan's normal population norms, the tolerances for prism loads for close viewing distances are larger than the tolerances for large viewing distances. Even though Percival's zone of comfort is normally determined per individual and other norms exist (Sheedy and Saladin, 1983), this reflection reveals that for a least a considerable part of the population, Percival's zone of comfort is not appropriate stereoscopic displays at large viewing distances. A third aspect is that people might adapt to changes in prism load, i.e., prism adaptation.

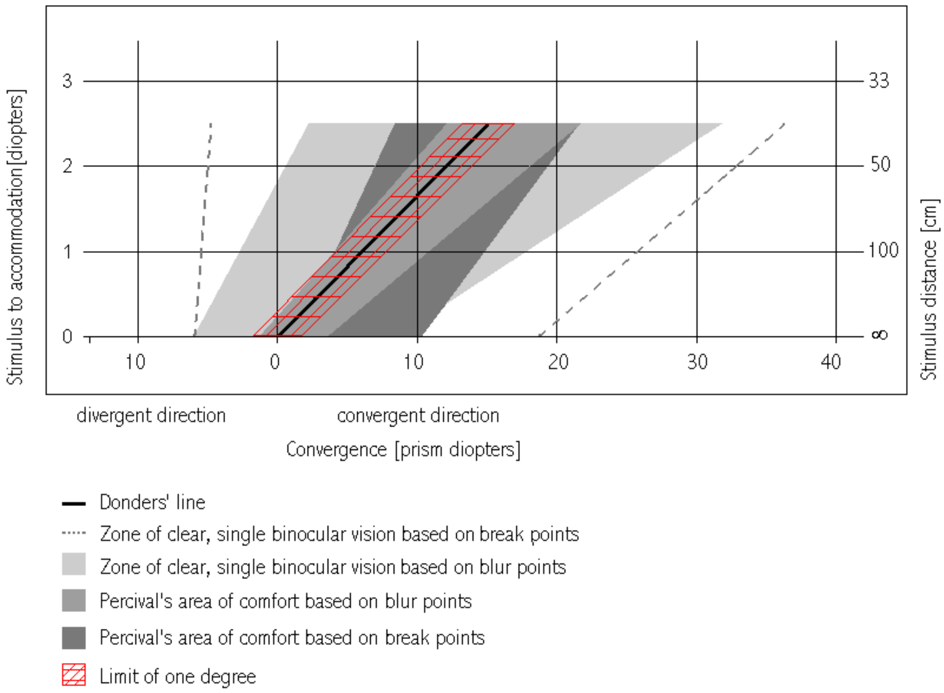


Figure 24. Different viewing zones with respect to comfortable viewing; "the zone of clear, single binocular vision"(Sheard, 1934),two different areas of comfort defined by Percival's criterion, one based on blur points (Sheard, 1934) and one based on break points (Emoto et al., 2005; Evans, 2007),and the zone formed by the one degree limit. The black solid line depicts Donders' line (Yano et al., 2004; Emoto et al., 2005; Hoffman et al., 2008).

We argue in favour of applying a screen disparity of one degree, in both divergent and convergent direction from the display plane, as a limit for a zone of comfort. Table 8 presents theoretical values for comfortable viewing expressed in distances for different viewing distances based on this limit. Note that although this limit is valid in theory, in practise no 3-D display can display the amount of depth at large viewing distances that results from one degree of disparity. Hence, the one-degree limit can be applied as a general limit of comfort for stereoscopic displays measured from the display plane, excluding the extensive list of factors that underlies the limit.

TABLE 8

LIMITS OF COMFORTABLE VIEWING AT DIFFERENT VIEWING DISTANCES CORRESPONDING TO ONE DEGREE OF SCREEN DISPARITY FOR BOTH CROSSED AND UNCROSSED DISPARITY. THE LIMITS SET THE AREA AROUND THE DISPLAY MEASURED FROM THE VIEWER.

VIEW.DISTANCE (MM)	Limits for comfortable viewing	
	Near (mm)	Far (mm)
500	440	580
1000	780	1400
2000	1300	4800
3000	1600	23000

To accept this one-degree limit as an applicable boundary for a zone of comfortable viewing, it is necessary to demonstrate and verify that stereoscopic image content beyond this limit results in asthenopia in contrast to within this limit. At larger screen disparities stereoscopic content is perceived sharply and fusion is preserved, though accompanied by the occurrence of asthenopia as a result of the increasing stress on the visual system, up to a point at which blur and double vision are perceived. A blurred image is expected to occur before a double image, as vergence seems to be dominant over accommodation, i.e., the visual system has a preference in avoiding diplopia before blurring (Edgar, 2007).

Beyond the zone of comfortable viewing

At increasing screen disparities beyond one degree, the oculomotor system operates under increasing stress to preserve fusion and provide sharply focused images. This statement was confirmed by Ukai and Kato (1999) who recorded the dynamic behaviour of the ocular near triad of participants viewing stereoscopic images. The screen disparity of a stereoscopic stimulus was increased stepwise from 0 to 1.6, 2.1 or 2.6 degrees. Vergence was evoked to preserve fusion and accommodation was elicited under the influence of vergence-accommodation away from the screen. This initial accommodation response, however, was followed by a correction in the opposite direction by the accommodation controller. For the step in screen disparity of 2.1 degrees, this correction was sufficient to correct the vergence response under the influence of accommodative-vergence. For the step in screen disparity of 2.6 degrees, however, the correction responses of accommodation and vergence repeated themselves and both systems became unstable and oscillated. Whether the conflict between accommodation and vergence resulted in double or blurred images was not verified, yet oscillations are indicative of fusion difficulty (Torri, Okada, Ukai, Wolffsohn and Gilmartin, 2008). Okada et al. (2006) revealed that when the accommodation-vergence system operates under stress, it continuously tries to find a more stable and less stressful state. Stereoscopic stimuli were varied by different levels of blur (i.e. accommodation) and different degrees of screen disparity (i.e. vergence). A shift in accommodation occurred towards the 3-D stimulus under the influence of vergence-driven accommodation that increased systematically with increased degrees of blur. This indicates a conflict between accommodation and vergence for sharp images displayed at screen disparities beyond the one-degree limit; the accommodation-vergence system is able to operate under stress, i.e., the CA/C and AC/A ratios adapt, and cross coupling still occurs, but continuously tries to resolve the stress. A follow-up study by Torii, Okada, Ukai, Wolffsohn and Gilmartin (2008) confirmed that viewing stereoscopic images with high spatial frequency components can be accompanied by dynamic oculomotor responses such as overshoots and oscillations, which can occur independently in accommodation and vergence systems and differ between participants. Recently, Fukushima, Torii, Ukai, Wolffsohn and Gilmartin (2009) attributed these differences between participants to CA/C ratio. More specifically, an initial convergence response induced by image disparity, generates a convergence-driven accommodation according to the CA/C ratio of the participants, after which the associated defocus steers accommodation to a balanced position between defocus- and convergence-induced accommodation. Hoffman et al. (2008) constructed a multi-focal 3-D display with separate left- and right-eye views per focal plane, enabling separate stimulation of vergence and accommodation for different focal distances. Stereoscopic stimuli were presented with various vergence and accommodation distances, from which two thirds of the distances were conflicting (ranging from 0.33 to 1.33 diopters). A questionnaire that followed an orientation detection task significantly revealed more visual discomfort for conflicting stimuli than for the non-conflicting ones. Nojiri, Yamanou, Hanazato and Okana (2003) 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. Objective measurements were not performed, therefore these finding could not be supported objectively. Yano et al. (2004) evaluated comfortable viewing for still images in relation to the range of screen disparity both subjectively, using a self-assessment test, and objectively, with pre- and post accommodation responses. The subjective evaluation revealed higher values for visual discomfort when images were displayed beyond one degree of screen disparity, which was confirmed by their objective measurements.

Within the zone of comfortable viewing

Within the zone of comfortable viewing, visual discomfort should not occur. Indeed, most stereoscopic stills are comfortable to view, nonetheless visual discomfort might occur as a consequence of much variation in screen disparity within this zone (Nojiri et al., 2003; Nojiri, Yamanoue, Ide, Yano and Okana, 2006). Yano, Ide, Mitsuhashi and Thwaites (2002) confirmed this finding with stereoscopic sequences (Yano, Ide, Mitsuhashi and Thwaites, 2002). A continuous subjective assessment revealed that visual discomfort was related to image content: visual comfort received local low evaluation scores for scenes with high degrees of screen disparity and high amounts of motion. In line with these findings, a follow-up experiment confirmed that discrete changes of motion in depth in stereoscopic sequences resulted in a decrease of the accommodation response and a significant decrease of visual comfort (Yano et al., 2004). Another study evaluated the effect of vergence load on Percival's area of comfort (Emoto et al., 2005). Though stereoscopic viewing through prisms differs from stereoscopic content on a 3-D display, it does affect the accommodation-vergence linkage. Vergence loads within Percival's area of comfort induced a lower degree of discomfort than loads outside this area. Temporally changing visual fields within this area, however, reduced the relative vergence limits, increased the latency of visually evoked cortical potentials and affected accommodation responses, but were subjectively not reported as yielding visual discomfort. To further clarify the effect on visual discomfort of changing screen disparity magnitudes in time, a relationship between the amount of screen disparity, object motion and visual comfort was verified (Speranza et al., 2006). Results showed that periodically changing screen disparity from crossed to uncrossed as well as the rate of this change negatively led to visual discomfort to a larger extent than the amount of disparity, even when it surpassed the one-degree limit.

It seems that visual discomfort increases when the demand on the oculomotor system increases as well. This occurs with screen disparities beyond one degree and with motion in depth within the zone of comfortable viewing. It is expected that prolonged viewing (exhausting the oculomotor system) and viewing at short distances (increasing of the relative exertion of accommodation) result in a further increase in visual demand, and thus in more visual discomfort. More detailed research is needed to clarify the relationship between accommodation and vergence with dynamic stereoscopic sequences within the DOF.

3.4.4. Stereoscopic distortions

Stereoscopic distortions result from several stages in the creation of 3-D content, namely content generation (choice of camera, camera configuration, 2-D-to-3-D conversion), coding and transmission (compression), rendering (multiple views rendered from a single view) and type of display. The literature describes several types of distortions that can induce visual discomfort and can occur simultaneously (Woods et al., 1995). Generation related distortions include keystone distortion, depth-plane curvature, puppet theatre effect, cardboard effect, and shear distortion. Display related distortions include picket fence effect, image flipping and crosstalk. They are not all discussed in detail here, as their technological causes and perceptual effects are well-understood. Recent detailed descriptions of these geometrical stereoscopic distortions are provided by Meesters et al. (2004) and IJsselsteijn et al. (2005). As crosstalk is an artifact that to some extent appears in nearly any 3-D display, it is briefly discussed separately.

Research mentioned crosstalk as the main display-related perceptual factor degrading image quality and causing visual discomfort (IJsselsteijn et al., 2005). Crosstalk is an artifact that results from the imperfect separation of the left and right eyes' view and as such, increases in impact with increasing screen disparity. It is used interchangeably with ghosting, though crosstalk denotes the electrical or optical mixing of left- and right-eye images (Siegel, 2001), which can result in perceived ghosting, but also in blurring. In some cases, however, crosstalk can also have some beneficial effect on image quality and visual comfort. Some autostereoscopic multi-view displays intentionally induce a certain amount of crosstalk to avoid a picket-fence effect (banding) and to minimize image flipping (the discrete transitions between neighbouring views) (van Berkel and Clarke, 1997). Small screen disparities limited to the fore- and background regions combined with crosstalk (up to 40%, i.e., 20% of each of the neighbouring views) are perceived as blur instead of ghosting (Siegel, 2001). Nonetheless, perception of depth is preserved (Seuntiëns, Meesters and IJsselsteijn, 2005). Thresholds of crosstalk increase as a consequence of broader spatial distribution of crosstalk, i.e, although the amount of crosstalk is the same, distributing it over multiple neighbouring views gradually smoothens the crosstalk to make it less visible (Kaptein and Heynderickx, 2007). These thresholds, however, also depend on the level of detail in images. Increasing detail in images, which is one of the main goals of imaging systems, i.e., optimizing image quality, has been found to decrease thresholds of crosstalk.

Furthermore, because crosstalk results in blurred objects to an extent related to their amount of screen disparity, it decreases the accommodation stimulus and as such, the accommodation-vergence conflict. On the other hand, blur is stated as one of the most important factors that determine viewing comfort (Kooi and Toet, 2004). In a worse-case scenario, unnatural blur can even facilitate or accelerate the development of accommodation difficulties or temporary nearsightedness (Peli, 1998). Hence, the optimal amount of crosstalk is still an issue of debate; the amount of induced depth should be a balance between annoying degrees of blur, image quality, perceived banding and clear transitions between views.

An artificial DOF

In real world situations, objects at distances both in front of and behind the fixation point are blurred to extents proportional to this distance, which if large enough, does not stimulate fusion. Blur in this sense can be defined as the perception of retinal defocus (Wang and Ciuffreda, 2006) and is a direct stimulus for accommodation. Sharpness enhancement is often implemented in display systems to improve image quality. Though a positive development, the lack of blur can cause visual discomfort due to different reasons. A stronger accommodation stimulus increases the accommodation-vergence conflict (Wann et al., 1995; Okada et al., 2006; Torii et al., 2008; Fukushima et al., 2009). In addition, objects with a screen disparity beyond the fusion limit still elicit an effort to fuse, whereas fusion is not possible due to the large retinal disparity (Talmi and Liu, 1999). And finally, the lack of blur removes the monocular depth cue of DOF (Peli, 1999).

Simulating DOF is said to minimize both these problems and to provide a more natural percept. Because limits of fusion increase with decreasing spatial frequency, artificially blurring images to a degree that corresponds to the amount of depth, can increase the range of fusion and reduce the conflict between accommodation and vergence. To conform to reality and avoid annoyance, objects fixated on must be displayed in full sharpness, whereas other regions must have a depth-dependent blur to preserve fusion of excessive parallax. This requires object-dependent depth information. Three essential steps are required for proper implementation of a simulated DOF: localization of the eye positions (Talmi and Liu, 1999), determination of the fixation point (Talmi and Liu, 1999) and implementation of blur filters to non-fixated layers (Blohm, Beldie, Schenke, Fazel and Pastoor, 1997). However, this procedure can also cause negative side effects. First, our visual system generally does not integrate retinal disparity and high amounts of blur, since they are active over different ranges (Mather and Smith, 2000). When the visual system is forced to do so, simulating DOF could lead to unnatural or uncomfortable viewing. Second, incorrect blurring of objects and edges can facilitate ambiguous depth perception. The amount of blur depends on the viewing distance and the polarity of the depth percept, i.e., in front and behind the fixation point. Different viewing distances and polarities can induce similar retinal defocus and as such, incorrect accommodation responses. And third, simulating such a DOF can have practical limitations with some autostereoscopic display technologies, e.g., in the case of multiple viewers that can concentrate on different parts of the image. Other research applied a different approach (Jones et al., 2001; Holliman, 2004). To avoid the entire tracking procedure another solution is to scale the scene depth range to our perceivable depth range. However, compressing or expanding the scene depth range can result in unnatural depth perception. An improved approach was introduced that compressed only the most outer regions, i.e., not the region of interest (Holliman, 2004). The solution has been implemented, but not yet evaluated perceptually.

3-D artifacts

To guarantee sufficient amounts of 3-D content for (auto)stereoscopic displays, (real-time) 2-D-to-3-D conversion is a promising method. This is especially true for digital television content, since research has demonstrated that generated depth only has to approach reality to create an acceptable 3-D percept (Meesters et al., 2004). Hence, development of these conversion algorithms is based on the assumption that geometrically accurate depth is not necessary and that a good depth impression on-screen will suffice. This quasi-depth-ordering process relies on assumptions, estimations and heuristic cues (Battiato, Curti, La Cascia, Tortora and Scordato, 2004; Redert, Berretty, Varekamp, van Geest and Bruijns, 2007; Tam, Vázquez and Speranza, 2007). These processes can result in artifacts that include spatial and temporal inconsistencies, e.g., objects or parts of objects that are assigned incorrect depth values and are, therefore, allocated to incorrect depth layers. This can lead to incorrect blurring and pixel rendering, and unnatural visualizations, e.g., flickering of (parts of) the image and turbulence around the edges.

Unnatural visualizations can also result from disocclusion. Image content that is unavailable in the original 2-D image because it is hidden behind occluding objects, can suddenly become visible in virtual views. Since no information of the occluded objects is available in the original image content, the missing areas (often referred to as *holes*), must be replaced with 'useful' colour information (Fehn, 2004). Different algorithms have been proposed for this hole-filling procedure (Fehn, 2004; Mark, McMillan and Bishop, 1997; Shade, Gortler, Li and Szeliski, 1998), yet all experience the same shortcoming, namely that the occluded area is never fully correct, but always interpolated from existing information. Hence, 2-D-to-3-D conversion cannot be fully accurate, and artifacts related specifically to the 2-D-to-3-D conversion and rendering process are likely to occur. Little is known about the impact of these artifacts on visual discomfort. In the case of misallocated objects for example, cue conflicts can occur between our stereopsis and other depth cues like familiarity. These conflicts are at least perceptually annoying, but when the visual system cannot satisfactorily resolve them, they are expected to cause visual fatigue as well.

3.5. Measurement methods

The indicators visual fatigue and visual discomfort are numerous and widespread (Sheedy et al. 2003; Blehm et al., 2005). They can be clustered into objective indicators for visual fatigue and subjective indicators for visual comfort. This section provides a more in-depth discussion of measurement methods and devices that are believed either to be suitable or promising in determining the degree of visual fatigue or visual discomfort.

3.5.1. Subjective measurement methods

Subjective assessment methods as a means to perceptually evaluate stereoscopic (as well as monoscopic) content are nowadays widely accepted and applied (Yano et al., 2004; Emoto et al., 2005; Speranza et al., 2006; Kooi and Toet, 2004). Visual discomfort and its dependence on individuals' self-appraisal must be evaluated on a perceptual basis (Meesters et al., 2004; Kooi and Toet, 2004). Three subjective methods can be distinguished, namely explorative studies, psychophysical scaling and questionnaires. According to Meesters et al. (2004) explorative studies can be used in the context of stereoscopic displays to 1) evoke unprimed perceptions, 2) evaluate the added value of stereoscopic displays both with and without predefined criteria, and 3) determine the attributes that underlie multidimensional concepts such as visual comfort. Psychophysical scaling enables engineers to enhance and optimize their systems based on quantified perceptual attributes such as image quality and visual comfort. Two types of applications can be distinguished, each with their own measurements methods. The first is performance-oriented, i.e., used to facilitate a certain task. The second is appreciation-oriented, i.e., used to establish a degree of appreciation. Recommendations for appreciation-oriented applications for stereoscopic displays are described in recommendations such as ITU-R BT.1438 (2000) and ITU-R BT.500 (2002).

Questionnaires have been extensively applied as a means to determine the degree of visual discomfort (Emoto et al., 2005; Emoto, Nojiri and Okano, 2004; Howarth, 1996; Sheedy et al., 2003). To our knowledge a generally accepted questionnaire that proved to be valid, sensitive, reliable and robust in determining the degree of visual discomfort of stereoscopic displays, has not yet been established. In clinical research, questionnaires enable evaluation of the degree of asthenopia due to visual deficits. In most cases these questionnaires are too extended for our purpose, since the assessment incorporates a wide range of mental, social and physiological aspects. In order to develop a questionnaire that measures the degree of visual discomfort caused by viewing stereoscopic content, consultation with clinical and eye care experts and interviews with users are required (Donovan, Brookes, Laidlaw, Hopper, Sparrow and Peters, 2003). Furthermore, questionnaires for evaluating the degree of specific visual deficits resulting in visual discomfort must also be taken into account. Sheedy et al. (2003) developed a questionnaire to measure the degree of asthenopia, but it is not specifically related to stereoscopic displays. We believe that any questionnaire evaluating stereoscopic content should incorporate as a minimum all the items that have been used in Sheedy et al.'s questionnaire: tired eyes, uncomfortable vision, headache, ache in or behind the eyes, eye irritation, pulling feeling of the eyes, blurred vision, dryness of the eyes, burning eyes, stress, neck pain and watery eyes. Depending on the purpose and application (e.g., stereoscopic computer games (Häkkinen, Pölönen, Takatalo, Nyman, 2006) or stereoscopic mobile phones usage (Häkkinen, 2004), it might be useful to include additional background information such as previous experience with similar applications or amount of near work during a typical day.

3.5.2. Objective measurement methods

The many indicators for visual fatigue are related to alterations in various characteristics of different visual functions (e.g., accommodative and vergence responses, pupillary dynamics, AC/A and CA/C ratios, fusion reserves, visual and stereo acuity and heterophoria). Alterations to these indicators can be quantified by implementing three different classes of measurements. The first class includes optometric instrument-based measurements that directly measure the indicators with optical instruments such as refractometers and pupil trackers. The second class consists of optometric clinical-based measurements that indirectly measure the indicators via prisms, lenses or vision charts. The third class contains brain-activity measurements in which indicators are measured as a function of brain activity.

Optometric instrument based measurements

In many studies, optometric devices have been applied in pre- and post-tests to determine the amount of change of an indicator for visual fatigue as a result of viewing stereoscopic content (Emoto et al., 2005; Yano et al., 2004; Yano et al., 2002). Binocular single vision and asthenopia have been related, however, to various aspects of the dynamics of the ocular triad. It is difficult to draw solid conclusions without simultaneous, continuous and direct measurement of the ocular triad of participants who view stereoscopic content (Ukai and Kato, 2002; Suryakumar and Bobier, 2004; Okada et al., 2006). A variety of commercially available oculomotor measurement devices are able to measure different parameters of the oculomotor system. The most familiar one is the autorefractor: an effective tool for measuring various aspects of the dynamic accommodative response and the objective refractive error of the eye. A major drawback of refractors is the inability to simultaneously measure the oculomotor triad dynamics (Suryakumar, Meyers, Erving and Bobier, 2007). Hunt, Wolfssohn and Gilmartin (2002) address photoretinoscopy, more specifically the PowerRefractor, as unique in allowing measurement of the oculomotor triad in both eyes simultaneously, continuously and remotely in a non-obtrusive manner. Comparisons with clinical methods and the more established open view autorefractors (e.g., the Nidek AR600-A and the Shin-Nippon SRW-5000) showed a similar average accuracy of accommodation measurement (Hunt et al., 2002; Allen, Radhakrishnan, O'Leary, 2003). The PowerRefractor has the advantage of open viewing, i.e., it allows an open field of view for natural binocular viewing without obtrusion of the device. As such, it can be used without a bite-bar or head strip, and allows easier use for measurement on visual systems of children or other less cooperative participants as well as a wider range of experimental applications. However, the accuracy of approximately two degrees of disparity for vergence measurements is too coarse to measure the effects of changes within the zone of comfortable viewing. Other solutions are 'simply' to combine an autorefractor with an eye tracking device to simultaneously record vergence eye movements and accommodation dynamics. Okada et al. (2006) applied a tracker on the left eye and a Shin-Nippon SRW-5000 on the right eye and Suryakumar, Meyers, Erving and Bobier (2003 & 2007) used a stereo eye tracker in synchronization with a custom build photorefractor allowing simultaneous high speed measurements of both vergence and accommodation. In general, optometric measurements are costly, time-consuming and are usually conducted with only small numbers of participants.

Optometric clinically based measurements

Clinical diagnoses to investigate and diagnose the degree of (binocular) visual anomalies are applied to patients who suffer from asthenopic complaints such as headaches or problems with focusing. These measurements are relatively cheap, concise, non-interventional, quantitative with a high sensitivity and specificity and applicable to a large group of participants. The number and diversity of clinical tests to detect specific visual deficits is enormous (Evans, 2007; Scheiman and Wick, 1994). However, due to an expected rapid reduction in the degree of visual fatigue after viewing stereoscopic content, only clinical tests that are able to diagnose the degree of visual fatigue with a fast measurement are useful. The following measurement protocol is proposed: 1) describe the general visual function of the participants in the unaffected state with the aim of establishing individual differences in visual aberrations and sensitivities, and 2) apply a set of clinical pre- and post-tests to determine possible alteration of the visual functions, i.e., the difference between the unaffected (pre-test) and the affected (post-test) state of certain visual functions, as a result of viewing stereoscopic displays.

The first step is a thorough optometric screening of participants in order to distinguish participants with normal vision (i.e., vision without any associated visual discomfort of non-strabismic binocular anomalies in normal viewing situations; see section 6.4) from those with visual deficits or increased sensitivity to visual discomfort. Both groups can serve different purposes; the group with visual deficits is more susceptible to visual fatigue, which is interesting from a clinical point of view. The group with normal vision reflects the visual behaviour of the majority of the population, which is interesting from a consumer's point of view. The screening also serves as a potential clarification for individual differences in the subsequent pre- and post-tests. It should include indicators such as visual acuity, stereo acuity, convergence ability and AC/A ratio.

The second step, i.e., a reliable and valid set of clinical pre- and post-tests, to our knowledge has not yet been established for stereoscopic displays. Assuming a difference between monoscopic and stereoscopic viewing, not all tests are equally appropriate to diagnose the effect of stereoscopic viewing. A few aspects should be accounted for when composing such a set. First, in order to address the impact of binocular depth on the visual system, a test should be able to distinguish conventional monoscopic viewing conditions from stereoscopic viewing conditions. Second, the tests should be relatively fast, as the recovery of the binocular visual system is usually quite rapid, which constrains the length of the test. Note that the recovery trajectory of the eyes after prolonged exposure to a stressful stimulus can be in itself be indicative of its functional plasticity and the severity of the visual strain. Thus, multiple measurements at different post-stimulus intervals can be needed. Third, ideally the set of tests should be applicable to all different types of displays, including autostereoscopic systems, and systems based on polaroid or shutter glasses. However, it is highly plausible that different displays, i.e., different principles of generating depth, differently affect the visual system. For example, measuring fusional amplitudes can be less relevant than measuring accommodation responses for autostereoscopic displays, which as a result of crosstalk, are limited in their amount of depth, but introduce high amounts of blur. Fourth, the display application should be taken into account. Vergence measurements are suitable for short and large viewing distances, i.e., desktop and TV applications respectively, whereas accommodation measurements are only suitable for desktop applications. And lastly, measurements themselves should not require too much visual effort or induce visual fatigue or visual discomfort on their own.

Some tests applied in the screening are expected to be applicable as clinical pre- and post-tests as well, e.g., binocular visual acuity or stereo acuity. For specific vergence measurements the clinical tests can include: 1) fusional reserves, which denotes the amount of vergence, both diverged and converged, that can be endured before blurring or double vision occurs while keeping accommodation constant, 2) vergence facility, which is the ability of the vergence system to respond efficiently and accurately to changing demands over time, and 3) fixation disparity, which refers to accuracy of vergence and relates visual stress to prism strength necessary to redirect perceived objects to corresponding parts of the retina (Emoto et al., 2004 & 2005; Evans, 2007; Scheiman and Wick, 1994). Specific accommodation measurements can include: 1) accommodation amplitude, which denotes the maximal range of accommodation, e.g., push-up method of Donders, 2) accommodation facility, which is the ability of the accommodative visual system to respond efficiently and accurately to changing demands over time, and 3) accommodation accuracy, which describes the difference between the accommodation necessary for a certain viewing distance and the measured accommodation (Emoto et al., 2005; Evans, 2007; Scheiman and Wick, 1994).

Brain activity measures

All sensory and high-level cognitive information is processed in the brain. As such, the neuronal activity in the brain also reflects visual fatigue as a consequence of viewing stereoscopic content. Brain activity measurements provide information on changes in brain activity as a result of simultaneous behaviour changes and provide knowledge that extends from better understanding of perceptual and cognitive processes to characterization of a variety of pathologies including specific visual disabilities (Nichols and Newsome, 1999; Pouratian, Sheth, Martin and Toga, 2003).

Most brain activity research related to depth perception concentrates on fundamental issues, such as identifying the exact pathways for binocular vision (Blake and Logothesis, 2002; Cumming and DeAngelis, 2001; Parker, 2007). Little work has been done on depth perception of stereoscopic content on 3-D displays and related aspects such as visual fatigue (Ciuffreda, 2009). This can be attributed to the fact that visual fatigue refers to multiple conflicting interactive visual modalities and that other evaluation tools are more practical. The few studies that have applied brain activity measurements, however, have revealed interesting results.

Emoto et al. (2005) used electroencephalography (EEG) to measure visually evoked cortical potentials, i.e., an potential evoked by sensory stimulation of the visual field. Visually evoked cortical potentials reflect fatigue of the interrelated extra-ocular muscles, intra-ocular muscles and central nerve of the brain. The P100 latency (positive component at approximately 100 ms latency) of the visually evoked cortical potential was used as a fatigue index. Delays of the P100 latency were found between pre- and post-exposure to different parallax settings. For temporally changing parallax, the delays were significant. Furthermore, as stated before, high correlations were found between P100 latencies and relative vergence limits. Li, Seo, Kham and Lee (2008) used background EEG and event related potentials to measure visual fatigue. The frequency spectrum of the background EEG signals is known to indicate the state of stress, i.e., higher frequencies starting at ±12 Hz denote stressful situations. Though stressful situations also delay the P300 latency of the event related potentials, they found that the delay was much stronger for the P700 latency. Results revealed that the power of the spectrum of the background EEG as well as the delay in the P700 latency depended on binocular parallax and presentation time, which was confirmed by subjective assessments. Hence, delays in the transmission of visual information measured with EEG seem to be an appropriate measure for visual fatigue.

To overcome limitations and exploit advantages in sensitivity and specificity, information with high-quality temporal resolutions (e.g. magneto encephalography (MEG) and EEG) can be superimposed on information with high-quality spatial resolution (e.g. functional magnetic resonance imaging (fMRI)) (Wandell and Dougherty, 2006; Dale and Halgren, 2001). Worth mentioning is that the magnetic field of the fMRI is much stronger than that of the brain, and as a consequence fMRI measurements cannot be performed with MEG and EEG measurements simultaneously. In addition, the added value of such high temporal information is questionable, since visual fatigue and accompanying symptoms correspond to much lower temporal resolutions. Nonetheless, Hagura, Nakajima, Owaki and Takeda (2006) combined brain activity data of high spatial and temporal resolution as a measurement tool to detect visual fatigue during 3-D experiments. In a preliminary study dipole data acquired by MEG was superimposed on a 3-D model composed by fMRI. As a result of viewing random dot stereograms, brain activity in the back left side of the brain was revealed. The isocontour maps of the dipole activity differed for different viewing periods. However, the isocontour maps were not clear enough to locate and identify exact activated locations. Hence, it seems that further investigation is required to apply MEG and fMRI as brain activity measurements for visual fatigue.

3.6. Discussion

Visual fatigue and visual discomfort are related to many different aspects of the human visual system, thus remain somewhat ambiguous concepts when used in a general sense. However, for the purpose of our current review, with respect to stereoscopic displays we define visual fatigue as a decrease in the performance of the binocular visual system as a consequence of physiological strain or stress resulting from excessive exertion of the visual system. It is the unnatural state of the visual system as a result from stereoscopic stimuli that can be objectively quantified in theory. Visual comfort is the subjective counterpart of visual fatigue and when formulated in this way, is the sensation that is assumed to reflect visual fatigue. With these definitions, I acknowledge that visual fatigue has been previously used as a subjective and objective measure (e.g.Yano et al., 2004; Yano et al. 2004; Suyama et al. 2005), yet distinguishes between the two concepts for clarification reasons. Note that a change within the binocular visual system itself does not necessarily indicate visual fatigue. The binocular visual system has some degree of plasticity and is able to adapt to altered viewing conditions, e.g., prism adaptation. In order to distinguish clinically significant visual fatigue from unproblematic, functional adaptations of the visual system, we need to establish relationships with subjective indicators of visual discomfort and monitor potential damage of the visual system as a result of prolonged viewing. The occurrence of visual discomfort alone, however, can be sufficient reason for further research. Firstly, consumers will be reluctant to purchase a display that induces visual discomfort, even if the visual discomfort is harmless in terms of visual fatigue. Secondly and more importantly, absence of visual fatigue related to short-term viewing (e.g., five minutes) might still compromise the binocular visual system when longer viewing durations are used (e.g., two hours) for extended periods (e.g., weeks, months or years). As such, carefully conducted long-term evaluations will be necessary to ensure that prolonged stereoscopic viewing does not induce any adverse side-effects to the visual system. Appropriately developed and validated questionnaires or other self-report measures can provide subjective indicators, provided they are proven to be sensitive, reliable, valid and robust. Their subsequent application in evaluative settings is relatively easy. Visual fatigue, however, in most cases concerns measurements with optometric devices on the visual system that are generally costly, time-consuming and are usually conducted with only small numbers of participants, making the results less reliable. Furthermore, optometric devices that measure all the modalities of the ocular triad are not yet commercially available and have to be custom-built. Brain activity measurements such as EEG, MEG and fMRI, have received increasing attention in the last decade, provide an interesting framework for cognitive neuroscience and a promising tool for researching the fundamental nature of asthenopia, yet remain impractical, and for most research facilities, far too costly for psychophysical experiments. Nonetheless, e.g. EEG measurements provided promising results in detecting visual fatigue. Clinical measurement methods on the other hand, are relatively cheap, concise, non-interventional, quantitative with a high sensitivity and specificity and applicable to a large group of participants. More research, however, is needed to determine which specific clinical methods can be used to quantify the degree of visual fatigue from stereoscopic displays. Multiple objective indicators are argued for the evaluation of visual fatigue since a single underlying factor, e.g., vergence insufficiency, can stimulate different anatomical locations and result in different sensations. Combined measurements of EEG and clinical methods provide an appropriate framework to measure visual fatigue, since latencies in EEGs and relative vergence limits correlate.

Ideally, we would like to arrive at a general and easily applicable indicator of visual fatigue and visual discomfort. When a robust relationship is established between visual discomfort and visual fatigue indicators, one might be used to substitute for the other, where appropriate. This would allow study of large groups of participants using easily applicable visual comfort measures. Moreover, it would apply to children as well, who can have some difficulties in filling in questionnaires. This latter group is of particular importance as they are expected to spend much time using 3-D applications, yet whose developing visual systems have not been extensively studied in relation to their physiological responses to 3-D television or gaming applications.

With respect to a zone of comfortable viewing, Percival's area of comfort seems not to be appropriate for stereoscopic displays. This area is determined by the use of prisms, which create stereoscopic content that perceptually differs from content on stereoscopic displays. We support the use of a maximum screen disparity that corresponds to a retinal disparity of one degree as a limit for a zone of comfortable viewing. Though this zone often serves as a rule-of-thumb, we acknowledge it here as a limit for a zone of comfortable viewing, analogous to auditory limits, i.e., it is not recommended to set the volume to the maximum level for extended periods. This still allows satisfactory depth perception for 3-D TV applications, though no 3-D display can display the amount of depth at large viewing distances that results from one degree of disparity. If viewers do not have some form of a binocular anomaly, the tolerances in our fusion and accommodation-vergence systems are able to adapt to conflicts within one degree of screen disparity (see section 6.4). Hence, this zone appears to prevent the 'classical' causes of visual discomfort, i.e., excessive screen disparity and accommodation-vergence conflict, from being perceptually annoying. Fusion is possible and blur is not perceived, hence, stereoscopic viewing should be comfortable within this limit. Beyond this limit clear and single binocular vision is still possible, yet not comfortable, up to a point at which blur and double vision are perceived. Hence, peak screen disparities can be induced in stereoscopic movies or games to increase the 3-D experience, but not too often or for extended periods. The description of the zone of comfortable viewing currently is dichotomous: a region that is comfortable and everything outside of that region is uncomfortable. I acknowledge that this description is certainly an over-simplification. Conflicts that are just outside the comfort zone are surely less than conflicts that are way outside the zone. Furthermore, some conflicts within the nominal zone are surely less comfortable than others. Hence, it makes more sense to describe the zone of comfort as continuous.

Even within the one-degree limit, however, there are certain factors that can still contribute to visual discomfort. To date, we have identified three such factors to be the most pertinent ones. The first factor is fast motion in depth. Though the mismatch between accommodation and vergence should not result in a conflict within one degree of disparity, continuously stressing the linkage by objects with motion and changing screen disparity can exhaust the AC linkage. This can cause visual discomfort that is expected to become more severe with prolonged viewing and at close viewing distances. The second factor concerns 3-D artifacts, resulting from insufficient depth information in the incoming data signal, yielding spatial and temporal inconsistencies. Such artifacts have not been subjected to much research yet, though inconsistencies, such as conflicts between depth cues and geometrical distortions have already proved to cause annoyance and visual discomfort. The third factor concerns unnatural blur. Blur can cause ambiguous and unnatural depth percepts. The lack of blur, i.e., an entirely sharp image, can reduce the range of fusion, thereby hampering fusion of some content and cause depth cue conflicts. Additionally, it can strengthen the accommodation stimulus, thereby increasing the mismatch between accommodation and vergence. A surplus of blur resulting from crosstalk, 2-D-to-3-D conversion and artificially induced DOF, causes annoyance, visual discomfort, can result in depth cue conflicts and can even facilitate or accelerate the development of accommodation difficulties or temporary nearsightedness.

3.7. Conclusion

In this chapter we have reviewed the concept of visual fatigue and its subjective counterpart, visual discomfort, in relation to stereoscopic display technology and image generation. To guarantee visual comfort in consumer applications, such as stereoscopic television, it is recommended to adhere to a limit of 'one degree of disparity', which still allows sufficient depth rendering for most application purposes. Within this zone of comfortable viewing, visual discomfort can still occur to an extent, however, which is likely to be caused by one or more of the following three factors: (1) fast motion in depth which assumably temporally changes demand of accommodation-vergence linkage, e.g., by (2) 3-D artifacts resulting from insufficient depth information in the incoming data signal yielding spatial and temporal inconsistencies, and (3) unnatural blur. In order to adequately characterize and understand visual fatigue and visual discomfort, multiple types of measurements, both objective and subjective, are needed.

-CHAPTER 4-

The impact of video characteristics

"Dolezal also reported that he was totally unable to follow the action in movies. Perhaps my inability to take in the whole movie screen at once explained why I had a hard time following the plots in many films. I tended to nod off when the action got complicated, like when the Mir space station blew up in the movie Armageddon. The faster the action in a movie, the quicker I started to snore, much to the amusement of my children….*
…. I like Star Wars, but until that night, I couldn't understand my family's fascination with the special effects. I was overwhelmed by the sense of space and volume created in the movie. Scenes of spaceships flying through the universe were fantastic! Skilled cinematographers had used monocular depth and motion cues to create scenes on the flat, two-dimensional movie screen that suggested dramatic depth. Before my vision transformed, I could not experience this sense of space and volume while watching a movie because I had never experienced this sense of space and volume in real life."

Susan Barry (Barry, 2009)

* author of *Living in a World Transformed*, in which he reflects upon his experiences perceiving the world without peripherical vision by wearing patches that cover part of his eyes for a week."

ABSTRACT

Some viewers of stereoscopic content experience visual discomfort that is intensified by certain video characteristics such as rapid in-scene motion and large changes in disparity. Two experiments (experiments 4 and 5) were designed with the primary objective of relating the impact of these video characteristics to the assessment of visual discomfort. In addition, experiment 4 compares a continuous assessment method with other assessment methods of visual comfort, and experiment 5 investigates the impact of subtitles on visual comfort.

Three 3-D movies were assessed in terms of visual comfort via a continuous assessment. The continuous assessment scores were directly compared to video characteristics that were derived from the 3-D movies. Additional assessment methods in experiment 4 included the assessment of six 10-second sequences captured from the 3-D movie and a single retrospective assessment of the entire 3-D movie. The two 3-D movies in experiment 5 were shown with or without subtitles.

Results show that the visual comfort of stereoscopic scenes can be predicted as a linear combination of screen disparity range and offset, changing screen disparity, and lateral motion. The specific contributions of these characteristics depend on the scene, yet more complex models are required to extend the comfort prediction to entire movies, incorporating different scenes. In addition, the results of experiment 5 reveal that subtitles required additional effort to keep vision comfortable and the results of experiment 4 show that the correlation between the assessment of the 10-second sequences captured from the 3-D movie and their corresponding parts within the continuous assessment is low, whereas the correlation between the retrospective assessment and the mean of the continuous assessment score over scene parts with a high screen disparity is higher.

Experiment 4 is based on:

Lambooij, M., IJsselsteijn, W. A., and Heynderickx, I. (2011). Visual discomfort of 3-D TV assessment methods and modelling. *Display: special issue image safety*. 32: 209-218.

Experiment 5 is based on:

Lambooij, M., Murdoch, M., IJsselsteijn, W. A., and Heynderickx, I. (2011). The impact of video characteristics and subtitles on the visual comfort of 3-D TV. *Displays*. submitted for publication.

4.1. Introduction

As the review in Chapter 3 demonstrates, in the last decade research concerning visual discomfort has gained considerable attention (Emoto et al., 2004, 2005; Hoffman et al., 2008; Kooi and Toet, 2005; Peli, 1998 and 1999; Speranza et al., 2006; Yano et al., 2004). A mismatch in accommodation and vergence in the human eye when watching stereoscopic content is often mentioned as a possible cause for visual discomfort. If viewers do not have some form of a binocular anomaly (Evans, 2007; Peli, 1999), the tolerances in our fusion and accommodation-vergence systems are able to adapt to conflicts within one degree of screen disparity. Hence, to guarantee a comfortable 3-D viewing experience it is recommended to adhere to the 'one degree of screen disparity' rule of thumb, which defines a zone of comfortable viewing. For most application purposes, e.g., 3-D movies and games and 3-D mobile phones, this limit allows satisfactory depth rendering. However, even within the one-degree limit, there are certain factors that can still contribute to visual discomfort. To date, we have identified three such factors that all relate to certain video characteristics: (1) fast motion in depth, (2) 3-D artifacts resulting from insufficient depth information in the incoming data signal, yielding spatial and temporal inconsistencies, and (3) unnatural amounts of blur. For a more extensive description of these factors I refer to Chapter 3. The perceptual impact of the three factors mentioned above is directly related to the amount of screen disparity or the amount of motion (Speranza et al., 2006; Yano et al., 2002), and as these quantities change, their impact fluctuates in time. A larger screen disparity puts more strain on the accommodation-vergence linkage, makes 3-D artifacts more visible and generates more blur. Higher speeds of in depth motion also stress the accommodation-vergence linkage more, whereas higher horizontal speeds make the viewer more sensitive to 3-D artifacts. It is also assumed that the degree of visual discomfort increases during prolonged viewing (Peli, 1998), which to our knowledge has never been confirmed. There are studies that explore long-term effects of visual fatigue in general, e.g., experienced when viewing chromatic displays (Matthews, Lovasik and Mertins, 1989) or while driving (Sullivan, 2008), but not many studies have explored long-term effects in relation to screen disparity. Where they do exist, their results are slightly contradictory; Pölönen et al. (2009) detected very moderate visual discomfort after the cinema movie U2-3-D, whereas Kuze and Ukai (2008) found significantly more visual discomfort after the movie 'Spy Kids: game over' in 3-D than in 2-D.

A relevant method to evaluate visual discomfort over time is to request viewers to assess the content continuously. This evaluation method requires participants to provide a real-time subjective rating of visual discomfort, using a slider or dial, from which the position is sampled at a fairly high frequency (e.g., 1-10 Hz) (Biocca, David and West, 1994). The advantage of this evaluation method is that the continuous set of ratings can be correlated with specific video characteristics that are known to be related to the pertinent determinants of visual discomfort. Fitting the continuous assessment scores of visual discomfort to these time-varying video characteristics provides more information on the impact of these pertinent determinants. The disadvantage of this evaluation method is that the duration of the assessment should be limited to a maximum of 30 minutes as stated within the ITU-recommendations (ITU, 2002). This is due to the possibility of strong visual fatigue caused by the assessment task itself. Since we also believe that concentration loss and mental fatigue might have an impact when assessing longer sequences, other assessment methods can be required for the perceived visual discomfort of e.g. feature-length 3-D movies. In the current research we try to gain insight into how people experience visual discomfort while watching 3-D movies and to what extent specific video characteristics induce visual discomfort in certain scenes.

4.2. Assessment methods

Subjective assessment methods as a means to evaluate stereoscopic as well as monoscopic content are nowadays widely accepted and applied. The Single Stimulus Quality Evaluation (SSQE) has been proven to be a valid method to obtain a quality judgment of a single (still) stimulus, but also to obtain continuous time-varying judgments of moving sequences. This latter method, referred to as Single Stimulus Continuous Quality Evaluation (SSCQE), is part of the ITU BT-500 recommendations (ITU, 2002), and is also mentioned in the ITU-R BT.1438, which specifically reflects evaluation of stereoscopic content (ITU, 2000). For an overview of its application, advantages and disadvantages, I refer to Biocca et al. (1994). The SSCQE was proposed by Hamberg and de Ridder to continuously evaluate the perceived quality of 2-D video sequences, 20 minutes in duration (Hamberg and de Ridder, 1995). IJsselsteijn, de Ridder, Hamberg, Bouwhuis and Freeman (1998) were the first to apply the SSCQE method to stereoscopic picture evaluation, continuously assessing presence, depth and naturalness for a stereoscopic TV programme (see also IJsselsteijn, 2004); other authors have applied the method since that time as well to evaluate 3-D image content (e.g., Yano et al., 2002). In sum, this method has been applied to assess different aspects of both 2-D and 3-D image content (Aldridge, Davidhoff, Ghanbari, Hands, Pearson, 1995; Aldridge, Hands, Pearson and Lodge, 1998; Hamberg and de Ridder, 1995; IJsselsteijn et al., 1998). The results show that continuous assessment is a consistent and reliable method to measure both 2-D and 3-D content in terms of image quality, naturalness, depth and presence. Moreover, continuous assessment has been demonstrated to be free from drift and fatigue effects for relatively short sequences (Aldridge et al., 1998). The implementation of continuous assessment, however, can also be limited to the assessment of fairly short sequences of e.g. ± 30 minutes, since fatigue, concentration loss and mental fatigue as a consequence of the assessment task can start to have an impact when judging feature-length movies of e.g., 2 hours (ITU, 2002).

Hence, for the assessment of feature-length movies, only two assessment methods mentioned in the ITU BT-500 recommendations that require less effort for the participants, both mentally and physically, exist (ITU, 2002). The first one is a retrospective assessment of the movie that should reflect the degree of visual discomfort of the entire movie. Due to the limited capacity of human working memory, however, participants form one way or another a temporally weighted average over the whole movie. Some research has shown that in this temporally weighted average, maximum weight is given to the last part, referred to as the *recency effect* (Aldridge et al., 1995). Additional research also revealed a *negative peak effect*, in which retrospective assessment is affected by the severity of negative peaks in the continuous assessment (Aldridge et al., 1998). Hence, it can be concluded that parts that induce more momentary visual discomfort as well as the last part of a movie can have a relatively large impact on the overall retrospective assessment, but the actual impact of the recency effect, the negative peaks and the rest of the movie on the retrospective assessment has not yet been fully established.

The second method is the assessment of multiple short sequences that are captured from different parts of the movie. Since the presence and severity of determinants of visual discomfort can differ substantially between scenes, a set of short sequences that incorporates these time-variant determinants could be representative for the entire movie.

Possible drawbacks of using short sequences, instead of the whole movie, are effects that are only present in the entire movie, such as: 1) the occurrence of visual discomfort as a result of prolonged viewing, 2) recalibration of the internal assessment scale during the 3-D movie, i.e., participants might alter the assessment criterion, and 3) context effects, e.g., preceding scenes affect the perception of following scenes.

4.3. Experiment 4: assessment methods and modelling

Within one degree of screen disparity visual discomfort can be induced by the three determinants mentioned in chapter 3 (excessive demand on the AC linkage, blur and 3-D artifacts), that are all related to some extent to specific video characteristics. The perceptual impact of excessive demand on the AC linkage is directly related to the amount of motion in depth, and so, to changes in screen disparity. The perceptual impact of 3-D artifacts and unnatural blur is directly related to the amount of screen disparity. Visual discomfort can also result from temporal inconsistencies in 3-D artifacts, the visibility of which is related to the amount of motion in the scene. Hence, one objective of our study is to determine the impact of motion, screen disparity and change in screen disparity on visual discomfort when assessing stereoscopic content. Time-variant video characteristics, such as derivatives of motion and screen disparity, are extracted from the image material with motion and depth estimation algorithms. The moment-to-moment values of these video characteristics are correlated with continuous visual comfort scores.

The second objective of our study is to compare the continuous assessment scores to the results of other assessment methods, such as the retrospective assessment and the assessment of multiple short sequences captured from the 3-D movie. The ITU recommendation implies that the assessment of short sequences of a 3-D movie or a retrospective assessment of visual discomfort is sufficient to generalize comfort experienced over an entire 3-D movie [22]. Consequently, a continuous assessment, which provides more detailed information yet requires more effort, time and analysis, is not necessary. In order to evaluate this statement, we wanted to compare visual discomfort experienced after the movie and for short sequences within the movie to the continuous scores given to these same scenes while watching the whole movie. To examine how the retrospective assessment method reflects visual discomfort with respect to the recency and negative peak effects, the amount of screen disparity was varied during the 3-D movie.

Hence, the first hypothesis states that the amount of motion, screen disparity and changes in screen disparity can be used to predict visual discomfort. The second hypothesis states that assessment of short sequences captured from the 3-D movie reflects the presence and severity of determinants of visual discomfort. The third hypothesis states that a retrospective assessment incorporates to some extent the recency effect and the negative peak effect.

4.3.1. Experimental set-up

Design

A 24-minute 3-D movie was evaluated in terms of visual comfort with three different assessment methods used for all participants: (1) a single assessment of six 10-second sequences captured from the 3-D movie (SA), (2) a continuous assessment of the 3-D movie (COA) and (3) a retrospective assessment of the entire 3-D movie (RA). Since in the specifications of the ITU BT-500 recommendations short sequences are not preceded by an entire movie, we intentionally chose to keep the order of the different measurement methods the same over all subjects, in order to avoid effects of mixing up the order in scenes between the short sequences and the entire movie. The order of the six short sequences was randomised. The initial maximum screen disparity of the 3-D movie was varied between participants (high and low IniDisp) and was halved for twelve participants after 70% of the movie had elapsed (full and half EndDisp). The experiment was a 2 x 2 (IniDisp x EndDisp) between-subjects design for 24 participants, divided over the conditions as outlined in Table 9.

TABLE 9
NUMBER OF PARTICIPANTS ASSIGNED TO SCREEN DISPARITY SETTINGS

		IniDisp	
		high	low
EndDisp	full	8	4
	half	9	3

Participants

Two female and twenty-two male, employees as well as graduate students working in a research environment, participated. Some of the participants (n=7) had previous experience with 3-D content. The average age was 31 years, ranging from 24 to 53, and all had a good visual acuity of ≥ 1.25 (tested with the Landolt C-test) and a good stereo acuity of ≤ 60 arc seconds (tested with the RANDOT stereo test).

Stimuli

The 24-minute 3-D movie that was used was converted from the feature-length stereoscopic HD movie 'Spy Kids 3-D: game over' to the WOWxv 2-D-plus-Depth format, i.e., a red-green-blue (RGB) image with a corresponding depth map both of which have a resolution of 960 x 544 (Redert et al., 2007). Figure 25 depicts a screen shot to illustrate the format.

Figure 25. Screenshot of the content. The content is stored in the WOWxv 2-D-plus-Depth format, i.e., a red-green-blue (RGB) image (left) with a corresponding depth map (right).

In the depth map, the grey value per pixel (ranging from 0 till 255 and thus enabling 256 different screen disparity values) indicates the relative screen disparity of the corresponding RGB pixel. The amount of maximal screen disparity (i.e. the screen disparity corresponding to a grey value of 255 in the depth map) can be set and altered by varying a gain factor, when rendering the left and right views. Figure 26 depicts the corresponding maximum distances of content in front of and behind the display for the low and high IniDisp setting.

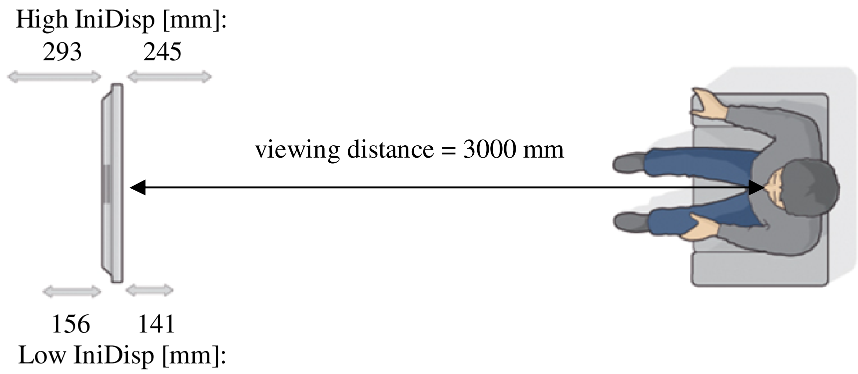


Figure 26. IniDisp settings visualized in maximum distances for content both in front of and behind the display screen.

In a pilot study a 3-D expert panel of four people watched 22 scenes in the 3-D movie and categorized them based on the level of activity of the content. More specifically, scenes with non-moving objects both in lateral and depth direction, e.g., talking people, were referred to as static scenes (n=12 in the entire movie), and scenes with moving objects both in lateral and depth direction, e.g., battles and races, were referred to as dynamic scenes (n=10 in the entire movie). The short sequences selected for the SA method included two static and four dynamic scenes. The distribution of the short sequences over the entire 3-D movie was nearly uniform: the sixth short sequence was captured from the last 30% of the movie and the other five short sequences were distributed over the remaining 70%. Per participant each short sequence was displayed with the same screen disparity as the corresponding part within the 3-D movie.

Equipment

All 3-D sequences were displayed on a 42" Philips autostereoscopic nine-view lenticular LCD display. For a detailed description, read Chapter 2.5.1. Experimental set-up.

For the assessments a hand held slider was used. The position of the slider could be adjusted along a graphical scale of a length of 10 cm having at regular distances the adjective terms [bad]-[poor]-[fair]-[good]-[excellent] according to the ITU recommendations (ITU, 2002). The position of the slider was sampled at 1000 digital values continuously at a sample rate of 2 Hz (every 0.5 second). Also, recording a single assessment was possible by pressing a button on the tool after which the current position of the slider was stored. A sensor was placed on the screen to detect luminance output of the screen and was used to synchronize the start of the movie with the start of the assessment.

The conversion algorithm that was used to convert the stereoscopic movie into the WOWvx 2-D-plus-Depth format combined several 2-D-to-3-D conversion algorithms that all had their own strengths and weaknesses. It relied on structure-from-motion, as well as on different assumptions, estimations and heuristic cues (Redert et al., 2007). As usual, the 2-D-to-3-D conversion did not generate artifact-free depth information; for some parts of the movie content was rendered at an inappropriate depth layer (all within the one-degree of disparity).

Procedure

The experiment was performed in an experimental room at Philips Research Laboratories where lighting conditions were held constant for all participants in all sessions. After completing the Landolt-C and RANDOT stereo test successfully, participants were seated at the optimized viewing distance of three meters behind a table on which the slider was positioned. Preceding the experiment they received a brief instruction concerning the course of the training and the experiment.

The participants were asked to assess stereoscopic content in terms of visual comfort with the slider. It was decided to assess the 3-D content in terms of visual comfort instead of visual discomfort. A pilot revealed that participants became confused by the negative aspect of the evaluation concept, that is, the fact that 'bad' visual discomfort is positive and 'good' visual discomfort is negative appears counterintuitive to many. Participants were asked to use the full range of the scale and were free to position their slider anywhere on the scale. Participants were requested not to move their head, because perception of image quality and depth could differ across the views of the multi-view display. Any questions concerning the procedure of the experiment were answered.

In order to familiarize participants with the assessment method and the stimuli, two one-minute 3-D video sequences were displayed. The first sequence had completely different content from what was used in the actual experiment and was incorporated to acquaint participants with the display and 3-D content in general. Participants did not have to assess this sequence. The second sequence was captured from the actual movie and was incorporated to familiarize participants with the slider. The participants had to assess the sequence continuously in terms of visual comfort by using the slider and were asked to give an overall assessment after the sequence by positioning the slider and pressing the button.

In the actual experiment, participants had to assess (1) the six short sequences by providing a single assessment score for each short sequence, (2) a continuous assessment of the 24-minute 3-D movie, and (3) a retrospective assessment score of the entire 3-D movie. The entire procedure is depicted in Figure 27. To determine correlations between the single assessments of the 10-second short sequences (hereafter referred to as SA1-SA6) and their corresponding assessments within the 3-D movie (hereafter referred to as COAS1-COAS6), the COA scores of these corresponding parts needed to be extracted from the total COA.

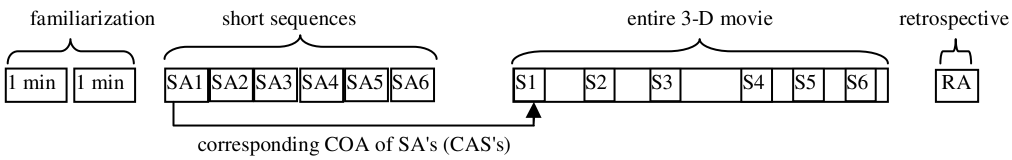


Figure 27. Time diagram of the procedure of Experiment 4. SA1-6 are the single assessments of the 10-second short sequences S1-6, COA is the continuous assessment, COAS is the corresponding assessments of the SA within the 3-D movie and RA is the retrospective assessment.

4.3.2. Results

Assessment methods

In this section the results are described in the following order: the validation of the continuous assessment, the analysis of the short sequences and the analysis of the retrospective assessment.

Continuous assessment

Figure 28 presents the raw visual comfort scores averaged over the participants for each of the four disparity conditions (see Table 9). The overall standard deviation and the time, at which the six 10-second sequences occurred in the movie are shown as well.

The change in COA, as shown in Figure 28 is relatively similar for all disparity conditions, which provides a first validation for this assessment method in terms of visual comfort. Especially in the first 70% of the movie, in which only IniDisp high and IniDisp low occur, negative and positive peaks occur at similar parts in the sequence. When 70% of the movie had progressed the screen disparity was halved for approximately half of the participants, after which four disparity conditions can be distinguished. In Figure 28 the 70% point is visualized by a time gap at approximately 17 minutes. Note that just before this point a large negative peak in visual comfort occurred, followed by large differences in COA between the four disparity conditions. Apparently, participants recovered differently from the dip in visual comfort depending on whether the disparity in the displayed content was higher or lower. After a few minutes these differences in COA between the disparity conditions diminished again. It appears that participants required some time to adapt to a certain change in comfort and/or disparity, but after a few minutes their visual comfort levels were back to average. A more global trend in adaptation to visual discomfort is seen during the course of the whole movie via a gradual increase in COA scores. Indeed, the visual comfort over the first 3 minutes of the movie is considerably lower than over the last 3 minutes, independent of the disparity condition. The level of visual comfort in general is relatively high, with the exception of certain parts or scenes with large negative peaks.

Figure 28 also depicts the standard deviation (STD) of the raw scores over all participants, which has a relative high mean value of 186. Peaks in the STD occur at approximately similar moments when drops in the visual comfort scores occur. Apparently, the video characteristics that cause the drops in visual comfort have different impacts on the participants. Hence, it seems that participants assess the time-variant video characteristics similarly, though with a different internal assessment range. More clarity is provided by a Cronbach's Alpha reliability analysis, which reflects the variability in the responses that is the result of differences between the participants and not of different video characteristics that were assessed. More specifically, the means and variances of the assessments per group can differ, but their covariance should be equal. The reliability testing revealed Cronbach's Alphas of 0.88, 0.89 and 0.77 for the full movie, the first 70% and the last 30% of the COA scores respectively (note, a Cronbach's Alphas of 0.70 is considered as 'acceptable' (Hair, Black, Babin, Anderson, Tatham, 2006)).

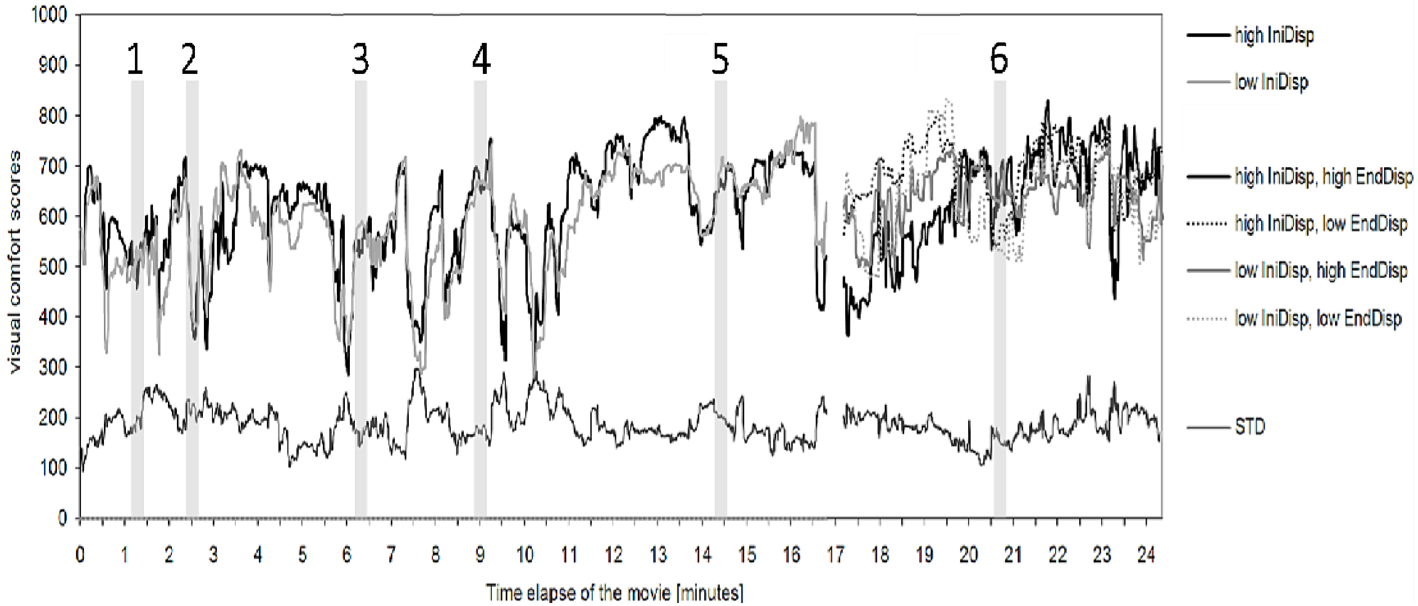


Figure 28. The raw visual comfort scores averaged over the participants per disparity condition (see Table 9) including the overall standard deviation. The x-axis represents the time lapse of the 3-D movie in minutes, the y-axis represents the total assessment range (a score of 1000 corresponds to comfortable and 0 to uncomfortable). The different lines represent the different screen disparity conditions and the standard deviation (STD). In the first 70% of the movie only high IniDisp (black line) and low IniDisp (grey line) can be distinguished and after 70% each line splits up in a high EndDisp (solid line) and low EndDisp (dotted line). The numbers 1 to 6 refer to the time positions, at which the six short

Short sequences

To determine correlations between the single assessments of the 10-second short sequences (hereafter referred to as SA1-SA6) and their corresponding assessments within the 3-D movie (hereafter referred to as COAS1-COAS6), the COA scores of these corresponding parts needed to be extracted from the total COA. Rather than extracting the entire 10 seconds, only the last 8 seconds were used under the assumption that participants required time to readjust their slider to the desired values. For each of the COASs the mean over the 16 sampled COA values (8 seconds sampled at 2 Hz) was determined. The correlation (calculated over N=24 participants) between the six SAs and their corresponding COASs are outlined in Table 10. All correlations are low and lack significance. Since the disparity conditions differed between participants, the correlations were also calculated per disparity condition. The results are similar; correlations are low and lack significance. These results are surprising, since one would expect that a particular (active) scene with a clear dip in visual comfort during the movie is also recognized as an uncomfortable scene when assessed in isolation. This is actually the basic assumption underlying all visual discomfort assessment studies with short sequences. Although not being significant, the data in Table 10 shows that indeed the correlation between the COA and SA scores is higher for the active than for the static scenes. Due to the large spread in scores, however, there is no significant difference ($p = .682$) in averaged SA scores between the static (574 ± 223) and active (559 ±264) scenes. The average COAS scores, however, significantly differed ($p < .05$) between the static (548 ± 147) and the active scenes (513 ± 181).

TABLE 10
CORRELATIONS (N=24) BETWEEN THE SCORES OF THE SHORT SEQUENCES (SA1-SA6) AND THEIR CORRESPONDING COA SCORES (COAS1-COAS6).

	Activity	Correlation	Sig.
SA1 - COAS1	active	0,339	0,106
SA2 - COAS2	active	0,360	0,084
SA3 - COAS3	static	-0,107	0,617
SA4 - COAS4	active	0,217	0,309
SA5 - COAS5	static	-0,124	0,562
SA6 - COAS6	active	0,289	0,170

Retrospective scores

To examine what part of the 3-D movie mainly determines the retrospective score (RA), the COA scores were averaged over various parts of the 3-D movie, i.e., over the entire movie, over its first 70%, over its last 30% and over its last 10 seconds. The resulting mean COA scores were correlated with the RA (calculated over N=24 participants). These correlations are outlined in Table 11. It seems that the retrospective score correlates better with the COA of the first 70% of the movie than with the COA over latter parts.

Since for part of the participants the disparity was halved during the last 30% of the movie, one would expect a larger correlation of the RA with the COA averaged over the last part of the movie for at least this group of participants. This is explicitly checked in Table 12, outlining the same correlations, but now per EndDisp setting. This table, however, shows that the correlation between the RA and the corresponding COA over the last part (i.e. the last 30% or 10 sec.) of the movie is particularly low when the screen disparity was halved. This observation can be related to the fact that the RA did not significantly differ between the EndDisp=full (669 ± 158) and EndDisp=half (610 ± 175) condition. In other words, the effect of halving the disparity only had a minimal effect on the RA scores. This will be further elaborated in paragraph 4.1.6. Discussion.

TABLE 11

CORRELATIONS (N=24) BETWEEN THE RETROSPECTIVE SCORE (RA) AND THE MEAN SCORE OVER DIFFERENT PARTS OF THE MOVIE, I.E., OVER THE ENTIRE MOVIE, ITS FIRST 70%, ITS LAST 30% AND ITS LAST 10 SECONDS.

	Correlation	Sig.
RA - entire COA	0,69	0,001
RA - first 70% of COA	0,72	0,001
RA - last 30% of COA	0,51	0,01
RA - last 10 sec of COA	0,54	0,01

TABLE 12

CORRELATIONS PER ENDDISP SETTING (N=12) BETWEEN THE RA AND THE MEAN SCORE OVER DIFFERENT PARTS OF THE MOVIE, I.E., OVER THE ENTIRE MOVIE, ITS FIRST 70%, ITS LAST 30% AND ITS LAST 10 SECONDS.

	EndDisp=full			EndDisp=half		
	N	Correlation	Sig.	N	Correlation	Sig.
RA - entire COA	12	0,70	0,01	12	0,79*	0,001
RA - first 70% of COA	12	0,68	0,03	12	0,84	0,001
RA - last 30% of COA	12	0,69	0,01	12	0,48*	0,11
RA - last 10 sec of COA	12	0,71	0,01	12	0,50*	0,10

* sessions in which the screen disparity was halved for the last 30% of the movie

Video characteristics

The three most pertinent determinants that reduce visual comfort within the screen disparity settings applied in this experiment are expected to be 3-D artifacts, unnatural blur and excessive demand on AC linkage. All three determinants are directly related to the amount of motion and (changes in) screen disparity in the content. The processing steps that were implemented for the extraction of motion and screen disparity characteristics from the image material used in the experiment are illustrated in Figure 28.

In the first step the incoming signal with a 2-D-plus-Depth format was split into separate frames each including a separate RGB and depth image to allow derivation of 2-D and 3-D image characteristics independently. In the second step motion estimation was applied to the RGB image. The 3-D Recursive Search Block matching method was used with two spatial and two temporal prediction vectors within the matching block to estimate the motion (de Haan, 2006). This process resulted in a motion vector per pixel. The depth map did not require pre-processing for the extraction of disparity characteristics, since the relative disparity information per pixel was already available in the grey-scale image. In the third step different potential motion and screen disparity derivatives were extracted from the motion and screen disparity information, respectively. Included were means, partials (e.g., minimum and maximum 10%) and ranges (difference between minimum and maximum 10%) over all pixels for each frame. The change in each derivative between frames was determined by their gradient. This step resulted in a value for each extracted derivative per frame. In the fourth step the frequency of the derivative values per frame and of the COA scores were down-sampled to a similar value. In the fourth step, the frequency of the derivative values per frame (i.e., 36000 frames in total) were temporally down-sampled from 25 Hz to 1 Hz to match the data frequency of the COA.

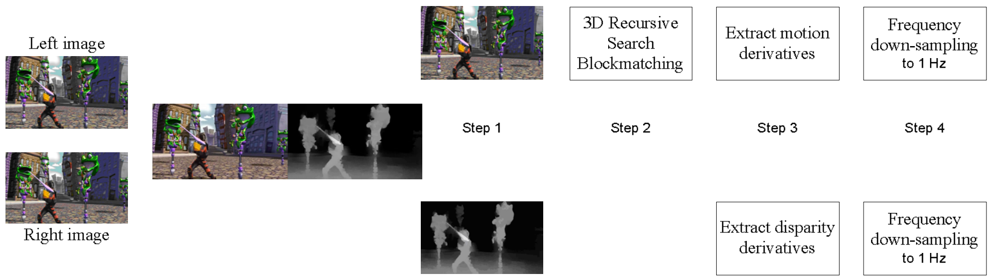


Figure 28. Flowchart representing the extraction of time-variant derivatives of motion and screen disparity.

Now a COA score per second could be regressed into different combinations of motion and screen disparity derivatives. The initial regression was based on previous research and incorporated motion, screen disparity and changing screen disparity, i.e., gradient of the screen disparity. The regression equation with the best fit, when incorporating only those predictors with a significant effect on the COA, is depicted in Equation 8:

$$COA = \alpha \cdot \overline{mo} + \beta \cdot \overline{de} + \delta(de_{\max 10\%} - de_{\min 10\%}) + \gamma[G(de_{\max 10\%} - de_{\min 10\%})] \tag{8}$$

with
COA = COA scores
$\overline{mo}$ = average amount of motion
$\overline{de}$ = average amount of screen disparity
$(de_{\max 10\%} - de_{\min 10\%})$ = screen disparity range
$[G(de_{\max 10\%} - de_{\min 10\%})]$ = gradient screen disparity range
$\alpha, \beta, \delta, \gamma$ = weights of the predictors

The explained variance, however, was low (R^2= .35). This might have been the result of the facts 1) that participants had some delay in setting the correct COA score, and/or 2) that the model was fitted to all disparity conditions simultaneously. As such, Figure 29 depicts the explained variance per disparity condition as a function of a time-shift between the COA and its predictors.

A few aspects are noteworthy. First, the behavior of the explained variance as function of the shift was similar for all disparity conditions, i.e., the R^2 increased for the first few seconds of time-shift and then dropped to almost zero.

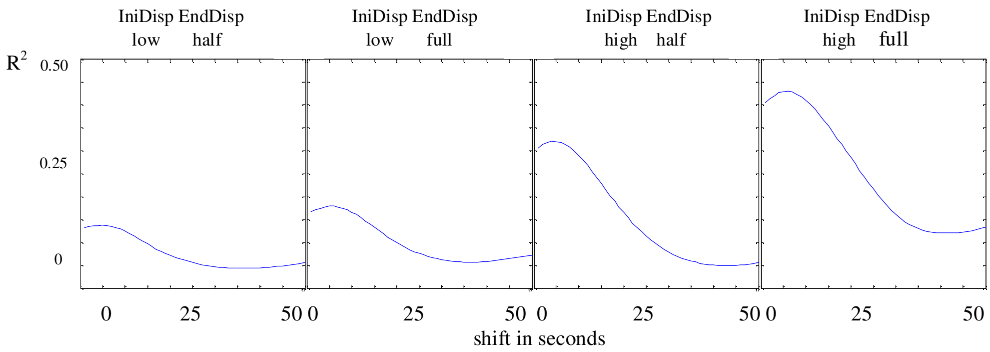


Figure 29. Explained variance per disparity condition as a function of a time-shift between the COA and its predictors in Equation 8. The x-axis represents the time shift in seconds and the y-axis represents the explained variance.

A second aspect is that the explained variances were significantly higher for high IniDisp conditions, though this might also be attributed to having more participants for these conditions. Although these aspects are promising, the R^2s were nevertheless still too low for the model to be valid. Since parts of the movie differed considerably in activity with respect to motion and screen disparity, it was highly plausible that certain derivatives were only appropriate for particular parts of the movie.

To evaluate the effect of activity in a scene, the 3-D movie was split into 22 scenes that differed in image content, motion and screen disparity characteristics. Per scene the COA used in the model was shifted with respect to the video characteristics with a maximum of five seconds to acquire the best model fit. A first explorative analysis revealed little consensus in the parameter weights of the regression analysis between the scenes. As mentioned in paragraph 4.1.4. Experimental set-up, the 22 scenes were clustered based on their activity in the content, i.e., 12 static and 10 dynamic scenes. Table 13 outlines the weighting coefficients and the corresponding R^2‘s for the best fits on the static and dynamic scenes separately. The weights of the regression terms are specified per screen disparity condition.

The mean visual comfort of static scenes can be predicted as a combination of contributions of 58% of average screen disparity, i.e., referred to as an offset, and contributions of 42% of screen disparity range. The visual comfort of dynamic scenes can be predicted as a combination of contributions of 70% of screen disparity range, 22% of motion and 8% of changes in screen disparity. Again, a few phenomena are noteworthy. First, not all scenes were included in each cluster to obtain the best fit. Some scenes (n=4) were difficult to refer to as either dynamic or static and including them in either cluster decreased the R^2 of the fit significantly. Other scenes (n=3) were obviously static or dynamic and including them also decreased the R^2 of the fit significantly. These scenes will be further considered in paragraph 4.3.3. Discussion. Second, for both the static and dynamic scenes the explained variance is low when IniDisp is low and is higher when IniDisp is high. This seems logical since the screen disparity has a large impact on the COA in the regression.

TABLE 13
THE WEIGHTING COEFFICIENTS OF THE REGRESSION TERMS RESULTING IN THE BEST FIT PER DISPARITY CONDITION FOR THE STATIC AND DYNAMIC SCENES SEPARATELY

Conditions		Static scenes (n=9)			Dynamic scenes (n=6)			
IniDisp	EndDisp	β	δ	R^2	α	δ	γ	R^2
low	half	49	51	0,33	34	48	18	0,42
low	full	45	55	0,42	25	58	17	0,32
high	half	49	51	0,53	14	75	11	0,65
high	full	86	14	0,61	24	72	4	0,66
mean		58	42	0,60	22	70	8	0,62

4.3.3. Discussion

With respect to the first hypothesis we found that visual comfort of stereoscopic scenes can be predicted as a combined effect of screen disparity range, screen disparity offset, changing screen disparity and lateral motion, of which the specific contributions depend on the activity of the scene. The specific weights in Table 13 do not apply to all 3-D movies. Since the units of the video characteristics differ, the weights are not free from scaling effects and reflect the relative ranges of the video characteristics. Hence, if a different disparity range was applied, the ‘good’-score in visual comfort would correspond to a different disparity. These video characteristics are assumed to directly impact visual discomfort via 3-D artifacts, unnatural blur (crosstalk) and a changing demand on AC linkage, of which the first two probably have the largest impact in this experimental setting. Note that the model is sufficiently reliable even though only moderate instead of extreme screen disparities were used, and as a consequence the whole movie was considered as reasonably comfortable. This setting was explicitly chosen to approach those in consumer applications.

According to our model, screen disparity had the largest impact on visual comfort. Additional evidence is found in two observations: 1) when the content was shown with a large screen disparity, i.e., IniDisp was high, the regression explained significantly more variance than when the content was shown with a low screen disparity, and 2) an R^2 of 0.6, as found for the high-disparity setting is relatively high for a non-specific evaluation criterion that is predicted by variations in motion and screen disparity.

There were some scenes, however, for which the COA scores of which could not be predicted. Four scenes were difficult to refer to as either active or static and including them decreased the R^2 of the fit significantly. Two of these scenes concerned content in outer-space of which the black background made it difficult to perceive relative depth of objects. Additionally, three scenes were obviously active, but not appropriate for the model to predict their visual comfort level. This inappropriateness was the result of an error during the 2-D-to-3-D conversion; the entire scene content was rendered at a single depth layer (note that here it concerns a 2-D-to-3-D conversion error instead of a 3-D artifact), and since there was no change in screen disparity range, the model was not applicable.

Our model is in line with previous research in which visual discomfort for non-moving images is related to high screen disparity and for moving images to rapidly moving objects and changing screen disparity (Yano et al., 2002 and 2004; Speranza et al., 2006). However, the finding of Speranza et al. that the rate of change in screen disparity has the largest impact on visual comfort is not confirmed; in our model changes in screen disparity did have a significant effect on the visual comfort of active scenes, but it was by far not the largest one.

An interesting aspect is that the stereoscopic content used in our experiment, i.e., nine-view with artifacts, differed from the content used in previous research, i.e, two-view artifact-free content. Whereas previous research mentioned only the changing demand on AC linkage as the main determinant for visual discomfort (Yano et al., 2002 and 2004), we believe that in the current experiment 3-D artifacts and unnatural blur were the main determinants. Converting from 2-D-to-3-D, and rendering and displaying the content, resulted in obvious 3-D artifacts, and unnatural blur became more present with increasing screen disparity in both static and dynamic scenes due to crosstalk. The interdependent impact of these three determinants on the visual comfort assessment remains indefinite, though we assume that this impact depends on generation, rendering and display technology. In practical terms, stereoscopic displays without crosstalk or stereoscopic content without 3-D artifacts can be rendered with more screen disparity before inducing visual discomfort in terms of unnatural blur and 3-D artifacts. In that case it can be assumed that instead of crosstalk and 3-D artifacts that had a high impact on visual comfort in this experiment, changing demand on AC linkage yet has the largest impact.

With respect to the second and third hypothesis, three assessment methods were addressed that can provide more insight into how visual discomfort is built up while watching 3-D content; a continuous assessment (COA), assessment of multiple short sequences captured from the movie (SA) and a retrospective assessment of the entire movie (RA).

The analysis of the COA reveals that participants are able to continuously assess visual comfort, which confirms that the COA is a consistent and reliable method. The average visual comfort level is relatively high for all groups, which can be attributed to the moderate screen disparity used. The global increase of comfort scores in COA from the start of the movie until the end indicates a gradual adaptation to visual (dis)comfort.

The assessment of short sequences captured from the 3-D movie is a poor reflection of how the same content is perceived in the 3-D movie: the SAs correlated poorly with their corresponding CASs. This can be attributed to context effects, i.e., preceding scenes affect the perception of subsequent scenes. Another explanation, which is strengthened by other results that will be discussed later, is that the experience of visual comfort changes with time. More specifically, the occurrence of visual discomfort as a result of prolonged viewing and a possible recalibration of the internal assessment scale during the course of the 3-D movie can affect the COA scores, but hardly the SA scores. This does not mean that the visual comfort experienced after 10 seconds is a poor reflection of visual comfort at that moment, but it has limited value to generalize conclusions drawn from the assessment of such short sequences to the entire movie.

Our results provide no evidence that the retrospective assessment is affected by a negative peak effect or a recency effect. Assuming that visual comfort is related to screen disparity as revealed by the regression model, there is only one point in the movie that can indicate a “negative” peak effect, i.e., the recover from a visual comfort dip when the disparity is halved for half of the participants when 70% of the movie had elapsed. If a negative peak effect would have affected the RA scores, the RA of the condition in which the screen disparity halved, would be significantly higher than the RA of the condition of which the screen disparity remained unchanged, i.e., lower screen disparity should result in higher visual comfort scores. This, however, is not the case. The same point in the movie can be used to evaluate the recency effect. More specifically, if a recency effect is present, one would expect the correlations between the RA and the COA averaged over the last 30%

or last 10 seconds to be high. These correlations, however, are low. A possible explanation would be that in some conditions the screen disparity was too low for participants to assess visual (dis)comfort; i.e. the COA for those parts represents noise rather than visual discomfort. Figure 28, however, suggests that this is not the case; even for the low disparity conditions there still is a high degree of consistency in scoring visual comfort dips. What can explain the low correlation between the RA and the COA averaged over the last part of the movie is that groups belonging to different disparity conditions just needed different periods to recover from the large negative peak in visual comfort, but after approximately three minutes all ended up with the same level of visual comfort.

We do not want to claim that retrospective visual comfort assessment is not affected by recency and/or negative peaks, yet their impact is not obviously present in our experimental results. Note that both effects are normally evaluated in sequences with clear and obvious image quality artifacts. In the current experiment, however, the amount of visual discomfort is not very high, and as such, both effects are very small, if present at all. The retrospective assessment reflects the average visual comfort of the entire movie and as such, is a better indicator than assessment of short sequences.

4.4. Experiment 5: video characteristics and subtitles

In Chapter 3 we identified three factors that can induce visual discomfort even within the one-degree of screen disparity limit: (1) excessive demand of accommodation-convergence linkage, e.g., by fast motion in depth, (2) 3-D artifacts resulting from insufficient depth information in the incoming data signal, yielding spatial and temporal inconsistencies, and (3) unnatural amounts of blur. In addition (4), subtitles can generate visual discomfort as a consequence of depth inconsistencies. Subtitles are often rendered at the display plane and induce depth cue conflicts by occluding 3-D objects that are rendered in front of the display plane. In the case of still images such rendering of subtitles was indicated to be conflicting, unclear and annoying (Pockett et al., 2009). Analogously, ticker-tapes, i.e., a strip of program information, or an advertisement, can result in similar visual discomfort.

The perceptual impact of the four factors mentioned above is directly related to the amount of screen disparity or the amount of motion, and as these quantities change, their impact fluctuates in time. Hence, the combination of motion and changes in screen disparity can induce visual discomfort, which is expected to become more severe with prolonged viewing. In Experiment 4, the direct impact of these video characteristics (except subtitles) was determined by relating it to a continuous assessment of visual comfort. It was concluded that the visual comfort of stereoscopic scenes could be predicted as a combined effect of (changes in) screen disparity range and offset and lateral motion of which the specific contributions depended on the activity of the scene. This conclusion, however, needs to be interpreted with caution for a number of reasons. Firstly, the 3-D movie comprised relatively simple screen disparity (i.e., objects of interest were rendered on the foreground layered over a flat background that was positioned statically behind the display plane). Secondly, this movie was displayed on an autostereoscopic display that induced high amounts of unnatural blur. Thirdly, subtitles were not included in the 3-D movie. It has often been stated that they induce visual discomfort, yet to our knowledge this has only been verified for still images (Pockett et al., 2009). And finally, the perception of the 3-D movie was only evaluated in terms of general visual comfort. No questions were asked concerning which video characteristics were incorporated in the continuous assessment.

In the current research the objective is two-fold; 1) to determine how the instantaneous visual comfort is affected by video characteristics and subtitles, and 2) to verify the prediction of visual comfort from video characteristics. The stereoscopic content used had high spatial and temporal complexity and was rendered on a high resolution two-view stereoscopic TV, i.e., a configuration that is present in home theatre sets. The first hypothesis states that video characteristics such as changing screen disparity and lateral motion can be used to predict the visual comfort of stereoscopic content with high spatial and temporal complexity. The second hypothesis states that subtitles negatively affect visual comfort.

4.4.1. Experimental set-up

Design

The experimental design consisted of a short screening of the participants, and an assessment of visual discomfort of two stereoscopic movies: a travel documentary (‘Travel’) and a sport movie (‘Sports’) that each consisted of distinctive scenes (Scene). Both movies were available with subtitles (Sub) and without subtitles (no Sub). The movies were continuously assessed in terms of visual comfort and were preceded and followed by an objective measurement of vergence facility and a subjective measurement with retrospective questionnaires. The order of showing the movies was varied as outlined in Table 14. Forty persons participated in the experiment. They each saw two movies in one session, where one movie had subtitles and the other not. As an example, ten participants started with the movie 'Sports' containing subtitles, and then saw the movie 'Travel' without subtitles, etc.

TABLE 14
ORDER OF MOVIES ASSIGNED TO PARTICIPANTS

		Sub	
		Sub	no Sub
Order	'Sports' – 'Travel'	10	10
	'Travel' – 'Sports'	10	10

Participants

The participants were forty employees and graduate students working in a research environment. Twenty-two were male and eighteen were female, with a mean age of 37 years (range between 22 and 65). The participants selected for the experiment had normal colour vision (as tested with the Ishihara colour-blindness test), good visual acuity (a visual acuity of ≥ 1 or 20/20 on the Landolt-C test), and good stereo acuity (≤ 30 seconds of arc on the RANDOT stereo test).

Stimuli

Two stereoscopic movies, 'Sports' and 'Travel', that differed in scene activity in line with the results of Experiment 4 were selected. The 'Sports' movie (889 seconds) contained four scenes with different types of sports: American football, boxing, racing and basketball. The 'Travel' movie (828 seconds) contained three scenes of a Spanish region each at different locations: landscapes, villages and the inside interior of a church and house. Both movies also had short introductory and concluding parts. Figure 31 depicts the relative motion and relative disparity data per frame (see section 4.3.2. Results Video Characteristics for the derivation of this data). The activity in the 'Sports' movie was high, including rapid camera and object motion, abrupt switches between close-ups and distant shots and fast changes in screen disparity, whereas the activity in the 'Travel' was low, with slow camera motion and static scenes, no object motion and no changes in screen disparity other than at scene switches. The average motion, disparity and disparity gradient were respectively 5.1, 0.8 and 4.6 times larger in the 'Sports' movie than in the 'Travel' movie.

Absolute screen disparity information was not available, yet the visual result was not considered excessive. At a short viewing distance (three meters from a 56" display, which subtends approximately 23 degrees of visual angle), the authors felt that the maximum screen disparity was just acceptable in terms of visual comfort. Dutch subtitles were added to each movie and rendered at the display plane where the readability and naturalness was highest (Pockett et al., 2009). For the 'Sports' movie, this was done by translating the English spoken commentary into Dutch. Since the 'Travel' movie contained no speech, subtitles were composed using information about the region such that the amount and frequency of on-screen text was approximately equivalent to the 'Sports' movie. The font (typeface Arial, size = 34, white) and position (centered at the lower part of the screen) of the subtitles were chosen to match the look of typical Dutch subtitles for motion pictures.

Both movies also had audio; for the 'Sports' movie the audio was similar to that of conventional sports reports, whereas the 'Travel' movie had relaxing background music.

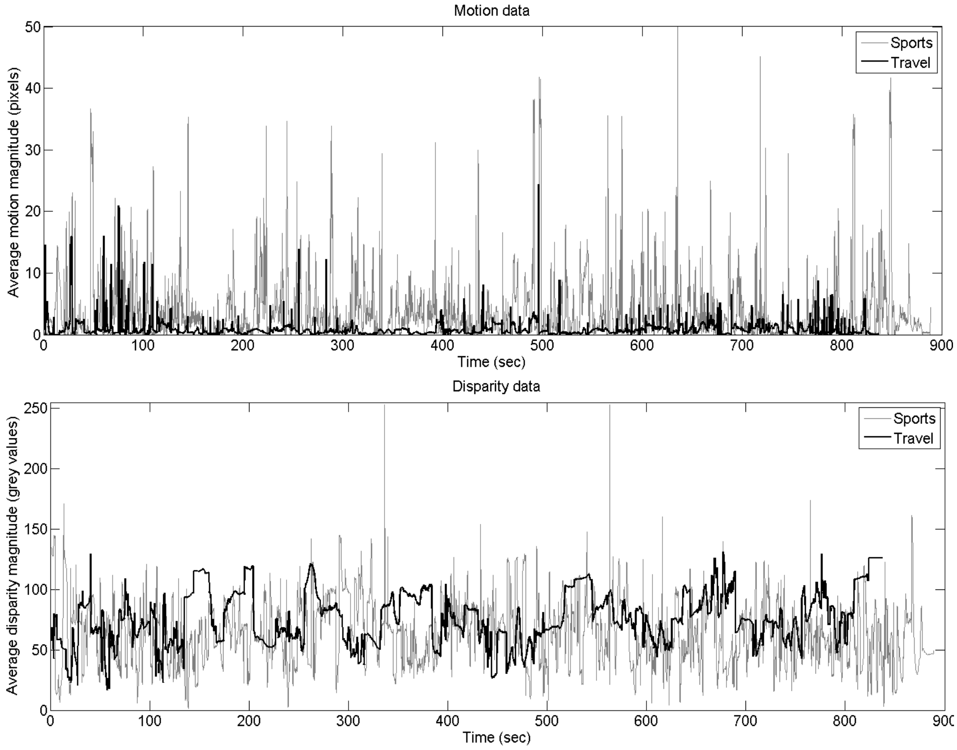


Figure 30. The mean amplitude of the motion vectors (upper graph) and relative disparity (lower graph) of the 'Sports' movie (grey lines) and 'Travel' movie (black lines). The x-axis represents the time elapse of the movie in number of frames and the y-axis represents the relative amount of motion and disparity (see Result section for the derivation of this data).

Equipment

The 3-D movies were displayed on a 56-inch Samsung 3-D display type HL-T5687S. The 3-D display is a Full HD rear projection monitor, with a dynamic contrast ratio of 10000:1 and a display frequency of 120 Hz. It utilizes temporal multiplexing (shutter glasses) to create stereoscopic depth.

The continuous assessments (COA) were performed with a hand-held slider. The position of the slider could be adjusted along a graphical scale of a length of 10 cm with the adjectives [bad]-[poor]-[fair]-[good]-[excellent] at regular intervals along the scale, in line with the ITU recommendations (ITU-R, 2002). The position of the slider was sampled at a rate of 2 Hz. A small luminance sensor that was connected to the slider was placed on the screen to synchronize the start of the movie with the start of the assessment.

Measurement methods

The impact of the 3-D movies on the binocular visual system was evaluated preceding the first movie (pre), after the first movie (intermediate) and after the second movie (post). For objective measurement, vergence facility was evaluated. Vergence facility is the ability of the fusional vergence system to respond efficiently and accurately to changing demands over time (Gall et al., 1998). It refers to the speed with which an individual can recover fusion or sustain binocular clear vision in the presence of rapid changes in vergence demand, i.e., the kind of changes that are present in 3-D movies. It is measured in cycles per minute that an object can be properly fused through alternating base-in and base-out prisms (4 prism convergent / 4 prism divergent) at a viewing distance of three meter (the viewing distance of the 3D display).

For the subjective measurement, questionnaires were used. The questionnaire consisted of three parts. The first part addressed general comfort issues and was an adjusted version of the Convergence Insufficiency Symptom Survey (CISS) that consisted of 15 items including general visual fatigue items as well as specific items concerning reading: *discomfort*, *loss of concentration*, *double vision*, *sleepiness*, *sharpness*, *exhaustion*, *appearance of moving words*, *slower reading*, *losing position in text*, *trouble remembering words*, *re-reading words*, *headache*, *pain in the eyes*, *strain in the eyes* and *irritated eyes* (Borsting et al., 2003; Sheedy et al., 2003). The second part of the questionnaire related to the overall experience of the movie (*3-D experience*, *naturalness*, *viewing experience*, *image quality*). The third part related to the video characteristics: four items addressed the impact of subtitles (*disturbing*, *interesting*, *necessary*, *readable*), and eight items addressed specific content characteristics (*speed of the images*, *amount of depth*, *scene-switches*, *framing of the content by the display, quality*, *camera motion*, *changes in depth*, *artifacts*). In the pre-questionnaire only the first part of the questionnaire was evaluated, and since subtitles were only present in one of the two movies per participant, the four items related to subtitles were only judged when subtitles were present.

The framework of the items was adjusted such that they were aimed at visual discomfort experienced whilst watching TV and whilst reading subtitles. All questionnaire parts were retrospective, i.e., it was explicitly stated that in the pre-questionnaire the items related to previous experiences while watching 2-D TV and in the intermediate- and post-questionnaires to the feelings experienced while watching the 3-D movies. The items in the first and third part of the questionnaire were assessed using a scale with the adjectives [never]-[seldom]-[occasionally]-[often]-[always]. The items in the second part were assessed using a scale with the adjectives [bad]-[poor]-[fair]-[good]-[excellent]. The adjectives were then transformed into numerical values ranging from zero to four.

Procedure

The experiment was performed in an experimental room at Philips Research Laboratories where lighting conditions were held constant for all participants in all sessions. Participants had to sign an informed consent containing information about the experiment and about the possible occurrence of (harmless) visual discomfort. After completing the Landolt-C, colour blindness test and RANDOT stereo test successfully, participants filled in the pre-questionnaire and performed the pre-vergence facility test. After the screening phase, participants were seated at a viewing distance of three meters on a comfortable chair and were given shutter glasses. The slider for the continuous assessment of visual comfort was positioned on a table next to the chair.

It was decided to assess the 3-D content in terms of visual comfort instead of visual discomfort. Experiment 4 revealed that participants became confused by the negative aspect of the evaluation concept; the fact that 'bad' visual discomfort is positive and 'good' visual discomfort is negative appeared counterintuitive to many. Participants were asked to use the full range of the scale and were free to position the slider anywhere on the scale. In order to familiarize participants with the assessment method, the display and the 3-D content, the participants were asked to continuously assess a one-minute 3-D video sequence in terms of visual comfort using the slider (this short sequence was completely different from the actual movies in the experiment). Remaining questions concerning the procedure of the experiment were answered. In the actual experiment, participants had to assess two 3-D movies continuously. After the first and second movie they performed the intermediate- and post-vergence facility test and filled in the intermediate- and post-questionnaire. Between the movies, participants had a few minutes of rest. At the end of the post-questionnaire, participants could add remarks and mention specific aspects that were not addressed before.

Statistical analyses

To find intrinsic correlations between the questionnaire items and to reveal if certain items shared similar underlying attributes of visual discomfort, principal component analyses were performed (PCA) (Hair et al. (2006). The PCA were combined with a non-orthogonal rotation method (Oblimin) to minimize the number of items with high factor loadings on more than one factor. For each participant, factor scores were calculated for each factor in each condition; i.e., the score given by the participant on each questionnaire item was weighted with the factor loading of that item, and then summed over all items of the factor. ANOVA could then be performed with these factor scores as dependent variables.

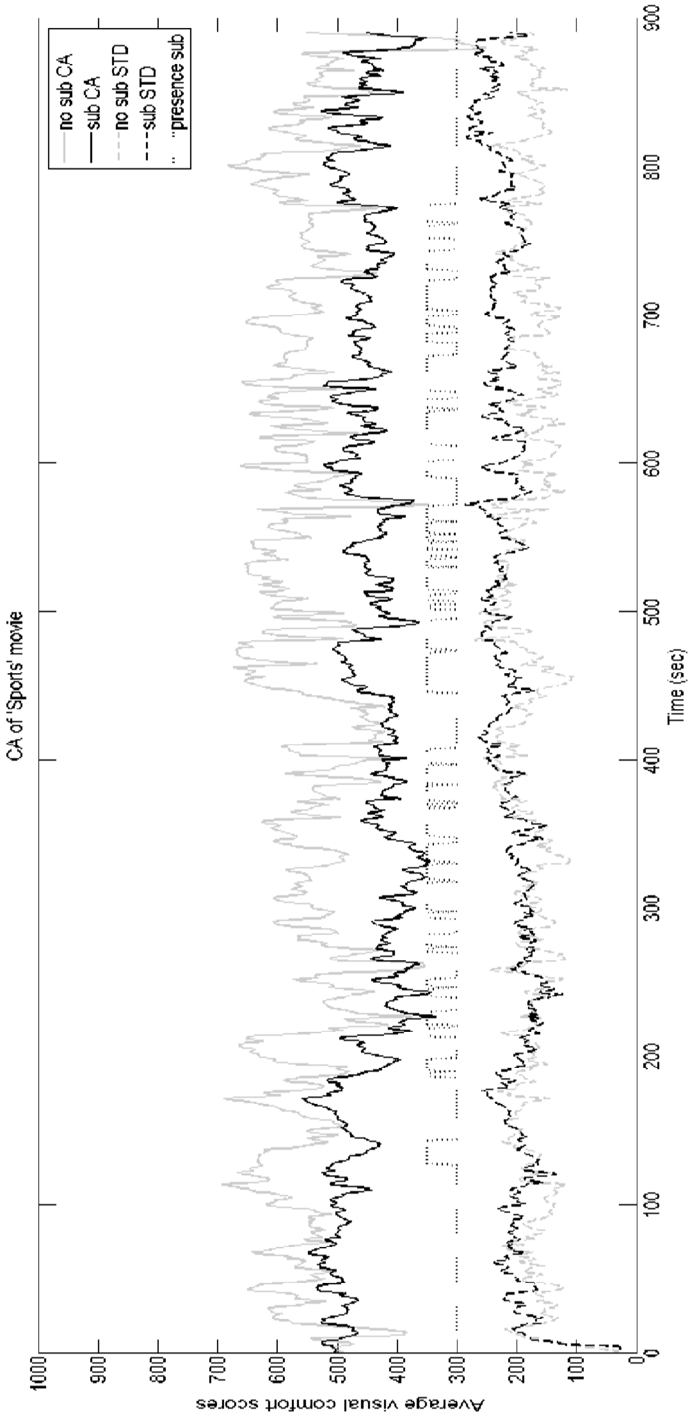
For reliability analysis Cronbach's Alphas were calculated. This analysis reflects the variability in the responses that is the result of differences between the participants and not of e.g. different video characteristics that were assessed. More specifically, the means and variances of the assessments per group may differ, but their covariance should be equal. Cronbach's Alpha of 0.70 is considered as 'acceptable' (Hair et al. (2006).

4.4.2. Results

The vergence facility differed between participants without exceeding norm values (Gall et al., 1998) and did not reveal significant changes between pre-, intermediate- and post-measurements. In other words, neither of the two 3-D movies was stressful enough to induce a significant change in this binocular characteristic.

Continuous assessment

Figure 32 presents the raw visual comfort scores averaged over the participants per Sub condition for the movies 'Sports' and 'Travel' respectively. The standard deviations (STD) and the time line of the subtitles are shown as well. Table 15 outlines the mean and standard deviation of the COA of each movie with and without subtitles.


CA of 'Sports' movie
Average visual comfort scores
Time (sec)
no sub CA
sub CA
no sub STD
sub STD
presence sub
0
100
200
300
400
500
600
700
800
900
1000

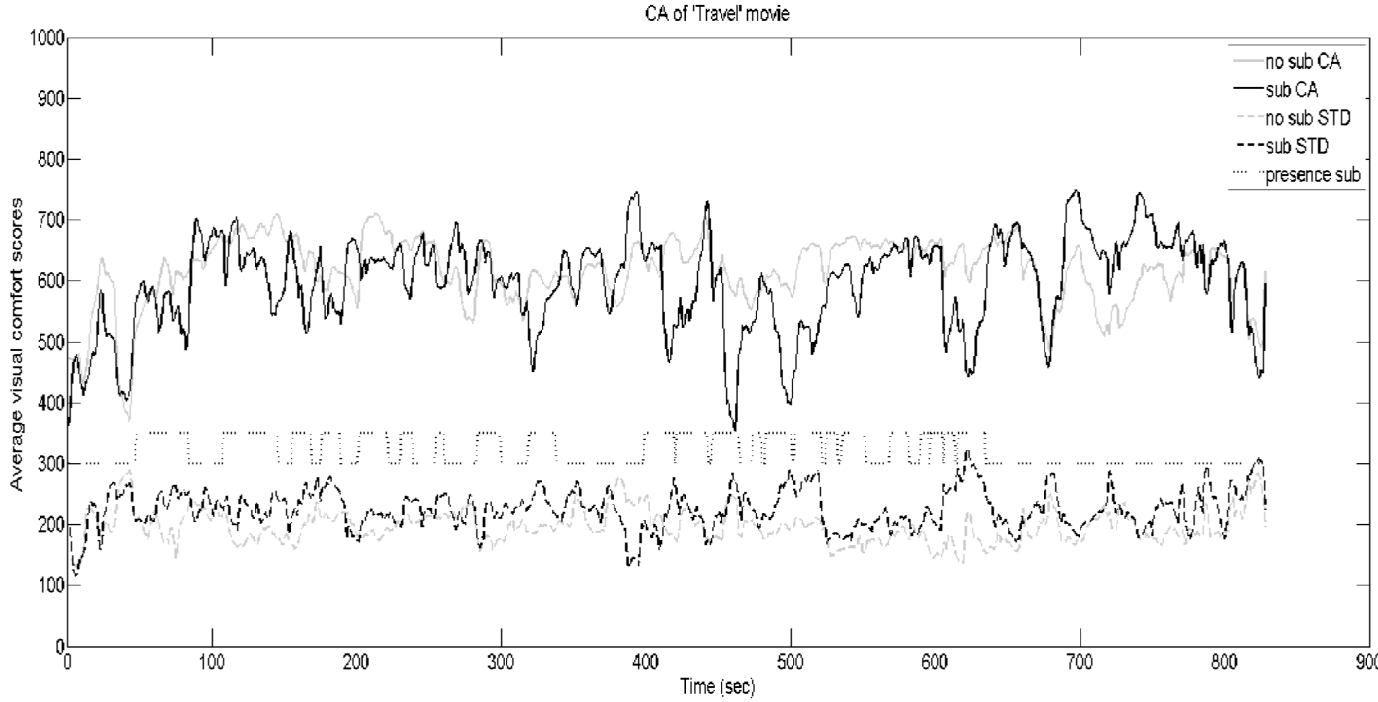


Figure 31. The raw visual comfort scores of the 'Sports' movie (upper) and 'Travel' movie (lower) averaged over the participants per subtitle condition including the overall standard deviation. The x-axes represent the elapsed time of the 3-D movie in seconds, the y-axes represent the total assessment range (a score of 1000 corresponds to comfortable and 0 to uncomfortable). The different lines represent the time line of the subtitles (thick solid line), the averaged scores for two different subtitle conditions and their accompanying standard deviation (STD).

TABLE 15
MEAN AND STANDARD DEVIATION OF THE COA OF THE 'SPORTS' AND 'TRAVEL' MOVIE WITH AND WITHOUT SUBTITLES

		Subtitle	
		yes	no
Movie	Sports	451 ± 211	553 ± 178
	Travel	597 ± 225	616 ± 196

Figure 32 shows that without subtitles (grey lines) the CA of the 'Sports' movie had more fluctuation (steep peaks and dips) than the CA of the 'Travel' movie (gradual peaks and dips). For the 'Sports' movie the inclusion of subtitles (black lines) seemed to flatten the peaks and dips in the CA, whereas the subtitles seemed to have the opposite effect for the 'Travel' movie.

Participants stated that the 'Sports' movie was difficult to follow because "a lot of stuff was happening" and "subtitles drew the attention from the content". The mean CA score was lower for both movies when subtitles were present (see Table II). For each movie an ANOVA with Sub as independent variable and the mean CA scores as dependent variables showed that Sub significantly reduced the comfort scores for the 'Sports' movie ($F(1, 48) = 5.701$, $p < .05$), yet not for the 'Travel' movie (p = .69). It must be noted, however, that it is questionable whether a mean CA score per participants is representative for the entire CA. In addition, Subs is a between-subject condition and since participants were asked to use the entire assessment range, these values need to be interpreted cautiously.

For both movies the STDs became higher with subtitles, which might indicate that not all participants assessed the content and its subtitles in a similar way. It must be noted that STDs in general are relatively high ranging (on average) from 178 to 225 (see Table 15). Peaks in the STD often occur at moments where the CA drops, indicating that video characteristics impact the visual discomfort of participants differently. The reliability tests revealed Cronbach's Alphas of 0.79 and 0.70 for the 'Travel' movie with and without subtitles, respectively, and Cronbach's Alphas of 0.76 and 0.48 for the 'Sports' movie with and without subtitle, respectively. This indicates a relative large inconsistency between the CA of participants of the 'Sports' movie without subtitles.

Analysis of the Questionnaire

Part 1: visual comfort

The CISS questionnaire (15 items) was used at all three stages (pre, intermediate, post; reflected by order) and was used to analyze how both 3-D movies were perceived in relation to conventional 2-D TV. Recall that in the pre-questionnaire only the first part of the questionnaire was evaluated (related to comfort) and it was explicitly stated that these items related to previous experiences while watching 2-D TV. Figure 32 visualizes how the different items responded on the Sub and Order variations including the 95% confidence intervals. Most of the items yielded their highest visual discomfort score when subtitles were present; even in that case, visual discomfort symptoms only occurred occasionally (score < 2).

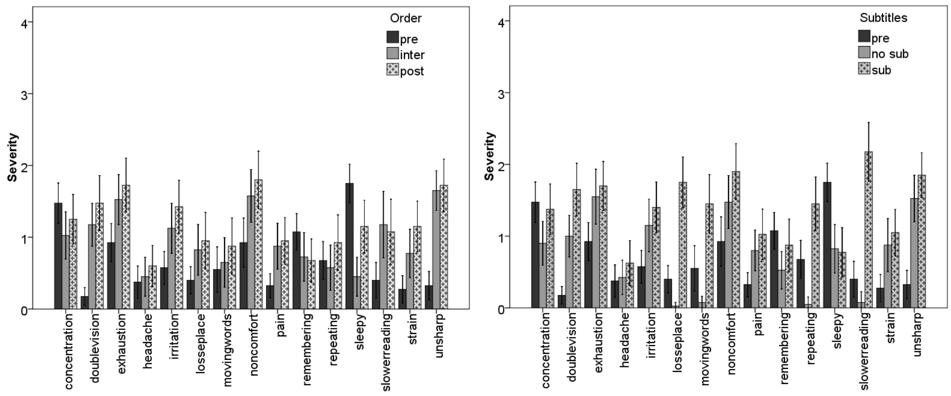


Figure 32. Questionnaire scores of the 15 items that relate to perceived visual discomfort including the 95% confidence intervals. The x-axis represents the different items, the y-axis the average questionnaire scores and the bars the pre-post conditions in the left graph and different Sub conditions in the right graph.

Since not all items were affected equally by changes in Order or Sub, a principal component analysis (PCA) was performed. The resulting PCA revealed three underlying factors that explained 43%, 19% and 11% of the variance of the data. Factor 1 received meaningful factor loadings (> .50 (Hair et al., 2006)) of the items *double vision*, *exhaustion*, *headache*, *irritated eyes*, *discomfort*, *pain in the eyes*, *strain in the eyes* and *sharpness*. Factor 2 consisted of *appearance of moving words*, *losing position in text*, *re-reading words* and *slower reading*. Factor 3 comprised of *loss of concentration*, *trouble remembering words* and *sleepiness*. Reliability testing revealed Cronbach's alphas of 0.89, 0.83 and 0.67 for factor 1, 2 and 3 respectively. An ANOVA was performed with Sub, Movie and Order as independent variables and factor scores as dependent variables. The items in factor 1, which incorporated the visual fatigue items, showed the impact of 3-D movies via a significant effect of the variable Order ($F(2, 111) = 14.799$, $p < .001$). The factor score largely increased between the pre- and intermediate-evaluation and only had a small increase between the intermediate- and post-evaluation. Factor 2, which incorporated the 'reading' items, revealed a significant difference between movies with subtitles (3-D Sub) and movies without subtitles (3-D no Sub) ($F(1, 111) = 139.996$, $p < .001$). Factor 3 did not reveal any significant effects.

Part 2: experiences

The intermediate- and post-questionnaires evaluated the 3-D movies in terms of their overall experience. The movie 'Travel' received lower scores in terms of *image quality* ($F(1, 72) = 4.027$, $p < .05$), higher in terms of *viewing experience* ($F(1, 72) = 8.112$, $p < .01$) and *3-D experience* ($F(1, 72) = 10.490$, $p < .01$), and similar scores in terms of *naturalness* ($p = .568$) as compared to the movie 'Sports'. The differences, however, were small, i.e., all four criteria for both movies scored (on average) between fair and good. None of the criteria was affected by the occurrence of subtitles.

Part 3: video characteristics

The intermediate- and post-questionnaires addressed which video characteristics (eight items) were incorporated in the CA. Figure 33 depicts the results of these eights items for both 3-D movies including the 95% confidence intervals.

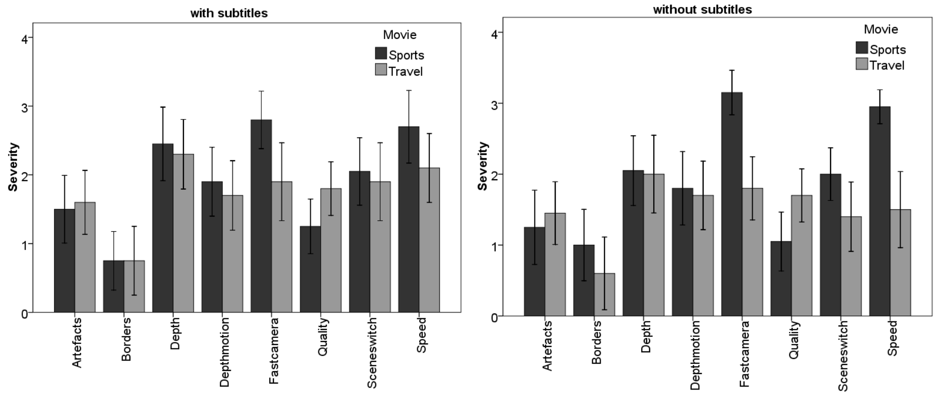


Figure 33. Questionnaire scores of the eight items that relate to the video characteristics that affect the COA including the 95% confidence intervals. The x-axes represent the different items, the y-axes the average questionnaire scores, the different lines both movies and the different graphs the Sub condition. The lines do not indicate correlations, but are incorporated for the ease of interpretation.

A PCA combined with a non-orthogonal rotation method (Oblimin) was performed, in which the item *border* was not included since it acted as floor effect, i.e., it was not incorporated in the CA, and decreased the reliability significantly. The resulting PCA revealed three underlying factors that explained 42%, 20% and 14% of the variance of the data.

Factor 1 received meaningful factor loadings of the items *fast camera*, *speed*, and *scene switches*. Factor 2 comprised the items *depth motion*, and *depth*, whereas factor 3 consisted of *quality* and *artefacts*. Reliability testing revealed Cronbach's alphas of 0.84, 0.79 and 0.74 for factor 1, 2 and 3, respectively. The items *fast camera*, *speed*, and *depth* received the highest score in the questionnaire, indicating that these items had a large impact on the assessment.

Video characteristics

The stereoscopic movies were converted to the WOWvx 2-D-plus-Depth format, i.e., a red-green-blue (RGB) image with a corresponding depth map (Redert et al. (2007). This format allowed extraction of specific video characteristics in the same four steps as those followed in Chapter 4.3.2. Results.

Now a COA score per second could be regressed to different combinations of motion and screen disparity derivatives. The initial regression was based on previous research and on the analysis of the third part of the questionnaire (video characteristics). The regression incorporated motion, screen disparity offset and range and changing screen disparity offset and range and did not include subtitles. The regression equation with the best fit, when incorporating only those predictors with a significant effect on the CA, is depicted in Equation 9:

$$COA = \alpha \cdot \overline{mo} + \beta \cdot \overline{de} + \delta(de_{max\,10\%} - de_{min\,10\%}) + \lambda\left[G(\overline{de})\right] + \gamma\left[G(de_{max\,10\%} - de_{min\,10\%})\right] \quad (9)$$

with

COA	= COA scores
$\overline{mo}$	= average amount of motion
$\overline{de}$	= average amount of screen disparity
$de_{max\,10\%} - de_{min\,10\%}$	= screen disparity range
$G(\overline{de})$	= gradient average amount of screen disparity
$G(de_{max\,10\%} - de_{min\,10\%})$	= gradient screen disparity range
$\alpha, \beta, \delta, \lambda, \gamma$	= weights of the predictors

The proportions of explained variances for each movie, however, were low (R^2 = .16 for the 'Travel' movie and R^2 = .10 for the 'Sports' movie). Since each movie comprised of several distinctive scenes, the model was fitted again per scene. Experiment 4 revealed that participants required some time in setting the correct CA value after scene switches. Based on this result, per scene the CA was shifted for a maximum of three seconds with respect to its predictors to improve the fit. Table 16 outlines the weights of each predictor and the corresponding R^2 values for each scene of the 'Sports' and 'Travel' movie. Table 16 shows that not all predictors were appropriate for each scene. Some predictors were redundant, indicated by the light-grey labels in Table 16. Redundancy refers to the fact that incorporating more predictors decreased the fit of the model, since the overlap in explained variance increased as a result of a high correlation between predictors. When a pair of predictors had a high correlation (> .70), the predictor that was labeled as redundant was that predictor of the pair that after exclusion resulted in the largest increase in R^2. Secondly, predictors could lack impact on the CA, which is indicated by the mid-grey labels in Table 16. Predictors that were not incorporated in the CA by participants had low weights and excluding them did not decrease R^2 significantly. Thirdly, predictors could have little or no variation, which is indicated by the dark-grey label in Table 16. When predictors lacked sufficient variation, they fell short of predicting the COA, resulting in low weights.

Even though redundancy, lack of impact and little variation in the predictors is accounted for, there is little consensus in weights between the various scenes, even within one movie. Not all scenes include all predictors and the predictors that are included differ significantly in size and polarity between scenes. Hence, Scene seems to act as a moderator on the weights. Since Scene was a categorical variable, dummy coding was applied to include Scene in the regression equation as a moderator. The resulting R^2 values were .87 for the 'Travel' movie and .78 for the 'Sports' movie. This indicates that the video characteristics can be used to predict visual comfort, yet differ between Scene. A more detailed analysis revealed that the predictors differed between scenes in their range, which affected how predictors related to the CA as well as how predictors related to each other. In general terms, this analogue can be mathematically represented by Equation 10.

TABLE 16
REGRESSION PARAMETERS FOR 'SPORTS' AND 'TRAVEL' MOVIE SEPARATELY.

Movie	Scene	Duration (sec)	α	β	δ	λ	γ	R^2
'Sports'	intro	83	-0.06*	0.37	0.8	-0.23	0.37	0.77
	football	239	-0.56	0.52	0.46	-0.05	0.22	0.64
	boxing	104	-0.29	-0.45	-0.42	-0.19	0.16	0.63
	racing	129	-0.34	0.77	0.41	-0.14	0.07	0.78
	basketball	203	-0.01	-0.85	-0.11	-0.18	0.05	0.65
	end	106	-0.37	-0.78	0.62	-0.15	0.05	0.79
'Travel'	intro	40	0.11	0.26	-0.69	-0.11	0.18	0.78
	villages	215	-0.24	0.4	-0.43	-0.08	0.07	0.83
	landscapes	267	0.70	-0.59	-0.64	0.02	0.09	0.79
	church	276	-0.98	0.01	0.08	-0.03	-0.08	0.90
	end	30	-0.08	0.15	1.14	0.02	0	0.90

redundancy
no impact
no variation

* a negative correlation indicates that when a video characteristic increases in occurrence, it decreases visual discomfort.

$$COA = \alpha_S \cdot \overline{mo} + \beta_S \cdot \overline{de} + \delta_S (de_{\max 10\%} - de_{\min 10\%}) + \lambda_S \left[B(\overline{de})\right] + \gamma_S \left[G(de_{\max 10\%} - de_{\min 10\%})\right]$$
(10)

with $\alpha_S, \beta_S, \delta_S, \lambda_S, \gamma_S$ = scene dependent weights of the predictors

4.4.3. Discussion

As stereoscopic cinema and television is rapidly gaining in popularity, a large number and diversity of people are enjoying stereoscopic content, or will be doing so in the near future. In order to guarantee an experience that is free of visual discomfort, research is needed to identify those factors that can be detrimental to the viewing experience. In this paper, we discuss elements of the stereoscopic video content that most negatively impact visual comfort. Using the continuous assessment method, we identify a number of motion and disparity related factors that influence visual comfort. Subsequently, we focus on how visual discomfort can be objectively predicted by video characteristics.

The objective measurement of the vergence facility did not reveal significant changes between pre-, intermediate- and post-measurements. The viewing periods of both 3-D movies were relatively short and the 3-D content had a average screen disparity well within the 1-degree limit that is expected to generate discomfort. Hence, the 3-D content was not really stressful, and a healthy visual system should be able to cope with these screen disparities at a typical TV viewing distance (3 meters). In addition, previous research already revealed that for non-moving images the vergence facility for large viewing distances lacks good specificity (the ability to correctly identify people without a specific condition) (Fortuin et al., 2010), and has poor repeatability (Gall et al., 1998). It is therefore not that surprising that vergence facility did not decrease after the movies. In future research other objective measurements to evaluate visual discomfort that might be more appropriate can be used (read Chapter 5).

With respect to the first hypothesis, we found that visual comfort could be predicted by specific video characteristics of 3-D movies, in which the contributions of these characteristics were scene-dependent. The experimental results indicated that participants incorporated two main video characteristics in the CA; one that was motion related (*fast camera*, *speed*, and *scene switches*) and the other that was disparity related (*depth motion*, and *depth*). These video characteristics, e.g., (change in) screen disparity offset and range and lateral motion, provided a good approximation for the experienced visual comfort using a simple linear combination resulting in R^2 values ranging from .65 to .90 per scene. The linear prediction model, however, was not sufficient to predict the visual comfort for the entire movie. More specifically, as a consequence of the inter-dependency between video characteristics and their non-linear relationship with visual comfort between scenes, more complex prediction models should be used. These results are in line with Experiment 4 in which the prediction of visual comfort by video characteristics depended on the activity in a scene. The fact that in the current research content with high spatial and temporal complexity is displayed on a full-HD stereo display, whereas in the previous research content with low spatial and temporal complexity was displayed on an autostereoscopic display with high degrees of crosstalk, suggests that the required complexity of the visual discomfort model holds for a broad range of content and displays. The same analysis was performed when subtitles were present in both 3-D movies, the results of which are not reported. Including subtitles altered the CA and thus also the weights of the predictors. Since we are not able yet to understand the impact of the motion- and depth-dependent predictors on instantaneous visual comfort, the results of the analysis with subtitles is even more complex and does not provide any additional insights at the moment. Note that similar problems arise when predicting image quality with objective image characteristics. More specifically, it is known which predictors are most important in estimating image quality (i.e., brightness, contrast, colour and sharpness), yet how these predictors need to be combined in such a way that they describe image quality for all sorts of content, remains unknown. Creating a general method for preemptive modification of video characteristics in order to maintain a comfortable viewing experience is not possible based on this work; more research is required. Hopefully, this result motivates more fundamental research, e.g. to determine which types of motion induce visual discomfort when different types of disparity (crossed vs. uncrossed) are present. However, a complete model is not necessary for all purposes. These results can be used to guide 3-D production companies to create more comfortable videos. For example, stereoscopic depth can be limited when 2-D motion is high. Or viewers can be made aware of the causes of potential visual discomfort.

With respect to the second hypothesis, it can be stated that including subtitles had a negative effect on visual comfort of stereoscopic movies. In addition, the reading of subtitles required more effort, which drew the attention away from the content. For dynamic content (e.g., fast camera and object motion, scene switches, changes of camera perspective, changes in depth) this makes it difficult to follow the movie, resulting in more double and/or unsharp images. For more passive content, it was stated by participants that is it not comfortable to perceive slow, steady background motion at a different depth layer than the subtitles, though future research should provide further clarification. In general, we believe that the impact of subtitles can be compared with their effect in stereoscopic stills, i.e., subtitles rendered at the display plane can induce depth cue conflicts by occluding the 3-D content. An unexplained aspect is that even though subtitles cause visual discomfort, they do not impact 3-D experience, naturalness, viewing experience and image quality. The results on the subtitles can be useful for video processing algorithms that can be implemented in consumer applications to sustain a comfortable 3-D viewing experience.

For example, screen disparity of the image content can be reduced in the area where subtitles are present, or subtitles can be presented with a small amount of screen disparity depending on the screen disparity of the image content they are occluding.

4.5 Conclusion

Viewing stereoscopic movies can reduce visual comfort as a consequence of multiple video characteristics. An experiment was performed with two main aims: 1) to identify these video characteristics reducing visual comfort, and 2) to use them to predict the experienced visual comfort. Additionally, the effect of subtitles was investigated, as they might cause a cue-conflict situation.

Results showed that participants incorporated lateral object and camera motion and (changes in) depth when evaluating visual comfort. For individual scenes, linear models of these video characteristics are sufficient to predict the visual comfort. This is not the case for an entire movie with different scenes, since these video characteristics are scene-dependent. In addition, subtitles remain an interesting problem; they reduce the visual comfort, yet do not affect other perceptions such as 3-D experience.

-CHAPTER 5-

A measurement protocol

"Although the exact role of vision in learning is a subject of intense debate, many scientific studies support a connection between vision and reading…. …. For about 50 percent of the time, the right eye is aimed about one of two letters to the right of the letters seen by the left eye. This doesn't present a problem to the reader because the image from the eyes are merged in the brain. What happens, however, if the two eyes register conflicting information? Since I was cross-eyed, I cross-fixated. When I was learning to read, my right eye saw letters located to the left of the letters I saw with my left eye. Although I am not dyslexic, I distinctly remember being in first grade and trying to figure out whether I was reading "saw" or "was".
She continues to argue her statement in more detail in one of her lectures. "If I looked at word, let's say pick out of a head the name "Sue". If I looked at the word with my right eye, I might see the "S" in the same place my left eye saw the "U". So was I seeing "Sue" of was I seeing "Use". I was not dyslexic. My guess is a lot of children out there who have problems learning to read, probably have binocular vision problems. A lot of children have binocular problems that are a little bit more subtle than mine, that remain unnoticed."

Susan Barry (Barry, 2009)

ABSTRACT

Some people have a sensitive binocular vision and their visual system is challenged too much under unnatural viewing conditions. Consequently, they may experience visual discomfort regardless whether visual discomfort factors are absent in the 3-D content. To adequately characterize the degree of objective visual fatigue and subjective visual discomfort, a measurement protocol was constructed that also incorporated differences in performance of the binocular system. Participants were categorized by their binocular vision; one subgroup with good binocular vision and one with moderate binocular vision. Results show that a combination of fusion range measurements and self-report is appropriate for evaluating visual complaints. The results also show that the ratio of number of words read in 3-D relative to number of words read in 2-D, is indicative of the binocular vision of people. This ratio is very promising, since 1) it can be used to explain inter-subject differences in results, 2) it provides additional information to stereo- and visual acuity tests, and 3) it can be implemented in 3D commercial displays to advise people with moderate binocular vision to lower the screen disparity range in order to avoid visual complaints.

Experiment 6 is based on:
Lambooij, M., Fortuin, M., IJsselsteijn, W. A., Evans, B. J. W., and Heynderickx, I. (2010). Measuring visual discomfort associated with 3-D displays. *Journal of Society of Information Displays*. 18:931-943.

Experiment 7 is based on:
Lambooij, M., IJsselsteijn, W.A., Fortuin, M., Evans, B. J. W., and Heynderickx, I. (2011). Susceptibility to visual discomfort of 3-D displays by visual performance measures. *IEEE Transactions on Circuits and Systems for Video Technology*. 21.

Experiment 8 is based on:
Lambooij, M., IJsselsteijn, W. A., Fortuin, M., and Heynderickx, I., (2011). Reading performance as screening tool for visual complaints from stereoscopic content. *Displays: special issues 3-D environment*. submitted for publication.

5.1. Experiment 6: Measuring Visual Fatigue and Visual Discomfort

ABSTRACT

Some people report visual discomfort when watching 3-D displays. A goal of this research was to identify methods to collect for objective signs of visual fatigue and subjective symptoms of visual discomfort. Previous research yielded contradictory results concerning such indicators. We hypothesize two potential causes for this: 1) not all clinical tests are equally appropriate to evaluate the visual discomfort associated with stereoscopic viewing, and 2) there is a natural variation in susceptibility to visual discomfort among people with normal vision.

To investigate these hypotheses, we designed an experiment, consisting of two parts. First, an optometric screening was used to differentiate participants with a moderate binocular status (MBIS) from those with a good binocular status (GBIS). Second, in a 2x2 within-subjects design (2-D vs. 3-D and MBIS vs. GBIS), a questionnaire and eight optometric tests (i.e. binocular acuity, aligning prism, fixation disparity, heterophoria, convergent and divergent fusional reserves, vergence facility, and accommodation response) were each administered before and immediately after a reading task.

Results revealed that only participants with a MBIS in 3-D conditions showed a clinically meaningful change in fusion range, experienced more visual discomfort and performed worse on the reading task. Our results indicate that a combination of fusion range measurements and self-report is appropriate for evaluating visual discomfort of stereoscopic stills and that people with an MBIS are more susceptible to visual discomfort associated with stereoscopic displays. We also propose that a simple measurement tool, i.e., the ratio of reading performance between 2-D and 3-D, can be appropriate to categorize people based on their binocular status.

5.1.1. Introduction

The main objective of the experiment is to determine the most appropriate indicators for visual fatigue and visual discomfort. In Chapter 3 it was stated that visual fatigue refers to a decrease in the performance of the binocular visual system as a consequence of physiological strain or stress resulting from excessive exertion. In theory, it can be quantified objectively. Visual discomfort is its subjective counterpart and is expected to provide an indication of visual fatigue. It is important to incorporate the evaluation of visual discomfort, since changes within the binocular visual system do not necessarily indicate visual fatigue. Only physiological changes that decrease the performance of the binocular visual system or that are accompanied by the experience of visual discomfort should be critically examined for their magnitude and subjective impact.

To determine the degree of visual fatigue and visual discomfort in a sensitive, accurate, reliable and valid way, multiple indicators for both components can be relevant. Indicators for visual discomfort can be provided by validated optometric questionnaires or other self-report measurements. Indicators for visual fatigue can be provided by clinical optometric measurement methods. These indicators for both visual fatigue and visual discomfort should be relatively easy to apply in evaluative settings and fulfil a number of additional requirements (Chapter 3). First, in order to address the impact of stereoscopic depth on the binocular visual system, the indicators should be able to distinguish stereoscopic viewing conditions from conventional monocular viewing conditions. Second, measurements should be relatively fast as the recovery of the binocular visual system is usually quite rapid. Third, the indicators should apply to different types of displays, e.g., autostereoscopic systems and systems based on polarised or shuttered glasses. And finally, measurements themselves should not require too much visual effort or induce visual fatigue or visual discomfort on their own.

Ideally, we would like to arrive at general indicators of visual fatigue and visual discomfort that can be implemented easily. When a robust relationship is established between visual discomfort and visual fatigue indicators, one indicator can be used to substitute for the other. This would allow for example, the evaluation of the binocular visual system for large groups of participants with simple subjective questionnaires, or the use of relatively simple objective measurements.

5.1.2. Background

Previous research already applied clinical optometric measurement methods, in combination with questionnaires, to determine the effect of stereoscopic devices on the visual system. The results, however, were not conclusive. Peli (1998) compared monoscopic and stereoscopic head-mounted displays (HMD) and a regular CRT on potential harmful effects to the visual system. In his study, the binocular disparity values did not exceed one degree, implying that he remained within the accepted zone of comfortable viewing as derived in Chapter 3 (Peli, 1998; Speranza, Tam, Renaud and Hur, 2006; Yano, Emoto and Mitsuhashi, 2004; Wöpking, 1995; Hiruma and Fukuda, 1993; Emoto, Nojiri and Okano, 2004; Iwasaki, Kubota and Tawara, 2009). He used a set of objective indicators (e.g. refraction, visual acuity, fixation disparity and fusion measurements) and a subjective questionnaire as pre- and post-measurements. Although no objective indicator revealed a significant or clinically meaningful effect of any of the display types, almost all items on the questionnaire indicated lower comfort scores for the stereoscopic HMD than for the other displays. Emoto, Nojiri and Okano (2004) evaluated changes within the binocular visual system as a consequence of viewing still images for 60 minutes in monocular and stereoscopic mode. Visual fatigue was evaluated using a pre- and post-measurement of the fusion range and the Accommodative Convergence / Accommodation ratio (AC/A ratio), i.e., the change in vergence due to accommodation per change in accommodation. Visual discomfort was evaluated via a questionnaire and a free-form in which participants could give their comments. To determine possible short term after-effects, fusion range was also measured after five and ten minutes rest. The results were related to the ability of the participants to free-fuse, because this ability might indicate that participants are accustomed to visual-perceptual problems related to stereoscopic image material. No differences were found between the pre- and post-measurements of the AC/A ratio, but the fusion range of participants who were unable to free-fuse significantly decreased in the convergent direction after stereoscopic viewing. The questionnaire revealed that five of the twelve participants experienced more visual discomfort, one experienced less visual discomfort and for six participants visual discomfort was similar under both stereoscopic and monoscopic viewing conditions. The subjective ratings were not related to the ability to free-fuse. More recently, Emoto, Niida and Okana (2005) performed an experiment to evaluate the visual fatigue of a one-hour movie stereoscopically, monocularly, and in a simulated stereoscopic condition. The simulated stereoscopic condition consisted of viewing monocular content through prisms. A prismatic manipulation evokes a whole-field vergence manipulation or change of absolute disparity, in contrast to 3-D content that incorporates relative disparity, i.e., disparity differences between objects within the retinal images. The strength of the prisms was set according to each participant’s individual Percival's area of comfort, which describes the range of prism loads that are conceptualised as unlikely to induce any discomfort [Peli, 1998; Sheard, 1934]. Visual fatigue was measured with a pre- and post-measurement of the fusion range and the accommodation response, whereas visual discomfort was measured with a post-questionnaire. Only the result of the item 'severe eye fatigue', was incorporated and indicated more visual discomfort when the prism load was varying or beyond Percival's area of comfort. Both the accommodation response and the fusion range were affected significantly by conditions with varying disparity in (simulated) stereoscopic conditions, whereas the fusion range also decreased significantly with fixed prism loads beyond as well as within Percival's area of comfort. No visual discomfort was perceived in the latter, which suggests that these changes in fusion range indicated functional adaptations to altered viewing situations. More recently, Pölönen, Salmimaa, Alltonen, Häkkinen and Takatalo (2009) performed subjective evaluation of the satisfaction of the full-feature cinematic stereoscopic movie U2-3-D. Although no scenes appeared in the movie with a high degree of disparity, fast changing disparity or strong apparent motion, some viewers experienced a mild degree of discomfort. Results, however, also revealed that some individuals were more susceptible to visual discomfort since they seemed to be extremely sensitive to 3-D-content.

5.1.3. The current study

Previous research revealed little consensus both between, as well as within, the indicators of visual fatigue and visual discomfort. Also large individual differences were revealed across the various studies. Since differences within single experiments also were revealed, variations in set-up of the experiments is not the only cause. We hypothesize two potential causes for these contradictions. The first cause is that not all tests are equally appropriate to evaluate the effect of stereoscopic viewing on visual fatigue and visual discomfort.

Stereoscopic viewing can impact on multiple components of the binocular system, e.g., vergence, accommodation and fusion, for which different optometric indicators are used. Since recovery from both visual fatigue and visual discomfort is expected to take place rapidly after viewing stereoscopic content, it makes no sense to compare all indicators after a long-term 3-D video, and it is too time-consuming to let each participant view one long-term video per indicator. Hence, a first logical step is to evaluate the impact of highly stressful short-term stereoscopic stimuli on multiple potential appropriate indicators for visual fatigue and visual discomfort to determine which ones are most appropriate. The second cause lies in the difference in binocular status between the participants. People differ in human binocular visual system characteristics, which directly determine their ability to perceive stereoscopic depth (Richards, 1970; Patterson & Fox, 1984). There is a paucity of epidemiological data on the prevalence of binocular vision anomalies, but these anomalies have been found to affect approximately 20% of patients consulting community optometrists, though somewhat deviating percentages have been reported as well (especially for people who often perform near work, higher percentages have been revealed) (Ciuffreda, 2002; Evans, 2007; Karania & Evans, 2006; Montes-Mico, 2001; Richman & Laudon, 2002; Scheiman, 1996; Stidwill, 1997). A distinction needs to be drawn between binocular anomalies that typically prevent stereopsis (principally, strabismus and amblyopia, i.e., squint and lazy eye respectively) and non-strabismic binocular anomalies that permit stereopsis but predispose the patient to visual discomfort (asthenopia). The prevalence of strabismus is about 2% (Williams at al., 2008) and of amblyopia about 3% (Kanodidou, 2011).This thesis is principally concerned with non-strabismic conditions. The associated visual discomfort of non-strabismic binocular anomalies, which may not be present in normal viewing situations, can become present or more severe in unnatural viewing situations, e.g., viewing stereoscopic content. Hence, some people can be more susceptible to visual discomfort than others, based on their binocular visual functioning. The purpose of the current study is to investigate both of these hypothesised causes. Hence, the first objective is to identify indicators for visual fatigue and visual discomfort that are appropriate to evaluate the impact of stressful short-term stereoscopic stimuli. The second objective is to investigate the direct relationship between binocular status of people with normal vision and the occurrence of visual fatigue and visual discomfort as a consequence of stereoscopic viewing. With 'normal vision' I mean vision without any associated visual discomfort of non-strabismic binocular anomalies in normal viewing situations (see section 6.4). A third objective is, assuming a relationship between binocular status and visual fatigue is present, to indentify an indicator that enables the categorization of participants based on their binocular status.

5.1.4. Experimental set-up

Design

The experimental design consisted of two steps: (1) an optometric screening (1 session), and (2) the measurement of visual fatigue and visual discomfort (2 sessions). The optometric screening facilitated differentiation of participants based on their binocular status (BIS): good BIS (GBIS) and moderate BIS (MBIS). The two sessions of the actual experiment differed in Dimension (2-D and 3-D), the order of which was randomised across participants. In each session visual fatigue and visual discomfort were evaluated as a pre- and post-measurement (Pre-Post); eight different optometric indicators were each administered before and immediately after a stimulus and a questionnaire containing 15 subjective items was administered before and after the first stimulus and after the entire session. Different versions of the Wilkins Rate of Reading test (WRRT) were used as stimuli. The screening session and each experimental session were scheduled on different days. In total, 47 participants completed the experiment.

Screening of participants

Prior to the experiment, an extensive optometric screening was carried out on 50 naive participants. This screening was performed for two reasons: (1) to exclude participants with eye diseases or severe binocular abnormalities (e.g., strabismus), and (2) to differentiate participants on their binocular status. An additional benefit was that since some of the screening tests were also used in the actual experiment, participants were familiarized with the optometric tests.

The indicators with their exclusion criteria are outlined in Table 17 and contained three subjective questionnaires (the first three tests in Table 17) and 15 objective indicators. The three questionnaires were the Convergence Insufficiency Symptom Survey (CISS) (Borsting, Rouse, Mitchell, Scheiman, Cotter, Cooper, Kulp and London, 2003), the Dry Eye Questionnaire (DEQ) (Schiffman, Christianson, Jacobsen, Hirsch and Reis, 2000), and the Visual Function Questionnaire (VFQ) (Mangione, Lee, Gutierrez, Spritzer, Berry and Hays, 2001). The CISS questionnaire was also used in the main experiment because it incorporates items that relate to visual discomfort in general (Sheedy, Hayes and Engle, 2003) and items to visual discomfort associated with reading tasks specifically.

For the objective indicators, a short explanation is given here, and for a more extended description, see Evans (2007). Visual acuity measures the smallest detail that can be resolved. Refractive error refers to the defocus of the eye. Stereopsis indicates the ability to perceive stereoscopic depth. Fixation disparity (or accuracy of vergence) is an angular measure of small misalignments of the visual axes of the eyes (measured here without a central fusional lock) and aligning prism, which was only applied in the main experiment and not in the screening, is the prism load that corrects the lesser fixation disparity that occurs in the presence of a central fusional lock (Figure 34). Heterophoria refers to misalignments of the visual axes of the eyes under dissociated conditions, i.e., viewing without any (central or peripheral) fusional lock. The cover test is one specific measurement of heterophoria, i.e., a dissociated situation created by covering one of the two eyes. Convergent and divergent fusional reserves indicate the amount of convergence and divergence respectively that can be induced before fusion is compromised. They are commonly characterized with prism loads at which binocular single vision is lost (break point) and recovered again (recovery point). The near point of convergence refers to the closest distance to which the eyes can converge, and the accommodation amplitude refers to the closest distance to which the eyes can focus. The accommodation response refers to the focus control of the eye, i.e., the accuracy of the accommodation. Accommodative and vergence facility indicate the ability to respond efficiently and accurately to changing demands over time of the accommodation and vergence systems respectively. Accommodative and vergence facilities are measured in cycles per minute that an object can be properly focused through alternating opposing lenses or be properly fused through alternating base-in and base-out prisms respectively. The slit lamp microscope is used to examine the eye and was used in this research to assess the extent of dry eyes. For more detailed results of each optometric test see Fortuin, Lambooij, IJsselsteijn and Heynderickx (2010).

TABLE 17

THE OPTOMETRIC INDICATORS, BOTH OBJECTIVE AND SUBJECTIVE, THAT WERE APPLIED IN THE SCREENING INCLUDING THEIR MEASUREMENTS ASPECTS AND THEIR EXCLUSION CRITERIA.

Indicators	Measurement aspects	Exclusion criteria
CISS	Convergence Insufficiency Symptom Score with 15 items	
DEQ	Dry Eye Questionnaire with 11 items	
VFQ	Visual Functioning Questionnaire with 25 items	
visual acuity	monocular and binocular	< 0.8 for both monocular and binocular
refractive error	monocular and between eyes	> 1 Diopter (D)
stereopsis	Randot Stereotest	> 60"
fixation disparity	angular adjustment to align Nonius markers without fusion lock	
heterophoria	Maddox rod (large distance)	vertical deviation > 1 Prism Diopter (PD)
heterophoria	Maddox wing (short distance)	vertical deviation > 1PD
cover test	(near and/or far) heterophoria or heterotropia (strabismus)	if strabismus*
convergent fusion range	break-recovery method (large distance)	Sheard's criterion**
divergent fusion range	break-recovery method (large distance)	Sheard's criterion**
near point of convergence		
accommodation amplitude	Binocular push up test	
accommodation response	MEM-retinoscopy	
accommodation facility	accommodation flipper binocular (± 2 D)	
vergence facility	prism flipper binocular (3 ΔBI /12 ΔBO)	
slitlamp microscope	fluorescein staining/ Break Up Time tearfilm / blepharitis	Grading: degree ≥ 2 (out of 4)

* strabismus is defined as a permanent deviation between the two eyes, i.e., a squint

* Sheard's criterion refers to a fusion range that still allows comfortable viewing (note that since breakpoints were measured instead of blur points the criterion was slightly modified (Evans, 2007))

Based on the exclusion criteria, three of the 50 participants were excluded from the experiment, resulting in 47 participants. Even though the binocular visual system of the remaining 47 participants was characterized as normal, there are undoubtly performance differences among the participants. Hence, based on their binocular status some participants can experience visual discomfort more severely or more rapidly while viewing stereoscopic content. The latter was investigated by constructing a subgroup by two means: 1) based on a modified algorithm proposed by Evans (1997, 2007) to evaluate the binocular status that underlies the degree of decompensated heterophoria, and 2) based on measurements that deviate more than one standard deviation from norm values for each optometric test (Fortuin, Lambooij, IJsselsteijn and Heynderickx, 2010). Both constructs were applied on the screening data only.

The algorithm of Evans (1997) evaluates people's binocular status that underlies degree of decompensated heterophoria. Decompensated heterophoria refers to misalignment of the visual axes of the eyes that is not compensated by fusion mechanisms. Efforts required to resolve excessive disparity, vergence or a combination of both, can lead to a decompensated phoria in a subgroup of individuals with a moderate binocular status (Evans, 2007). Especially in unnatural viewing situations, e.g., viewing stereoscopic content, fusion mechanism can become inadequate to compensate heterophoria, and people can experience visual discomfort. Since we are interested in the same subgroup of people, i.e., with moderate binocular status, the algorithm can be an appropriate method to distinguish participants based on their binocular functioning. The algorithm is outlined in Table 18. It computes a score representative for the binocular status. This score is the accumulative value of ten single scores that each relate to the result of an optometric measurement of binocular instability. The algorithm was modified by using fixation disparity instead of aligning prisms for numbers 4 and 5 and by excluding indicators numbers 6 and 8. Foveal suppression (nr 6) concerns a test for close viewing distances and as such was not appropriate for our experiment (the viewing distance in our experiment was three meters). Percival's criterion (nr 8) is defined as the middle third of the "zone of clear, single binocular vision" (Sheard, 1934; Evans, 2007). This zone can be determined by measuring the blur points of our fusion range, i.e., the points at which *clear* vision is lost. In the screening, however, the fusion range was determined by measuring the break points, i.e., the points at which binocular *single* vision is lost. Even more, Chapter 3 revealed that Percival's criterion is not suited for large viewing distances. The total algorithm score (excluding nr 6 and 8) ranged from 0 to 13; a score of 4 or lower could be labelled as a moderate binocular status (Evans, 1997). When using this threshold, the algorithm divided the participants into 41 who had a good binocular status (GBIS) and 6 who had a moderate binocular status (MBIS).

Although the Evans algorithm is designed to summate the risk factors for decompensated heterophoria, it is not likely to be a perfect indicator in relation to visual discomfort associated with screen disparity, which manifests idiosyncratically in different people (Evans, 2007). The second approach to categorize participants on their binocular status was based on indicator values in the screening that deviated more than one standard deviation of general normative values (Peli, 1998; Fortuin et al., 2010; Gall, Wick and Bedell, 1998). In total six indicators were found with such deviations of individual values; heterophoria, convergence fusion range break and recovery value, divergent fusion range break value, fixation disparity without fusion lock and vergence facility. Participants were assigned a MBIS if four or more of these six indicators had values that deviated more than one standard deviation from the norm. When using this threshold, this approach identified seven participants with a MBIS of which five had already been identified by the previous algorithm. The combination of both approaches, i.e., based on the algorithm and norm values, divided the participants in 38 with GBIS and 9 persons with MBIS. Neither the participants, nor the experimenter knew to which subgroup a participant belonged to during the experiment.

TABLE 18

ALGORITHM PROPOSED BY EVANS (1997) APPLIED TO DIFFERENTIATE BETWEEN PARTICIPANTS WITH RESPECT TO THEIR BINOCULAR STATUS BASED ON OBJECTIVE AND SUBJECTIVE OPTOMETRIC INDICATORS. INDICATOR NUMBER 4 AND 5 WERE ALTERED AND NUMBER 6 AND 8 WERE EXCLUDED FROM THE ALGORITHM AS EXPLAINED IN THE TEXT.

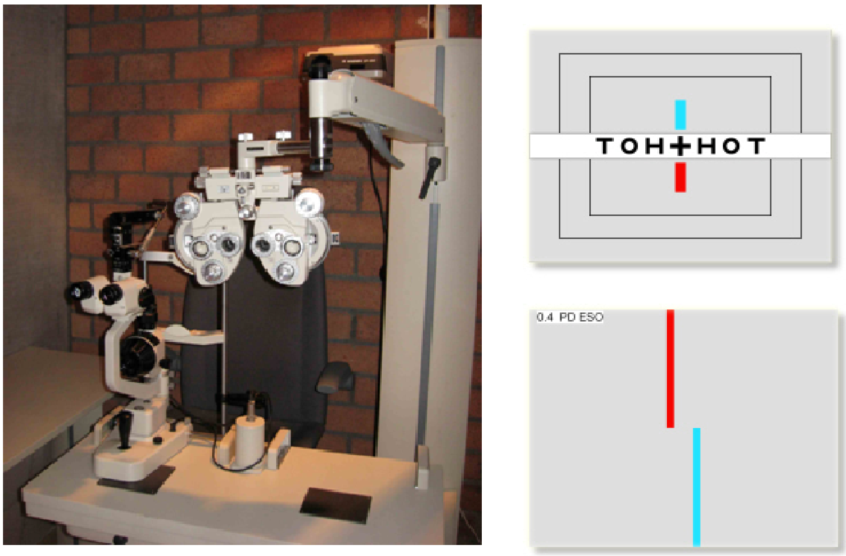
Nr	Indicators	Sign or symptom	Score
1	symptomatic heterophoria	questionnaire symptoms: VFQ (if so, +3 or +2 or +1 if borderline)	3
2	cover testing	heterophoria detected	1
3	cover testing	recovery rapid and smooth (if so, +2 or +1 if borderline)	2
4	fixation disparity	>1.2 prism diopter patients under 40* (Figure 34B)	2
5	fixation disparity	stable	1
6	*foveal suppression*	*foveal suppression (Mallet): >3'*	*2*
7	Sheard's criterion	failed	2
8	*Percival's criterion*	*Percival's criterion*	*1*
9	dissociated heterophoria	unstable (if > 4 prism diopter)	1
10	fusion range	< 20 prism diopter	1

* Since normative values were not known to us for large viewing distances measured with the Test Chart 2000, one standard deviation was applied as a cut-off, which corresponded to 1.2 prism diopter.

Equipment

Figure 34A is a photograph of the work unit that was used for the optometric measurements during the screening. It included a control console, an examination chair, a double sliding instrument table, a projector column and a phoropter. The phoropter arm contained prisms and lenses to evaluate the visual functions described in the screening. The program Test Chart 2000 was used to generate visual stimuli to facilitate the measurements of the visual functionality of the participants both during screening and experiment. It included a range of vision assessment tools of which two are illustrated in Figure 34B. During the screening these tools were displayed via the projector column and during the experiment on a separate CRT monitor placed at the same distance as the stereoscopic displays (three meters).

The stimuli for the actual experiment were rendered on a stereoscopic display that consisted of two 20" CRT monitors with a resolution of 720 x 576 mounted perpendicularly to each other. The dual monitor system displayed the right and left image at the same time using a semi-transparent mirror in between and a polarization filter in front of each screen. Polarised glasses were required to direct the correct view to the correct eye with very little cross-talk in the stereo pair (less than 0.1% with linear polarization filters) (Pastoor and Wopking, 1997).



A B

Figure 34. A) Photograph of the workstation used during screening measurement including a control console, an examination chair, a double sliding instrument table, a projector column and a phoropter. B) Screenshots of two visual assessment tools of the program Test Chart 2000. The upper is used for aligning prism with fusion lock combined with Risley prisms and the lower for fixation disparity without fusion lock with adjustable nonius lines.

Stimuli

The stimuli were different passages of the Wilkins Rate of Reading test (WRRT; Wilkins et al., 1996) that were randomly assigned to conditions. Visual deficits can impair reading ability, which can be properly assessed with the WRRT (Wilkins et al., 1996; Wilkins, 2002; Jeanes et al., 1997; Kriss and Evans, 2005; O'Leary and Evans, 2006). It was originally developed to assess the reading speed alleviated by using coloured filters of a specific tint in reading (Wilkins et al., 1996). The WRRT consists of a meaningless passage of seemingly random words; ten lines with on each line the same 15 words distributed randomly ("you for the and not see my play come is look dog cat to up") that participants were asked to read 'out loud' as rapidly as possible for 60 seconds. Since common simple words are used, poor readers can perform the task. The text is independent of any syntactic and semantic constraints and because participants do not know which words come next this requires them to keep the text in focus. Another consequence of its meaningless character is that readers do not have a sense of failure when making errors. Figure 35 depicts a screen shot of a stimulus. Since the screen size was too small to display both the entire left- and right-eye images in 3-D mode, the stimuli were slightly modified. Instead of 15 words per line, 10 words per line were distributed randomly. Only the text was presented with stereoscopic depth, whereas the frame with the circles was presented at zero disparity. The frame was added around the periphery to improve the perception of stereoscopic depth and facilitate faster and easier fusion. The horizontal visual angle of the text was 6.27 degrees and the vertical visual angle of the text, inner and outer frame were 4.44, 4.95 and 5.75 degrees respectively. The amount of relative screen disparity between the WRRT and the frame in the 3-D condition was set to 1.5 degrees, since stressing the visual system is the simplest way to evaluate its relationship with asthenopia (note that one degree limits the zone of comfortable viewing). This resulted in the text floating at approximately 133 cm in front of the viewer (and 167 cm from the display plane), whereas the surrounding frame with circles was displayed at the display plane (at 300 cm).

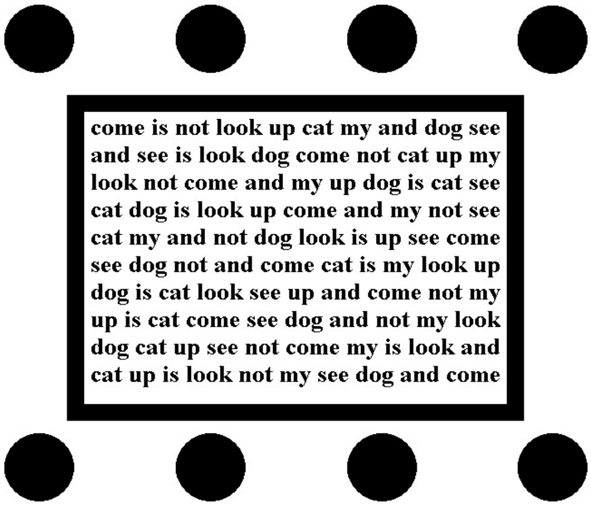


Figure 35. Screen shot of a stimulus; a modified passage of the Wilkins Rate of Reading test (Wilkins et al., 1996).

Questionnaire items

The Dry Eye Questionnaire consist of the three parts that each address specific eye problems experienced during the last week. The first part includes five asthenopic items (*sensitive to light*, *gritty eyes*, *painfail or sore eyes*, *blurred vision* and *poor vision*), the second part includes four items relating to problems with eye performance (*reading*, *driving at night*, *working with a computer* and *watching TV*), and the last part refers to unpleasant situations (*windy conditions*, *areas with low humidity* and *areas with air-conditioning*). All items have to be assessed on a scale labelled with adjective terms [all the time]-[most of the time]-[half of the time]-[some of the time]-[none of the time].

The Visual Function Questionnaire-25 addresses 12 subscales: *general health* (1 question), *general vision* (1 question), *near vision* (3 questions), *distance vision* (3 questions), *driving* (2 questions), *peripheral vision* (1 question), *colour vision* (1 question), *ocular pain* (2 questions), *role limitations* (2 questions), *dependency* (3 questions), *social function* (2 questions), and *mental health* (4 questions).

The Convergence Insufficiency Symptom Survey consists of 15 items including general asthenopic items as well as specific items concerning reading: *discomfort*, *loss of concentration*, *double vision*, *sleepiness*, *sharpness*, *exhaustion*, *appearance of moving words*, *slower reading*, *losing position in text*, *trouble remembering words*, *re-reading words*, *headache*, *pain in the eyes*, *strain in the eyes* and *irritated eyes* (Borsting et al., 2003). All items had to be assessed on a scale labelled with the adjective terms [never]-[seldom]-[occasionally]-[often]-[always], which were transformed into numerical values ranging from zero to four.

Procedure

Participants were provided with an informed consent statement containing information about the screening, the experiment, and the possible occurrence of visual discomfort. In addition, the statement included a version of the WRRT to familiarize participants with the specific reading task. After signing the statement, the participants proceeded with the optometric screening. The tests applied in this screening are outlined in Table 19, and took about 45 minutes to complete.

Those participants who completed the screening successfully (n=47), participated in the experiment. They were seated at a viewing distance of three meters and received a brief instruction about the course of the experiment. All questions concerning the procedure were answered, after which the experiment started. Table 19 provides an overview of the procedure. The first column outlines the order of the subjective and objective indicators. Participants filled in the CISS questionnaire twice; once after the first stimulus and once after the last stimulus (note that during screening the CISS questionnaire was also used to assess the subjective state of the participants prior to the experiment). In between eight objective indicators were measured, each before and after a stimulus (i.e., one of the passages of the WRRT). After each post-measurement a rest of at least one minute was inserted. The participants were asked to read the text 'out loud' as rapidly as possible for 60 seconds. The number of words correctly read in one minute was noted. The indicators that were used are incorporated in the second column of Table 19. Their order which depends on the amount of effort each indicator induces, is described in the third column of Table 19. The objective indicators were divided into three blocks (as indicated in the first column), to avoid visual fatigue induced by the tests themselves as much as possible. Since the indicators in the first block did not require any visual effort, they were used to start with and were randomly administered within this block. The order of the indicators in the second block depended on the participant’s direction of the heterophoria, i.e., convergent or divergent. The participants first performed the fusion test in the direction opposite to their heterophoria (for compensation), followed by the fusion test in the same direction as their heterophoria. The two indicators in the last block could require some visual effort, and therefore, were postponed to the end of the experiment. They were again mutually randomised within their block. The fourth column of Table 19 provides the tests used to measure the indicators. For a description of these tests see Evans (2007). The participants performed the experiment two times; both in a 2-D and a 3-D session that were scheduled on different days of which the order was counterbalanced. Each session lasted about 45 minutes.

5.1.5. Results

The step-size in screen disparity from 0 to 1.5 degrees appeared too large for eight participants. They were unable to fuse the 3-D stimulus and their data were not incorporated in the analysis, which resulted in 32 participants with GBIS and 7 participants with MBIS. An ANOVA was used to analyze the effects of BIS, Dimension and Pre-Post on the objective, subjective and performance measurements.

TABLE 19

OVERVIEW OF THE PROCEDURE INCLUDING THE ARRANGEMENT OF OBJECTIVE AND SUBJECTIVE INDICTORS, THE OBJECTIVE INDICATORS AND THEIR ORDER AND THE TESTS USED TO MEASURE ALL INDICATORS.

Test order	Indicators	Randomization	Tests	
Pre-subjective	-questionnaire (after first WRRT)		-CISS	
Block 1 Objective	-visual acuity binocular	randomised within block 1	-Log Mar	Test Chart 2000
	-aligning prism (with fusion lock)		-Risley prism	
	-fixation disparity (without fusion lock)		-adjustment of nonius lines	
Block 2 Objective	-heterophoria		-Maddow rod*	Test Chart 2000
	-divergent fusion range		-break / recovery by Risley rotary prism	
	-convergent fusion range			
Block 3 Objective	-vergence facility	randomised wihtin block 3	-prism flipper	
	-accommodation response		-MEM-retinoscopy	
Post-subjective	-CISS (after last reading task)		- Questionnaire	

* not evaluated with Test Chart 2000

Subjective visual discomfort

A few aspects of the analysis are noteworthy. First, most of the CISS items yielded a low average visual discomfort score, referring to moderate levels of visual discomfort. Second, not all items were affected equally by the variations in Dimension, BIS and Pre-Post. A principal component analysis (PCA) was performed on the 15 items to find intrinsic correlations and to reveal if certain items shared similar underlying attributes of visual discomfort (Hair, Black, Babin, Anderson and Tatham, 2006). The PCA was combined with a non-orthogonal rotation method (Oblimin) to minimize the number of items with high factor loadings on more than one factor (i.e., a linear combination of the original 15 items). The resulting PCA revealed two underlying factors that explained 48% and 20% of the variance of the data. Factor 1 received meaningful factor loadings (>.50 (Hair et al., 2006)) of the items *discomfort*, *double vision*, *movement of words*, *slower reading*, *sharpness*, *losing position in text* and *rereading words*. Factor 2 consisted of *exhaustion*, *pain in the eyes*, *irritated eyes* and *strain in the eyes*. Figure 36 depicts the factor loading of each item on the two factors. A reliability test was performed that analyzed the internal consistency between the items belonging to a given factor, and revealed Cronbach's alphas of 0.91 and 0.81 for factor 1 and 2 (a Cronbach's alpha of > 0.70 is considered acceptable (Hair et al., 2006). The items *headache*, *concentration problems*, *sleepiness* and *remembering words* received low factor loading on both factor 1 and factor 2 and based on reliability analysis it was decided to exclude these items from further analysis.

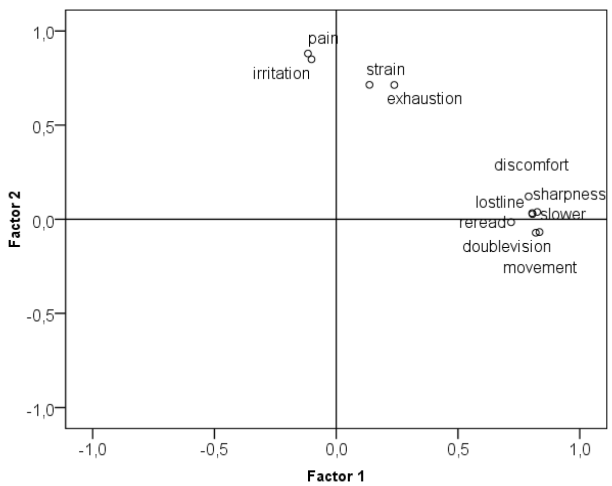


Figure 36. Factor loadings of all visual discomfort items on the two factors. The x-axis represents factor 1 and the y-axis represents factor 2.

For each participant a factor score was calculated for each factor in each condition; i.e., the score given by the participant on each questionnaire item was weighted with the factor loading of that item (i.e., with its relative importance on a given factor), and then summed over all items of the factor (Hair et al., 2006). An ANOVA with BIS, Dimension and Pre-Post as independent variables and the factor scores as dependent variables revealed that factor 1 was only affected significantly by a change in Dimension ($F(1, 188) = 46.266$, $p <$

.001) and factor 2 was affected significantly by BIS ($F(1, 188) = 11.172$, $p < .01$) and by Pre-Post ($F(1, 188) = 8.714$, $p < .01$). The results are depicted in Figure 37.

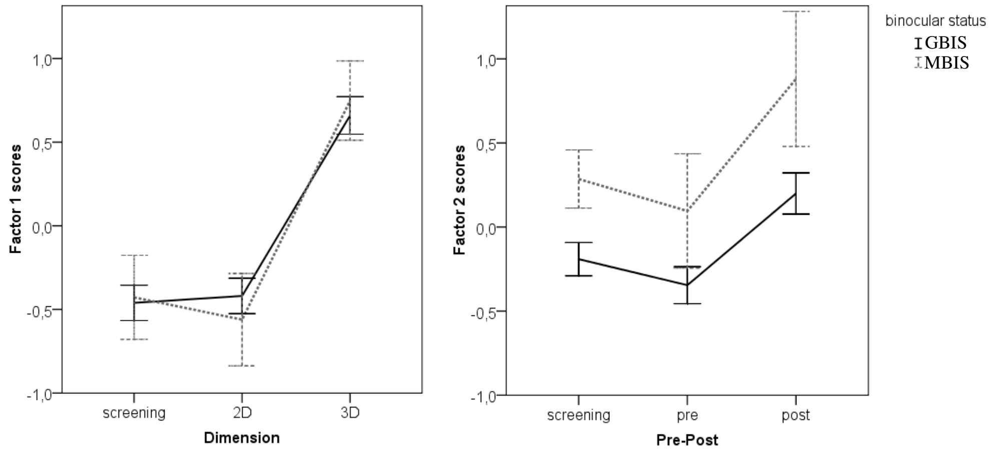
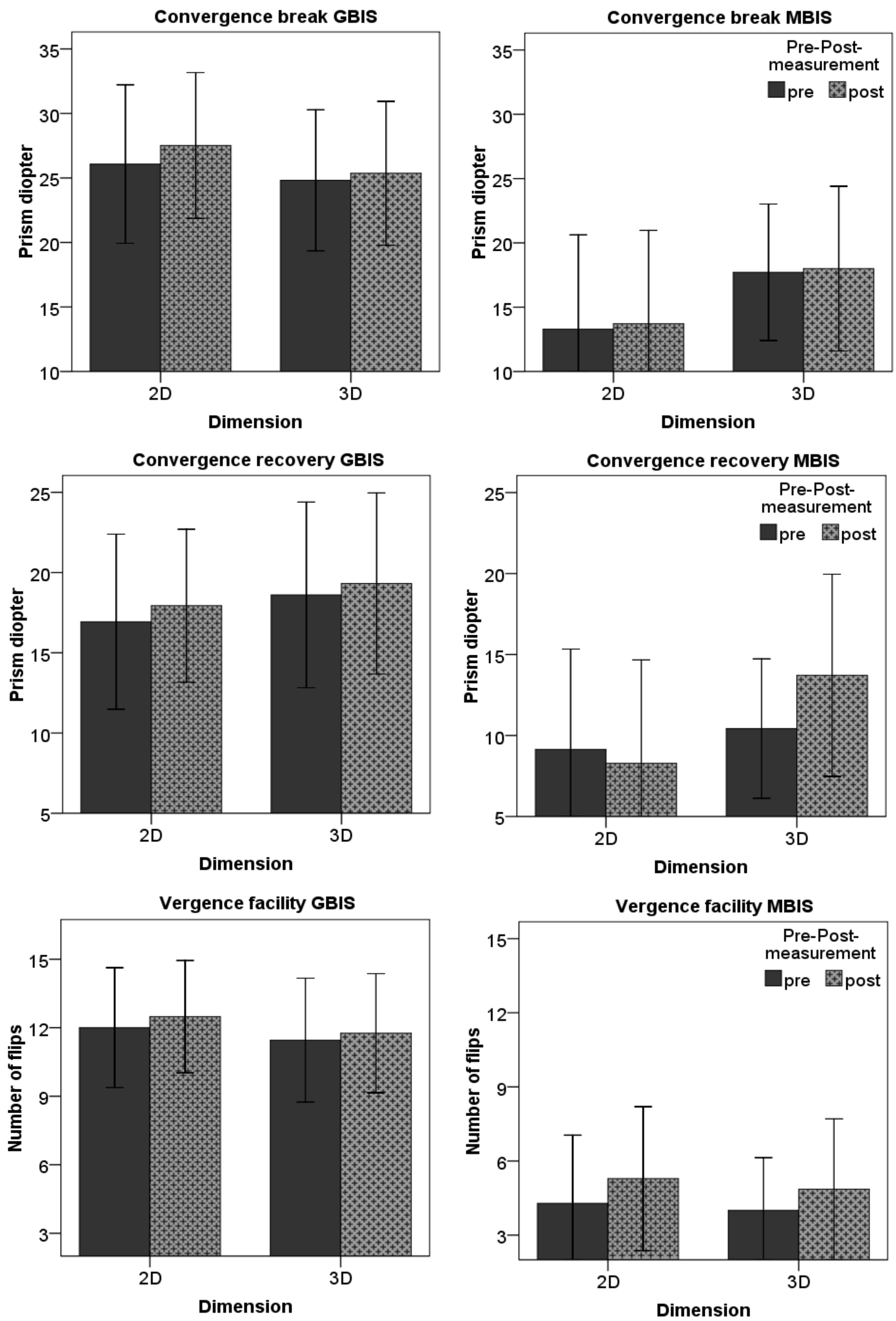


Figure 37. Mean visual comfort results in standard standardized scores with 95% confidence intervals of the two sets of questionnaire items that relate to two underlying factors as a function of the independent variables. The x-axes represent the variation in Dimension and Pre-Post for factor 1 and 2 respectively, the y-axes represent the standardized factor scores and the different lines represent the groups with different BIS.

Objective visual fatigue

When all participants were pooled, none of the indicators could differentiate significantly between the 2-D and 3-D conditions or between the pre- and post-measurements. Not surprisingly, some indicators did significantly and/or clinically distinguish between participants with a GBIS and MBIS, i.e., participants with a MBIS had lower convergent (p < .01) and divergent (p < 0.01) fusion values and vergence facility (p < 0.001). Other indicators were only suggestive of poorer visual functioning for participants who had MBIS (fixation disparity both with and without fusion lock and the heterophoria). Figure 38 depicts the clinically interesting results of fusion range and vergence facility when participants were categorized on their BIS. Participants with a GBIS showed no clinically meaningful changes in fusion range in both convergent and divergent directions between 2-D and 3-D and pre- and post-measurements. Participants with a MBIS showed clinically meaningful increases of four prism diopters between the 2-D and 3-D conditions of the convergence break measurement. The measurements of convergence recovery values for participants with MBIS seems to indicate a small pre-post-effect in the 3-D condition, i.e., the post-measurement after the 3-D condition is four prism diopters higher than 3-D pre-measurement and the 2-D pre- and post-measurements. The divergent break and recovery values of participants with a MBIS seems to react in the opposite direction, i.e., from 2-D to 3-D conditions the divergent break and recovery range reveal a small yet clear decrease. These differences were however not significant.


Convergence break GBIS
Prism diopter
35
30
25
20
15
10
2D
3D
Dimension
Convergence break MBIS
Pre-Post-
measurement
pre
post
Prism diopter
35
30
25
20
15
10
2D
3D
Dimension
Convergence recovery GBIS
Prism diopter
25
20
15
10
5
2D
3D
Dimension
Convergence recovery MBIS
Pre-Post-
measurement
pre
post
Prism diopter
25
20
15
10
5
2D
3D
Dimension
Vergence facility GBIS
Number of flips
15
12
9
6
3
2D
3D
Dimension
Vergence facility MBIS
Pre-Post-
measurement
pre
post
Number of flips
15
12
9
6
3
2D
3D
Dimension

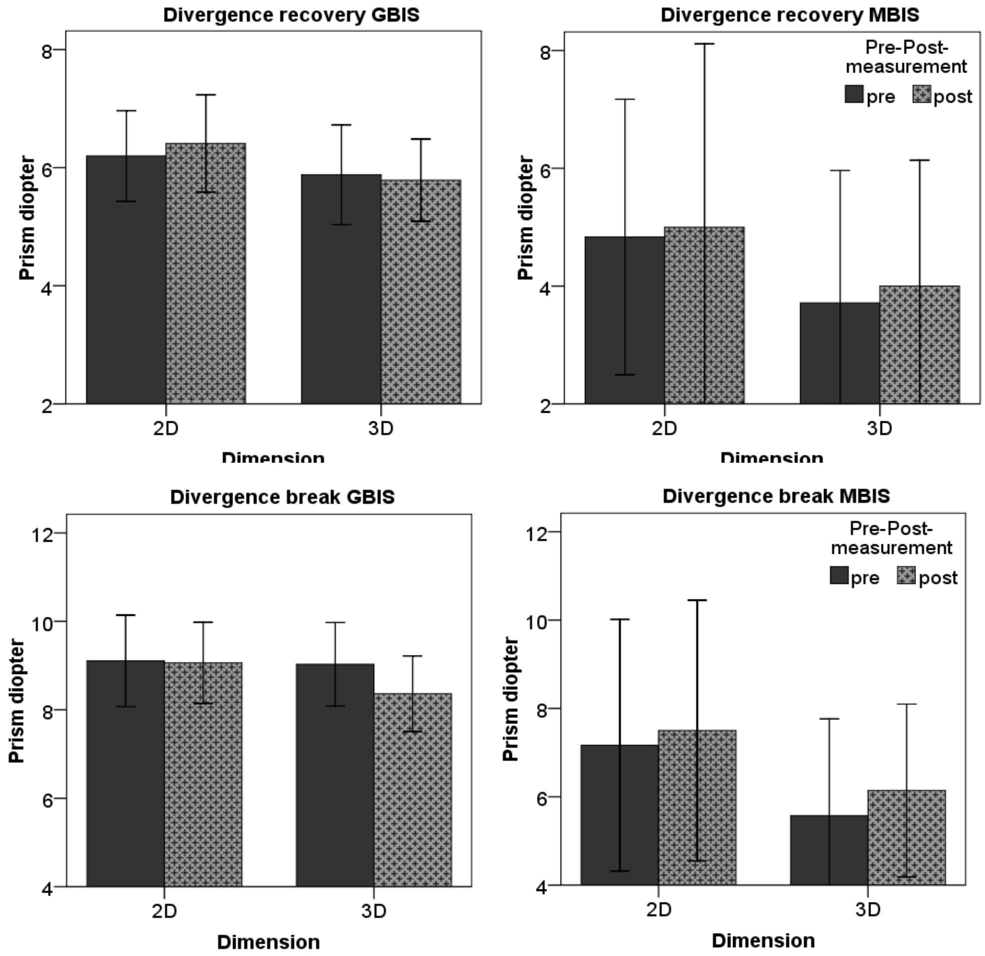


Figure 38. Mean values and their 95% confidence intervals of the convergent and divergent fusion ranges and vergence facility for each BIS group. The x-axes represent the variation in dimension, the y-axes represent the mean measurement values and the colours represent the pre- and post-measurements.

Word count

Participants performed the WRRT a total of ten times; two for the subjective pre- and post measurement and eight for the eight different objective indicators. Figure 39 depicts the results. The number of words participants could read in 60 seconds differed significantly between 2-D and 3-D ($F(1, 759) = 124.198$, $p < .001$), between participants with GBIS and MBIS ($F(1, 759) = 95.608$, $p < .001$) and the difference between 2-D and 3-D was larger for participants with MBIS ($F(1, 759) = 4.872$, $p < .05$). The chronological order of the WRRT (the stage in the experiment when it was carried out) did not have an effect on the number of words read ($F(9, 759) = 0.330$, $p = .965$), indicating consistent results.

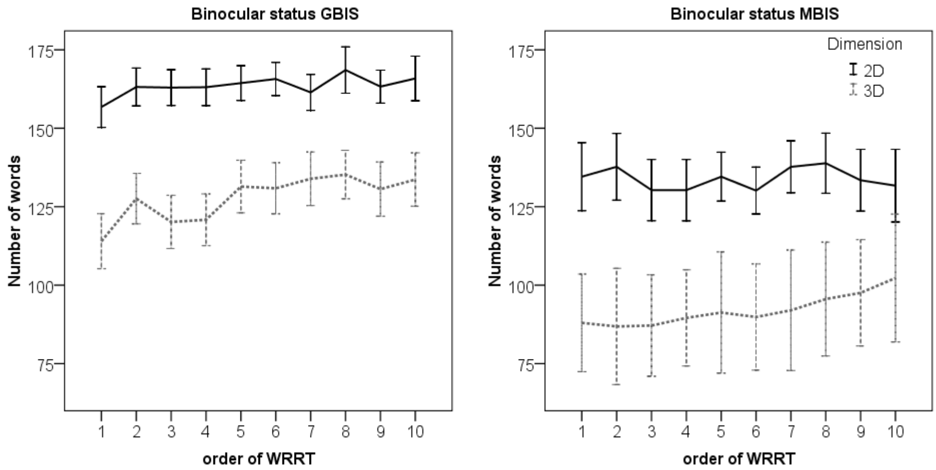


Figure 39. The mean and 95% confidence intervals of number of words read by participants with GBIS (left graph) and by participants with MBIS (right graph). The x-axes represent WRRT in chronological order, the y-axes represent the number of words and the different lines represent the dimension (2-D or 3-D).

Participants with MBIS read fewer words both in the 2-D and 3-D condition and their reading performance is affected more by the occurrence of screen disparity, i.e., the difference between 2-D and 3-D reading performance is larger for this subgroup. Hence, to analyze this impact of screen disparity, the ratio between the 2-D and 3-D reading performance is determined. Recall that the algorithm used to categorize participants based on their BIS computes a score between 0 and 13 of which a score of 5 or higher indicates MBIS. The categorization based on deviating values of more than one standard deviation of the norm, categorized participants with a MBIS when at least four of the six indicators deviated more than one standard deviation. Figure 40 depicts the ratio of the WRRT between 2-D and 3-D as a function of both categorizations methods as well as of the overall distinction in BIS.

5.1.6. Discussion

Some people report visual discomfort when watching 3-D displays, and previous research has revealed a lack of consensus in indicators to evaluate visual discomfort. We performed an experiment to identify methods for detecting objective signs of visual fatigue and subjective symptoms of visual discomfort. We hypothesized that 1) not all clinical tests are equally appropriate to evaluate the effect of stereoscopic viewing in terms of visual fatigue and visual discomfort, and 2) there is a natural variation in susceptibility to visual discomfort amongst people with normal vision. In this experiment 39 participants were categorized on their binocular status. The effect of a reading task in 2-D and 3-D on their binocular status was tested with eight objective optometric indicators and one questionnaire with 15 items.

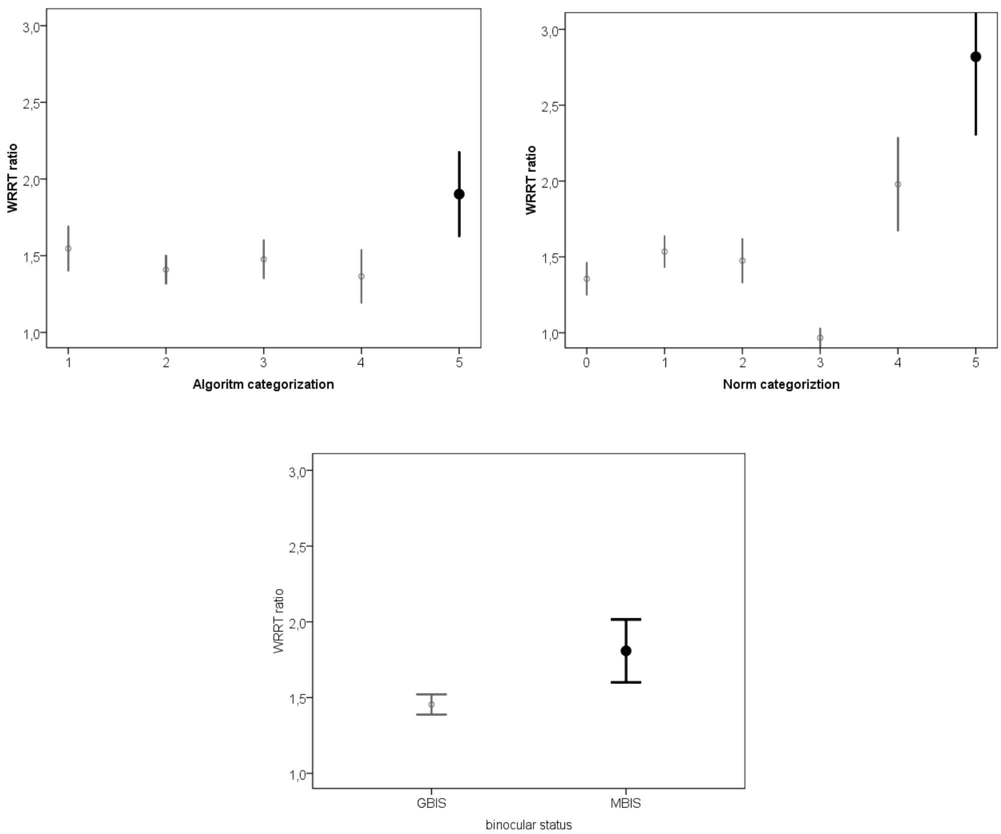


Figure 40. Categorization of participants based on optometric algorithm (upper left graph), norm values (upper right graph) and BIS (bottom graph) including 95% confidence intervals. The x-axes represent the algorithm score and the y-axes represent the ratio of number of words read between 2-D and 3-D sessions. The ratio values of the degree of categorization that labelled participants as MBIS are indicated in solid black.

Our results demonstrate that 1) a combination of fusion range measurements and self-report is appropriate for evaluating visual discomfort of stereoscopic stills, and 2) that people with a moderate binocular status are more susceptible to visual discomfort associated with stereoscopic displays. Additionally, we also describe a relatively simple measurement tool (the ratio of reading performance between 2-D and 3-D) that identifies these people who are more susceptible to visual discomfort associated with stereoscopic content.

With respect to the first hypothesis, most objective indicators did not reveal any change in visual functionality caused by 3-D stimuli. Under certain specific circumstances, however, some indicators did show clinically meaningful changes after the 3-D stimuli. Hence, these indicators are more sensitive to changes in the binocular visual system associated with stereoscopic displays. These specific circumstances underline the support of the second hypothesis, i.e., the fact that certain people respond differently to stereoscopic content as a result of differences in their binocular visual system. Only people with a relatively moderate binocular status revealed changes in some objective indicators. Given that the binocular visual system has some degree of plasticity and is able to adapt to altered viewing conditions, only objective changes that decrease the performance of the binocular visual system or that are accompanied by the experience of visual discomfort can be regarded as visual fatigue. Since this was the case, we regard these changes as visual fatigue. Hence, people's degree of susceptibility to visual discomfort associated with stereoscopic displays relates to their binocular visual system. In particular, a combination of convergence fusion range measurement and self-report is appropriate for measuring visual fatigue and visual discomfort associated with 3-D displays. Finally, the performance of the WRRT indicated that the ratio of the number of words read in 2-D and 3-D has potential as a simple measurement tool to identify people with moderate binocular status, i.e., who are susceptible to visual discomfort. By using the ratio as an indicator the impact of screen disparity is taken into account, whereas general reading disabilities are excluded. Participants with a moderate binocular status had a significantly higher ratio than participants with a good binocular status. This ratio has potential 1) as an indicator of moderate binocular status in perceptual research additionally to stereo- and visual acuity tests, and 2) in consumers' applications to set individual norms for comfortable screen disparities based on viewers' binocular status. As a potential supplementary tool the prism flipper to measure vergence facility can be used; it could be easily used in consumer applications. Although both hypotheses tested in this experiment are confirmed, some aspects need clarification.

Inclusion criteria

All participants who were included in the experiment passed the inclusion criteria. More specifically, people with large refractive deviations, amblyopia, stereo blindness and children and older people, were not included in the experiment. Hence, these results can only be generalised to people without gross binocular vision anomalies and can not extend to the entire population. From another perspective, one can also state that even though the subgroup contained only people without gross binocular vision anomalies, they could be differentiated based on their binocular status which determined their ability to perceive stereoscopic content comfortably.

Second display

Although not described herein, the experiment was also performed on a 42" Philips autostereoscopic nine-view lenticular LCD display (van Berkel and Clarke, 1997) (results are described in a conference contribution (Lambooij, Fortuin, IJsselsteijn and Heynderickx, 2009b) and in a paper concerning the impact of stereoscopic viewing on individuals (Fortuin et al., 2010)). Though similar effects were obtained, they were smaller, which could be attributed to two reasons; fewer participants performed the experiment on this display (n=19), and the amount of screen disparity of this display type is limited, i.e., a factor of four less than the screen disparity set on the two-view display.

Evans' algorithm

We implemented the algorithm of Evans (1997) to assess the binocular status that underlies decompensated heterophoria, since we were interested in the same subgroup, i.e., people with moderate binocular functioning. Especially in unnatural viewing situations, e.g., viewing stereoscopic content, this subgroup of people can experience visual discomfort that is associated with stereoscopic content. Efforts required of people with moderate binocular functioning to resolve excessive disparity, vergence, or a combination of both, in theory can lead to visual discomfort that is also associated with a decompensated heterophoria in susceptible individuals. Our results confirmed that participants with a moderate binocular status were more sensitive to visual fatigue and visual discomfort, and that the visual discomfort was similar to that experienced by participants with decompensated heterophoria. This might indicate that these participants have less plasticity in their binocular visual system to overcome unnatural viewing situations, which is also suggested by results of vergence facility. This is in line with previous research in which significant changes were found within the fusion range only for participants who were unable to free-fuse stereoscopic stimuli (Emoto et al., 2004).

The algorithm was modified by using fixation disparity without a fusion lock instead of aligning prisms with a fusion lock. The presence of fixation disparity and more specifically the magnitude of the prismatic correction that is required to eliminate the fixation disparity (aligning prism) during near vision is a good predictor of symptoms that are attributable to a decompensated heterophoria (Jenkins et al., 1989; Pickwell et al., 1991; Karania and Evans, 2006). When fixation disparity is assessed using instruments with a good foveal and peripheral fusion lock, the measurement reveals more accurate and smaller fixation disparity than without a fusion lock (Brautaset and Jennings, 2006; Evans, 2008; Ukwade, 2000). This explains the 20% of the fixation disparity measures that exceeded the normative cut-off value that was derived from Peli (1998) of 0.1 PD. Since we know of no norms for fixation disparity without a fusion lock as measured with the TC2000 for far distances, we used one standard deviation (which was 1.2 PD) as a criterion. This relative high cut-off value minimizes overestimation of participants with a MBIS based on the different fixation disparity measurement used in Evans' algorithm.

Fusion range

The fact that the fusion range was the only objective indicator that was sensitive enough to reveal visual fatigue associated with stereoscopic viewing, can be attributed to the type of stimuli. The stimuli that were used in the 3-D session were short-term stereoscopic stimuli with excessive screen disparity. The rationale behind the application of such a stressful short-term stimulus was to induce some measurable level of visual fatigue and visual discomfort. Hence, the combination of fusion range and the specific questionnaire items can be appropriate only for evaluation of stereoscopic stills with a specific stressor. For example, if the words in the reading tasks had in turn crossed and uncrossed disparity, also vergence facility might indicate differences between 2-D and 3-D conditions, or if the stimuli had 3-D artifacts, other questionnaire items might be more appropriate.

In addition, the effects in objective indicators are barely clinically relevant. Longer or more stressful stimuli could be used that have more profound impacts on the visual system, though this raises ethical issues since long-term visual discomfort and headaches might be induced. The convergent fusion range of participants with moderate binocular status increased and the divergent fusion slightly decreased in 3-D sessions. In general, the difference between their pre-fusion values in 2-D and 3-D was larger than the difference between their pre- and post fusion values in the 3-D session. This can indicate an accumulative change in fusion as a consequence of the five preceding 3-D stimuli, used for the five preceding optical measurements. More specifically, this tendency of the fusion range of participants with a moderate binocular status becomes interesting since it appeared to shift towards the direction of the stimulus. The increase in convergence fusion range and decrease of divergence fusion correspond to a crossed screen disparity, i.e., an object rendered in front of the display plane. It is important to note that this change in fusion only occurred during the 3-D session for participants with a moderate binocular status.

Objective indicators

It should be noted that the relationship between objective indicators for visual fatigue and subjective indicators for visual discomfort is still difficult to establish. We stated in the introduction that only changes that decrease the performance of the binocular visual system or that are accompanied by the experience of visual discomfort can be referred to as visual fatigue. The performance of the WRRT could be related to experimental conditions, and participants seemed very capable in indicating specific visual discomfort items that could be related to experimental conditions as well. The detection of visual fatigue with pre- and post-measurement of optometric indicators, however, was complicated since changes within the binocular visual system as a result of stressful short-term stereoscopic still images have a rapid deterioration. This rapid deterioration can have been the reason why Peli (1998) did not find any clinically meaningful visual fatigue, because he performed all his tests as a set before and after a stimulus. Long-term stressful stimuli might impose a larger impact on the visual functioning, though as stated before, this raises ethical issues. The objective changes in the visual functioning of participants with a moderate binocular status can indicate harmless functional adaptations to an altered environment instead of visual fatigue, though the opposite is more plausible. The categorization based on binocular status by itself means that the binocular visual systems of participants with moderate binocular status has a low degree of plasticity, whereas the binocular visual systems of participants with good binocular status have larger buffers or more adaptability in their functioning. The changes in visual functions during the experiment were only measured within the group of participants with moderate binocular status and were accompanied by the occurrence of more subjective visual discomfort in terms of pain, exhaustion, strain and irritation and by a poorer performance in the WRRT. As such, it is referred to as visual fatigue instead of as a functional adaptation to a change in the environment or an improvement of the visual performance. The fact that no visual fatigue was measured for participants with good binocular status indicates that no objective changes were present, that these objective changes disappeared more rapidly once the 3-D stimulus was gone or that these changes were smaller. Another possibility is that the objective indicators lack specificity (ability to correctly identify people without a specific condition) and sensitivity (ability to correctly identify people with a specific condition), yet since effects were measured in the binocular system of participants with moderate binocular status, this possibility seems less likely. It is unclear which of these hypotheses is correct: direct measurements whilst performing the WRRT would provide more detailed information, though such measurement methods are very expensive and not within our reach.

WRRT

The WRRT when used as described in this research has potential as a useful measurement tool to differentiate people based on their binocular status. Such a tool can be easily implemented in 3-D consumer applications as advice or warning to decrease screen disparity settings to reduce visual discomfort while watching 3-D television. Such a warning could be part of future international standardization of guidelines on image safety (van Nes, 2009). Future research, however, is required to 1) validate this new use of the WRRT, and 2) validate the WRRT as measurement tool for close viewing distance since accommodation can start to have an impact as well.

5.1.7. Conclusion

We made a first attempt to find the most appropriate objective, subjective and performance indicators to evaluate visual discomfort associated with 3-D-TV. An initial optometric screening allowed us to identify people with a moderate binocular status. For this group of people, the fusion range is the only objective indicator that revealed a clinically meaningful change within the binocular visual system due to stressful short-term stereoscopic viewing. Specific questionnaire items indicate more visual discomfort for this subgroup, whereas other items indicate more visual discomfort in 3-D than in 2-D. Hence, we suggest that a combination of fusional range measurement and self-report is appropriate for evaluating visual discomfort related to 3-D-TV. We also constructed a simple measurement tool, i.e., the ratio of the number of words read in the 2-D and 3-D viewing conditions, to distinguish participants based on their binocular status for large viewing distances. Though the value of this tool needs to be confirmed in future research, it can be useful 1) to serve as a binocular ability test in perceptual research additionally to stereo- and visual acuity tests and 2) in consumers' applications to set individual norms for comfortable screen disparities based on viewers' binocular status.

5.2. Experiment 7: Performance Measures of Visual Discomfort

ABSTRACT

We recently revealed that people with some signs of binocular dysfunctioning are susceptible to visual discomfort associated with viewing stereoscopic content at large viewing distances. Two performance measurements enabled to distinguish people by their binocular status (BIS): the ratio of performance of the Wilkins Rate of Reading Test (WRRT) between 2-D and 3-D, and the vergence facility. In this paper, both measures are hypothesized to be also appropriate to identify people by their BIS for near-view stereoscopic content.

An experiment was designed in two parts. Firstly, an extensive optometric screening was carried out to differentiate visually asymptomatic young adults with good BIS (GBIS) (N=27) from those with a moderate BIS (MBIS) (N=6). Secondly, participants had to perform the WRRT at close viewing distance under five screen disparity settings (-1.5, -0.75, 0, 0.75 and 1.5 degrees), and each WRRT was immediately followed by a questionnaire.

The results reveal that the ratio of the WRRT between 0 and -1.5 screen disparity is an appropriate indicator of participants with MBIS in comparison with participants with GBIS. Subjective asthenopic items and objective vergence facility reinforce this distinction in subgroups. In addition, the results show that 0.75 degrees of screen disparity is already problematic for people with MBIS. Based on these results we conclude that the WRRT ratio has potential as a BIS test in perceptual research in addition to stereo- and visual acuity tests, or in consumer applications to provide individual settings for comfortable screen disparities based on viewers' BIS.

5.2.1. Introduction

Since a relatively large viewing distance (three meters) was used in Experiment 6, the impact of accommodative dysfunctions was minimized. In the current experiment we focus on stereoscopic near work, which complicates the analysis of the relationship between people's binocular status (BIS) and the occurrence of visual discomfort as a consequence of screen disparity; accommodative dysfunctions might have a perceptual impact as well. Accommodative anomalies (e.g., accommodative insufficiency, accommodative infacility, and accommodative spasm) affect approximately 16% of patients consulting community optometrists (Ciuffreda, 2002; Scheiman, 1996). The consequent visual discomfort is diverse and non-specific, and can become present, or more severe, in unnatural viewing situations, such as when performing near work. The underlying determinants that can cause difficulties in near vision stereoscopic displays can be established via an optometric screening. For close viewing distances it can be possible, as with binocular dysfunctions for large viewing distances to find an analogue (e.g., rate of reading) that will indicate clinically relevant accommodative anomalies.

The reported research here aims to 1) provide further data on the proportion of people with a moderate BIS who are more susceptible to visual discomfort associated with modern stereoscopic displays, 2) confirm whether those with this susceptibility can be detected by the ratio in reading performance between 2-D and 3-D, 3) confirm that vergence facility is useful as a supplementary indicator to distinguish levels of susceptibility, and 4) analyze the effect of accommodative dysfunctions on stereoscopic near work. If participants who are prone to visual discomfort as a consequence of a moderate BIS associated with stereoscopic viewing can be detected by a simple test of performance in a reading task then this has two important implications. First, such a test could serve as a BIS test in perceptual research, perhaps complementing the information provided by stereo acuity and visual acuity tests. Second, this test could be a useful tool in consumers' applications to individually set comfortable screen disparities based on viewers' BIS.

5.2.2. Background

Binocular status

To obtain clear, binocular single vision, the visual system maintains accommodation and convergence that are appropriate for the viewing distance. *Vergence* can be defined as the movement of the eyes in opposite directions to maintain fixation on an object that is moving in depth. The vergence system is primarily retinal disparity-driven (Hung, 2001; Ciuffreda, 2002) and resolves mismatches in the fusion system. In most cases, the vergence system compensates for binocular anomalies such as heterophoria (the tendency of the eyes to misalign when they are dissociated) or fixation disparity (the minimum angular measure of misalignment of the visual axes of the eyes). People with inadequate fusion or excessive heterophoria can require more effort to maintain fusion. This can lead to visual discomfort when viewing stereoscopic content (Blehm et al., 2005; Evans, 2007; Sheedy et al., 2003; Steinman et al., 2000). In this case the heterophoria is described as decompensated and can be associated with clinical signs such as fixation disparity (Jenkins, Pickwell, Yekta, 1989; Karania and Evans, 2006; for a review see Evans, 2007)

Accommodation concerns the alteration of the crystalline lens to obtain and maintain the object of interest with the highest attainable resolution focused on the fovea. The accommodation system is primarily driven by retinal blur and can compensate for small errors in specific accommodative dysfunctions (Ciuffreda, 2002; Lambooij et al., 2009; Hung, 2001; Peli, 1999). In unnatural or stressful viewing situations, such as stereoscopic near work, people with smaller tolerance in their accommodative functioning can expend so much effort that visual discomfort can arise (Blehm et al., 2005; Dillon and Emurian, 1996; Iribarren et al., 2001; Owens and Wolf-Kelly, 1987; Scheiman, 1996; Sheedy et al., 2003; Sheedy and Parsons, 1990).

The literature described above generally supports the need for clinically testing people before they participate in stereoscopic perception experiments, and ideally before they view stereoscopic full-feature movies or participate in other simulated 3-D viewing environments. This is because simulated 3-D content creates artificial viewing conditions, most notably by creating a different stimulus for vergence and accommodation (Hoffman et al., 2008).

For diagnosing binocular dysfunctions no single test is 100% effective or objective (Evans, 2008). In view of this, Evans (1997; 2007) constructed an algorithm (described in section 3.2 *Screening of participants*) for diagnosing decompensated heterophoria that computes a score representative for a person's binocular status (BIS). In Experiment 6 this algorithm was applied to discriminate amongst people with normal binocular functioning (i.e., without any associated visual discomfort of non-strabismic binocular anomalies in normal viewing situations) those with a moderate BIS, and consequently, with a higher susceptibility to visual discomfort associated with viewing stereoscopic displays (see section 6.4). Seven of the 39 participants (18%) were identified as people with a moderate BIS. After short-term stressful stereoscopic reading tasks at a viewing distance of three meters these seven participants revealed objective signs of visual fatigue and indicated more visual discomfort in contrast to those participants with a good BIS.

In case of stereoscopic near work, participants should be screened for both binocular and accommodative functioning. Dysfunctions in either systems can cause similar, non-specific, visual discomfort including blurred, double or distorted vision, difficulty in changing focus, headache, aching eyes, sore eyes and irritated eyes (Sheedy and Parsons, 1990; Evans, 2008; Ciuffreda, 2002; Scheiman, 1996). Hence, it is not that straightforward to determine whether a symptom associated with stereoscopic near work is attributable to a binocular vision anomaly, an accommodative dysfunction, to both, or maybe even another condition, e.g., dry eyes.

Reading performance

In modern societies, in which people often want to read as accurate and fast as possible, a common symptom in people with visual dysfunctions is a poorer reading performance (O'Leary and Evans, 2006). It is widely reported that visual deficits (in e.g., visual acuity, fixation disparity, fusion, accommodation, vergence, visual stress) can impair reading ability and induce visual discomfort when reading (Buzzelli 1991; Jeanes et al., 1997; O'Leary and Evans, 2006; Legge, 2007; Wilkins et al., 2004). For people with accommodative dysfunctions text will be out of focus, which can impair reading performance since blur impedes resolving detail. For people with binocular dysfunctions, text can appear doubled, blurred or distorted due to vergence errors while making saccades or making return sweeps at the end of the line (O'Leary and Evans, 2006; Evans, 2007).

The extent of these reading impairments can be properly assessed with the Wilkins Rate of Reading Test (WRRT; Wilkins et al., 1996; Wilkins, 2002; Kriss and Evans, 2005; O'Leary and Evans, 2006). The WRRT will be discussed in more detail in paragraph 5.2.4.

Experimental set-up. It was originally developed to assess symptoms that are alleviated by using coloured filters of a specific tint while reading (Wilkins et al., 1996). Jeanes et al. (1997) reported that children who benefit from using coloured filters, showed an improvement in performance of 8% with the WRRT. O'Leary and Evans (2006) found that people who required optometric corrections for decompensated heterophoria, improved by at least 5% with the WRRT when using that correction. Experiment 6 revealed that people with a moderate BIS classified by Evans' algorithm (Evans, 1997) showed a significantly higher ratio in performance of the WRRT between 2-D and 3-D (number of words read in 3-D relative to number of words read in 2-D and referred to as WRRT-ratio), than people with good BIS. Participants with a moderate BIS had on average a WRRT-ratio of 1.8, whereas participants with a good BIS had on average a WRRT-ratio of 1.5. In other words, the adverse effect of screen disparity on reading performance was larger for people with a moderate BIS than with a good BIS.

Facility measurement

An indication of a normal visual system is proper adaptation and sufficient facility to respond to alterations in the viewing environment (Steinman et al., 2000). Vergence facility is the ability of the fusional vergence system to respond efficiently and accurately to changing demands over time (Gall et al., 1998). Vergence facility testing improves the diagnosis of binocular dysfunctions (Gall et al., 1998) and is an indicator of readiness for binocular visual tasks (Scheiman, 1996; Melville and Firth, 2002). The results of Experiment 6 revealed that participants with a moderate BIS, who experienced more visual discomfort associated with stereoscopic reading tasks, had a lower vergence facility than participants with a good BIS. Analogously, accommodation facility is the ability of the accommodative visual system to respond efficiently and accurately to changing demands over time (Garcia et al., 2000). It is used as an indicator for readiness of near vision tasks (Scheiman, 1996; Garcia et al., 2000). Testing accommodation facility under binocular test conditions provides 1) a direct evaluation of the dynamics of accommodative responses similar to under monocular test conditions, and 2) information about the coupling between accommodation and vergence (Garcia et al., 2000; Evans, 2007; Gall et al., 1998). As such it is also referred to as interactive facility and seems appropriate for stereoscopic near work.

5.2.3. The current study

The current experiment aims to construct relatively simple performance measurements that can be used to identify a subgroup of participants who might be more susceptible to visual discomfort associated with stereoscopic displays at shorter viewing distance. Participants with normal visual functioning (i.e., without any associated visual discomfort of non-strabismic binocular anomalies in normal viewing situations) are categorized via an extensive optometric screening into two subgroups: one with a good BIS and one with a moderate BIS (see section 6.4). It is noteworthy that the original aim was to construct a subgroup of participants with accommodative dysfunctions as well, yet all participants appeared to have proper accommodative status. Three relatively simple indicators are evaluated in terms of their ability to confirm this distinction in subgroups: the WRRT-ratio, the vergence facility and the accommodation facility. The former two showed potential to categorize people based on their BIS, which determined their susceptibility to visual discomfort associated with viewing stereoscopic displays for a large viewing distance in Experiment 6. The latter is incorporated since accommodative dysfunctions can also impact the ability to comfortably view (stereoscopic) content for close viewing distances. The facility measurements are two tests that are not included in Evans' algorithm and are thought to be most relevant as potential predictors of problems when viewing 3-D displays.

The first hypothesis states that the WRRT-ratio is effective to distinguish people that are more susceptible to visual discomfort associated with stereoscopic displays for close viewing distances. The second hypothesis states that vergence facility measured at large viewing distances distinguishes people by their susceptibility to visual discomfort associated with stereoscopic near work. The third hypothesis states that for near vision, accommodation facility is an appropriate indicator to differentiate people's susceptibility to visual discomfort associated with stereoscopic displays.

5.2.4. Experimental set-up

Design

The experimental design consisted of two parts: (1) an optometric screening and (2) the performance of the WRRT for close viewing distance. The optometric screening included accommodation and vergence facility measurements and facilitated categorization of participants based on their BIS: good BIS (GBIS, N=27) and moderate BIS (MBIS, N=6). Neither the participants, not the experimenter did know to which subgroup a participant belonged during the experiment.

In the second part the WRRT was performed via a Wheatstone viewer under five screen disparity settings (Disparity): 1.5, 0.75, 0, -0.75 and -1.5 degrees, where a '-' sign indicates crossed screen disparity. Different WRRTs were assigned randomly to the five screen disparity settings of which the order was randomised across participants. After each WRRT, the subjective visual discomfort was evaluated with a questionnaire containing 15 subjective items (Item). The entire experiment (screening part and experimental parts) was scheduled on one day.

Screening of participants

An extensive optometric screening was carried out on 33 naive visually asymptomatic participants of age range 18 to 36 years and with refractive errors no larger than 0.50 diopter (D). This screening was performed for two reasons: (1) to exclude participants with eye diseases or severe binocular abnormalities such as partial (stereo) blindness, strabismus and amblyopia, and (2) to differentiate participants in subgroups based on their visual functioning.

Subgroups were assembled based on the optometric screening carried out at near vision distances (40 to 70 cm). The MBIS subgroup was established based on an algorithm proposed by Evans (1997 and 2007). An explanation of the algorithm is provided in Experiment 6 and we refer to Evans (2007) for a more extensive description. Since in Experiment 6 the algorithm was slightly modified, an explanation is provided of the algorithm as it was implemented in the current experiment. The algorithm is outlined in Table 20. The algorithm evaluates the binocular status by computing a score that is the cumulative value of ten single scores that each relate to the result of a single optometric indicator of decompensated heterophoria or binocular instability. The total algorithm score ranges from 0 to 16; a score of 0 is normal, 1-3 indicates minimal signs of decompensated heterophoria but still classified as normal, 6-16 as decompensated heterophoria, and 4-5 as suspect (borderline) decompensated heterophoria (Evans, 1997 and 2007). For the present research, a score 6 or higher was classified as MBIS and 5 or lower as GBIS. When using this threshold, the algorithm classified six participants to the MBIS subgroup. To detect the symptoms that are suggestive of binocular dysfunctions (the first indicator in the algorithm), the Convergence Insufficiency Symptom Survey (CISS) was used that incorporates general asthenopic items (Borsting et al., 2003; Sheedy et al., 2003). Since the CISS questionnaire also incorporates items that relate to reading tasks specifically, it was also used in the actual experiment.

For objective indicators, a short explanation is given here only for those that differ from Experiment 6. In contrast to Experiment 6 instead of fixation disparity, aligning prism is assessed in the presence of a central fusional lock. This provides a more accurate indication of the objective eye position than without a central fusion lock (Brautaset and Jennings, 2006), and is a good predictor of symptomatic heterophoria during near vision (Yekta et al., 1987; Jenkins et al., 1986; Pickwell et al., 1991; Karania and Evans, 2006). Foveal suppression refers to a very small suppression area that occurs in the foveal region. Sheard's criterion states that the fusional range that opposes the heterophoria should at least twice the degree of the heterophoria (Sheedy and Saladin, 1977; 1978).

For the detection of accommodative dysfunctions suggestions have been made about which accommodation indicators should be included in an optometric screening (Scheiman, 1996), yet an objective protocol of diagnosis to clearly distinguish the subgroup with poor accommodative status is not available to our knowledge. Table 21 lists the accommodative indicators used in the screening. The accommodation amplitude refers to the shortest distance on which the eyes can focus. Relative accommodation refers to the maximum amount by which the accommodation can change for a given degree of vergence measured in diopters, which can be positive or negative. In other words, this is the amount by which accommodation can be increased or relaxed, in response to negative or positive lenses respectively, before blur arises while keeping vergence constant. The accommodation response refers to the focus control of the eye, (i.e., the accuracy of the accommodation in direction and size). Accommodation facility, which has been explained in detail, is the rate (in cycles per minute) at which the accommodation can be changed to maintain clear vision through alternating opposing lenses. None of the participants had deviating values from the norm for more than one of these indicators, which made it infeasible to create an accommodative subgroup. Table 21 lists the mean and standard deviation of all accommodative indicator values for the MBIS and GBIS subgroup and the normal values (Evans, 2007).

The last part of the screening concerned the measurement of the vergence facility. Vergence facility is the rate (in cycles per minute) at which vergence can be changed so that an object is properly fused through alternating base-in and base-out prisms. To prevent subjects requiring a high degree of accommodation, measurement of the vergence facility was performed at a viewing distance of 2.5 meter. The commonly used prism combination to test vergence facility, i.e. 12 base out / 3 base in, was used in this experiment (Gall et al., 1998).

Equipment

The work unit that was used for the optometric measurements during the screening was the same as used in Experiment 6. The Mallett Near Vision Unit (IOO Sales Ltd. London, UK) was used to facilitate the measurements of the visual functionality of the participants during screening for close viewing distances. The Mallett Near Vision Unit is shown in Figure 41.

TABLE 21

ALGORITHM PROPOSED BY EVANS (1997) APPLIED TO DIFFERENTIATE BETWEEN PARTICIPANTS WITH RESPECT TO THEIR BINOCULAR STATUS BASED ON OBJECTIVE AND SUBJECTIVE OPTOMETRIC INDICATORS (EVANS (1997; 2007)).

Nr	Indicators	Sign or symptom	Score
1	symptomatic heterophoria	One or more of the questionnaire symptoms (if so, +3 or +2 or +1 if borderline)	3
2	cover test	heterophoria detected	1
3	cover test	absence of rapid and smooth recovery (if so, +2 or +1 if borderline)	2
4	fixation disparity with fusion lock	aligning prism (Mallett): >1 Δ patients under 40 or >2 PD patients over 40	2
5	fixation disparity with fusion lock	aligning prism (Mallett): unstable	1
6	foveal suppression	foveal suppression (Mallet): >3 arc minutes,	2
7	Sheard's criterion	Failed	2
8	Percival's criterion	Failed	1
9	dissociated heterophoria	unstable (if range > 4 PD)	1
10	fusion amplitude	< 20 PD	1

TABLE 20

MEAN VALUES AND STANDARD DEVIATIONS OF ACCOMMODATION INDICATORS OF THE NORMAL FUNCTIONING (NF) AND VERGENCE DYSFUNCTION (VD) SUBGROUP IN DIOPTERS AS WELL AS NORMAL VALUES.

	GBIS	MBIS	Norm
Indicators	mean ± STD*	mean ± STD*	
accommodation amplitude (binocular push up test)	9.20 ± 1.67	9.44 ± 0.86	15 – (0.25 x age)
negative relative accommodation	2.44 ± 0.50	2.38 ± 0.85	2.00 ± 0.50
positive relative accommodation	2.50 ± 1.11	2.13 ± 0.95	2.37 ± 1.12
accommodation response (MEM-retinoscopy**)	0.48 ± 0.37	0.67 ± 0.26	0.35 ± 0.35
accommodation facility (± 2 D lenses)	14.07 ± 4.44	13.00 ± 5.14	8 ± 5***

* units are in diopters except accommodation facility that is expressed in number of cycles

** MEM abbreviates monocular estimation method

*** (Zellers et al., 1984)

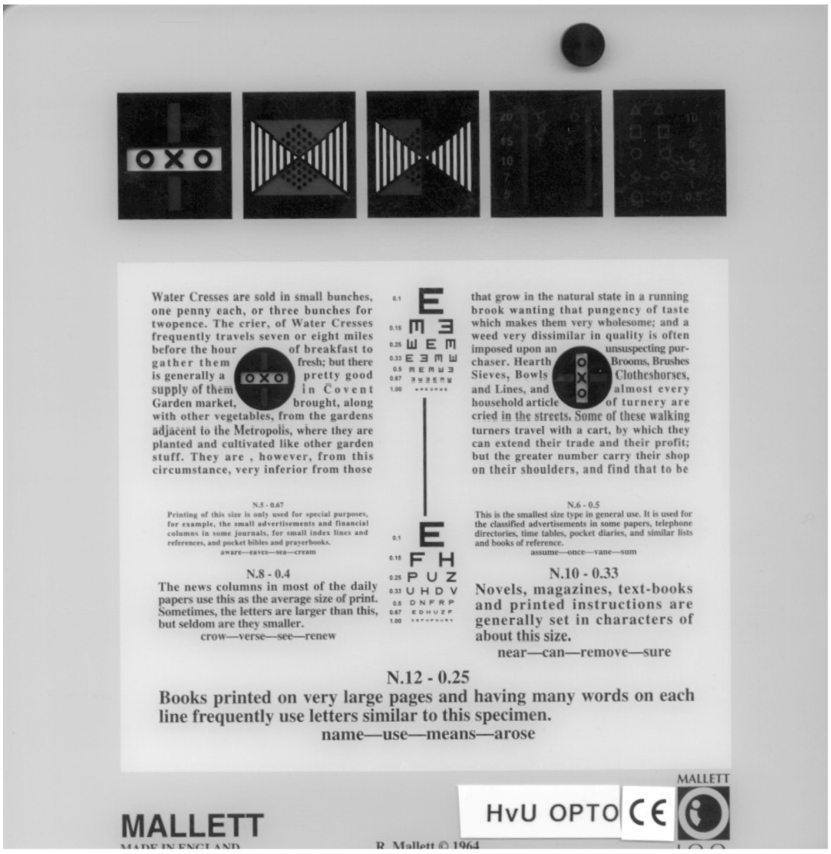


Figure 41. Photograph of the Mallett Near Vision Unit on which the polarized Nonius lines were aligned by using prism glasses while the subject viewed the small OXO test figure (the instrument is backward illuminated). Test instructions followed guidelines of Karania and Evans (2006).

To display the stimuli a Screenscope (mirror stereoscope) was used to direct the left- and right-eye image of a side-by-side displayed stereo pair to the appropriate eyes as depicted in Figure 10. For a more detailed description read Chapter 2.4.1. Experimental set-up.

Stimuli

The stimuli were five different passages of the Wilkins Rate of Reading test (WRRT; Wilkins et al., 1996) that were randomly assigned to conditions. For a complete description of the WRRT see Chapter 5.1.4. Experimental set-up. Since in that Chapter a modified version was used, Figure 42 depicts a screen shot of a stimulus as it was used in the current experiment. The visual angle of the text, inner and outer frame were 6.45, 10.08 and 11.53 degrees in horizontal direction and 3.17, 3.69 and 5.67 degrees in vertical direction respectively.

Questionnaire items

For subjective evaluation the CISS questionnaire was used and for a description read Chapter 5.1.4, Experimental set-up.

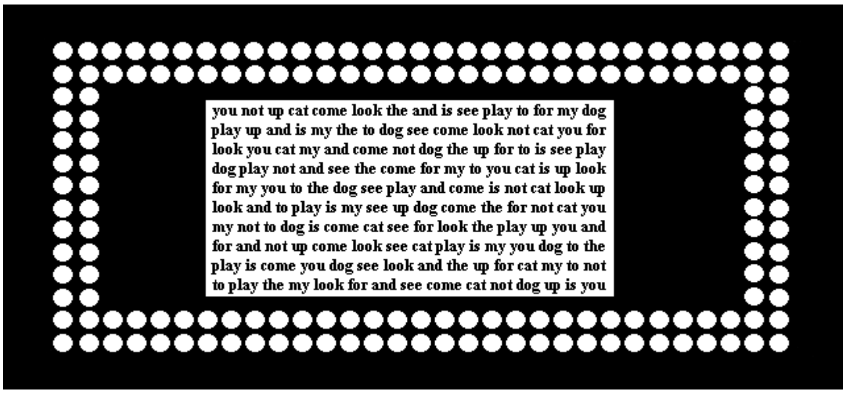


Figure 42. Screen shot of a stimulus; a passage of the Wilkins Rate of Reading test (Wilkins et al., 1996).

Procedure

Participants were provided with an informed consent statement containing information about the screening and the experiment, such as the general procedure and the possible occurrence of visual discomfort. After signing the statement, they proceeded with the optometric screening, which required about 25 minutes to complete. They received a brief instruction about the specific course of the experiment that incorporated a version of the WRRT to familiarize participants with the specific reading task. All questions concerning the procedure of the experiment were answered, after which the experiment started.

Participants had to properly rest their head against the Screenscope in order to maintain the appropriate viewing distance, and to obtain a clear and single image. The initial setting of the stimulus was always with zero screen disparity, yet participants were able to increase the screen disparity stepwise by pressing keypad number 8 (up arrow). Except in the 2-D condition, depending on the condition of the stimulus the WRRT changed in stereoscopic depth: either in crossed or uncrossed disparity. If participants had trouble fusing the stimuli, keypad number 2 (down arrow) could be used to decrease the screen disparity. This stepwise altering of the screen disparity to set any of the five screen disparities (-1.5, -.075, 0, 0.75 and 1.5 degrees) was included since Experiment 6 revealed that not all participants were able to fuse 1.5 degree of screen disparity in a single step. Hence, participants could increase the screen disparity to the value that belonged to that condition. Once the appropriate screen disparity was set, the participants were asked to read the words 'out loud' as rapidly as possible for 60 seconds. It was emphasized that the text needed to be in focus (sharp) and single. Each reading task was followed by the CISS questionnaire and a period of approximately two minutes to relax the eyes. The participants were instructed that each time the CISS questionnaire was carried out, their responses should relate to symptoms during the experimental condition that had just taken place. This entire procedure was performed five times (i.e., for five different disparity conditions) with the conditions randomised across participants. The experiment session required about 20 minutes.

5.2.5. Results

As explained in paragraph 5.2.4, Experimental set-up, two subgroups were identified based on their visual functioning; participants with a good binocular status (N= 27) and participants with a moderate binocular status (N=6); referred to as GBIS and MBIS, respectively.

Subjective results

Most of the CISS items yielded an averaged visual discomfort score around two, referring to moderate levels of visual discomfort. Not all items were affected equally by changes in Disparity or experienced similarly by the two subgroups of participants. A principal component analysis (PCA) was performed on the 15 CISS items to find intrinsic correlations and to reveal if certain items shared similar underlying attributes of visual discomfort (Hair et al., 2006). The PCA was combined with a non-orthogonal rotation method (Oblimin) to minimize the number of items with high factor loadings on more than one factor. The resulting PCA revealed two underlying factors that explained 39% and 19% of the variance of the data. Factor 1 received meaningful factor loadings (>.50 (Hair et al., 2006)) of the items *moving words, slower reading, sharpness, loss of position in text, rereading words* and *double vision* and relates to miss-perception of text. Factor 2 consisted of *exhaustion, headache, pain, strain* and *discomfort* and relates to asthenopia. Reliability testing revealed Cronbach's alphas of 0.84 and 0.75 for factor 1 and 2 respectively (a Cronbach's alpha of > 0.70 is considered acceptable (Hair et al., 2006)). The items *concentration problems, sleepiness, irritated eyes* and *remembering words* received a low factor loading on both factor 1 and factor 2. Based on the reliability analysis it was decided to exclude these items from further analysis. For each participant a factor score was calculated for each factor in each condition; i.e., the score given by the participant on each questionnaire item was weighted with the factor loading of that item, and then summed over all items of the factor (Hair et al., 2006). An ANOVA with BIS and Disparity as independent variables and the factor scores as dependent variables revealed that factor 1 was significantly affected by Disparity ($F(1, 155) = 2.553$, $p < .05$). Factor 2 was significantly affected by BIS ($F(1, 155) = 14.810$, $p < .001$), Disparity ($F(4, 155) = 4.747$, $p < .05$) and the interaction between BIS and Disparity ($F(4, 155) = 2.780$, $p < .05$). The results are depicted in Figure 43 for factor 1 (left graph) and factor 2 (right graph). Noteworthy to mention is the significant increase in factor 2 for the MBIS-group already at a disparity of 0.75 degrees.

Accommodation and vergence facility

Figure 44 depicts the average measurement results of the accommodation facility (left panel) and vergence facility (right panel) per BIS subgroup including the 95% confidence intervals and norm values.

A MANOVA revealed no statistically or clinically significant differences between the two different BIS subgroups in terms of their accommodation facility (p= .337) or vergence facility (p= .202).

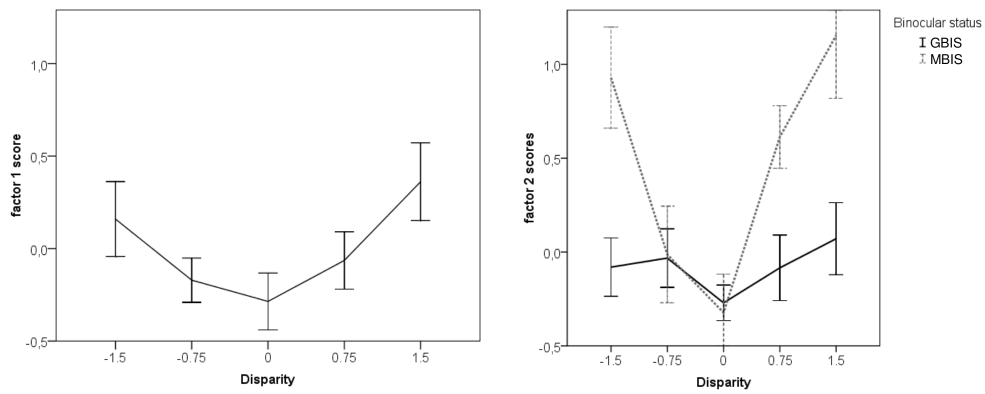


Figure 43. Mean visual comfort factor results of the two sets of questionnaire items. The x-axes represent the different screen disparities, the y-axes the standardized factor scores and the different lines in the right-hand panel the two subgroups.

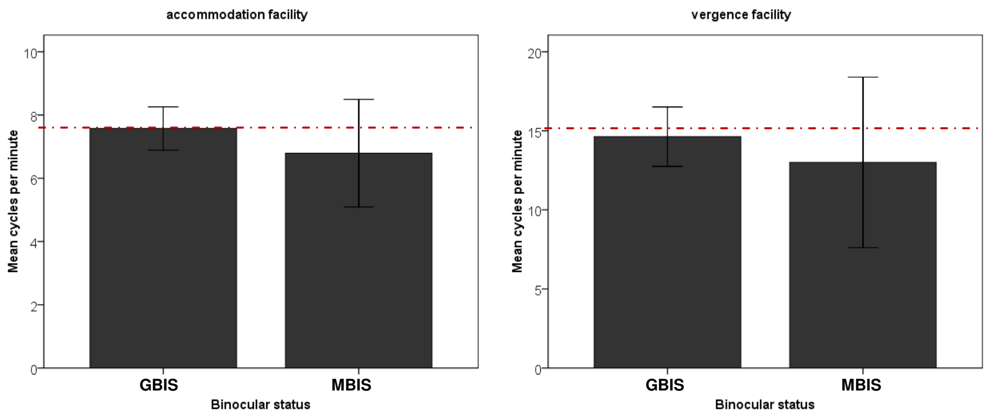


Figure 44. Accommodation facility scores (left panel) and vergence facility scores (right panel). The x-axes represent the BIS subgroups and the y-axes the number of cycles per minute. The dotted lines represent norm values; 7.7 cycles per 60 seconds for accommodation facility (Zellers et al., 1984) and 15 cycles per 60 seconds for vergence facility (Gall et al., 1998). It should be noted that the two norms were obtained in different ways.

WRRT performance

Figure 45A depicts the number of words read as a function of Disparity and BIS, whereas Figure 45B depicts the ratio of words read in 2-D compared to 3-D as a function of disparity and BIS.

Concerning the data illustrated in Figure 45A, an ANOVA revealed that the performance at the WRRT (i.e., number of words read) did not significantly differ between levels of Disparity ($F(4, 155) = 2.070$, $p = 0.087$) or between the BIS subgroups (($F(1, 155) = 1.211$, $p = 0.273$).

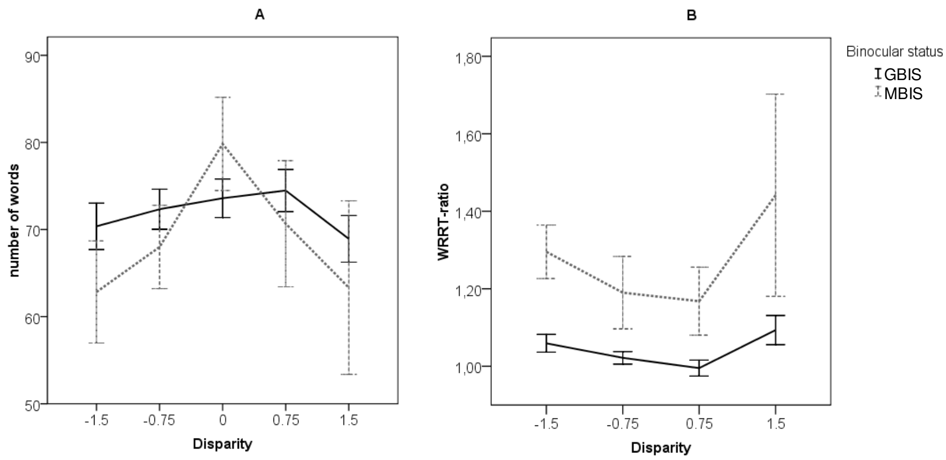


Figure 45. Panel A. Number of words read per minute as function of disparity and binocular status. Panel B. Ratio of number of words read in 2-D and 3-D as function of disparity and binocular status.

There is an effect visible, however, of Disparity on the performance of the WRRT for the MBIS subgroup, but this was not significant. It is plausible that this is due to the small number of participants in this group (N=6). The analysis of the ratio of the number of words read in 2-D and 3-D, i.e., WRRT-ratio in Figure 45B, revealed that BIS ($F(1, 124) = 23.805$, $p < .001$) and Disparity ($F(3, 124) = 3.751$, $p < .05$) had a significant effects on this ratio.

A reverse analysis should reveal if this WRRT-ratio can be used to predict the BIS, i.e., distinguish the MBIS subgroup from the GBIS group. For the reverse analysis only the WRRT-ratio of the 1.5 crossed screen disparity was used, i.e., -1.5 degrees. This screen disparity condition appeared most appropriate, since it revealed a large difference in WRRT-ratio between both subgroups with a small variance. The condition of the 1.5 uncrossed screen disparity caused problems for most of the participants, which could be attributed to the fact that it is easier for the human visual system to converge than to diverge. Hence, the WRRT-ratio in the uncrossed 1.5 screen disparity setting is not so appropriate for the reverse analysis. Since the implementation of a tool should be as simple as possible, a cut-off value of 1.25 in the WRRT-ratio was chosen which is the mean plus one STD of all participants, i.e., 1.10 + 0.15. Table 22 depicts the distinction accomplished by the WRRT-ratio when applying this limit.

TABLE 22
RESULTS OF REVERSE ANALYSIS OF THE WRRT-RATIO

	WRRT-ratio pass	WRRT-ratio fail
GBIS	25	2**
MBIS	1*	5

* false positive
** false negative

The discriminating performance of the WRRT-ratio criterion can be reflected by the sensitivity index d' (Rotello, Masson and Verde, 2008). The sensitivity index is defined as the standardized distance between the means of both subgroups and is calculated by the difference between the normal deviates of the *hit rates* and *false positives*. Calculation of d' entails the assumption that both subgroups have stochastic responses that are normally distributed and that all quadrants of the contingency matrix have equal impacts. The criterion of the WRRT-ratio results in a 92.6% *hit rate* of the GBIS group and 16.7% *false positives* of the MBIS group. The resulting normal deviates are 1.446 below the mean and 0.967 above the mean for respectively the GBIS group and the MBIS group, resulting in a sensitivity index d' (or discriminating performance) of 2.414.

To what extent this sensitivity index is associated with the maximal discriminative power, can be demonstrated by calculating the expected value of the WRRT-ratio criterion. The expected value reflects the performance of a test by positively incorporating the *hit rates* and the *correct rejections* and negatively incorporating the *false positives* and the *false negatives*. By varying the *false positives* and *false negatives*, the optimal expected value can be determined and thus the highest discriminative power. At the WRRT-ratio criterion found the sensitivity index d' of 2.4 had an expected value of 0.82, which was 1.2% less than the optimal expected value of 0.83.

5.2.6. Discussion

Viewing stereoscopic displays can cause visual discomfort (asthenopia) for a relatively limited number of people. We performed an experiment to identify easy applicable indicators of visual discomfort in relation to stereoscopic 3-D displays. Participants were asked to perform the Wilkins Rate of Reading Test (WRRT) at various levels of screen disparity. Based on Evans' (2007) criterion participants were differentiated into two subgroups: those with a good binocular status (GBIS) and those with a moderate BIS (MBIS). Our results reveal that participants with a MBIS experience more asthenopic complaints (already at a screen disparity of 0.75 degrees) and have a poorer reading performance at higher screen disparities than participants with GBIS. This indicates that the ratio of number of words read between 2-D and 3-D crossed screen disparity (WRRT-ratio) is able to categorize people based on their binocular status.

These results are in line with those of Experiment 6 in which the WRRT-ratio also significantly differed between subgroups with a GBIS and a MBIS. The cut-off ratio to distinguish between subgroups in the current experiment, however, is lower (1.25 compared to an average WRRT-ratio of 1.5 in Experiment 6). This can be associated with the difference in viewing distance. A close viewing distance gives participants a greater opportunity to adjust accommodation, which can account for the lower ratio for all participants. Even so, the ratio remains a useful predictor for subjects with binocular vision anomalies.

In research concerning reading difficulties, causal relationships with visual problems are not fully understood. Poor reading performance can be easily correlated with visual problems, yet it is more difficult to claim causality (Evans et al., 1996); people without visual problems can have reading difficulties as well, and people with visual problems can have no reading difficulties at all (Beech and Singleton, 1997). It is not the intention of the experiment to establish causality between reading performance and people's binocular status. Reading speed, however, depends strongly on image quality (Aberson and Bouwhuis, 1997). Since the virtual reading distance changes in accordance with the virtual stimulus size, the quality remained the same for all conditions. The WRRT-ratio only reflects the impact of screen disparity. For good stereoscopic depth perception, precise coordinated alignment of the two eyes is required (Cole and Boisvert, 1974; Ukwade et al., 2003). Hence, people with deficits in their binocular system (inadequate fusion, a decompensated heterophoria or a large fixation disparity) have more problems with saccades or with return sweeps to the beginning of the next line (Evans, 2007).

With respect to the second hypothesis, the vergence facility is poorer in the MBIS subgroup than in the GBIS subgroup. The average performance in both groups, however, does not reach the norm of 15 cycles per minute (Gall et al., 1998), which is in line with the results of Experiment 6. Since previous research also reveals that the vergence facility for large viewing distances lacks good specificity (the ability to correctly identify people without a specific visual dysfunctioning) (Fortuin et al., 2010), and has poor repeatability (Gall et al., 1998) the tool appears inadequate to differentiate subgroups by itself.

With respect to the third hypothesis, differences in accommodation facility could not enable identification of participants with a MBIS. All participants had similar accommodative functions; meaning none of the accommodative functions were more than one standard deviation below the mean for normal subjects (as reported by Zellers and colleagues in Zellers et al., 1984).

What can appear counter-intuitive is that on one hand we claim that there is no single method to identify people with binocular dysfunctions, while on the other hand we claim a single performance tool to categorize people according to their BIS. Though Evans' algorithm makes a reliable and valid distinction in subgroups based on binocular status, it is a multifactorial test that is complex, time-consuming and requires equipment that is not available in most research facilities or in any home environments. The WRRT-ratio is a much simpler measurement. The criterion of an abnormal WRRT-ratio had a 93% specificity (ability to correctly identify people without a MBIS) and an 83% sensitivity (ability to correctly identify people with a MBIS), and a high discriminating power; a sensitivity index of 2.4. There is little reason to change this criterion in order to obtain a higher sensitivity index, since the expected value can be maximally increased by 1.2%. Since it must be acknowledged that our research has thus far included fairly modest sample sizes (only 6 subjects in the MBIS group) further research with more participants is necessary to determine the stability of this sensitivity. In addition, since only young and healthy people were included, people with large refractive errors, strabismus and amblyopia, children, older people or people with visual impairment, should also be included.

Hence, the WRRT-ratio seems appropriate for detecting people who are susceptible to visual discomfort associated with stereoscopic displays. Indeed, we believe that this WRRT-ratio can have potential for future international standardization of guidelines on image safety (van Nes, 2009). For example, participants in 3-D perceptual research most often are not visually screened. This WRRT-ratio-test is indicative of BIS and can be added to stereo- and visual acuity tests that are the tests that are often used for screening. Future research should reveal if participants with a MBIS who are identified via the WRRT-ratio indeed have lower comfortable screen disparity limits than people with normal binocular status. If so, the WRRT-ratio-test can be used in consumer applications to set individual levels for comfortable screen disparities based on viewers' binocular status.

5.2.7. Conclusion

The home consumer market is progressing towards 3-D movies and games in the comfort of the living room. Research concerning visual discomfort related to stereoscopic displays is socially broadly based, since part of the population has some binocular deficit which could lead to visual discomfort when viewing stereoscopic content. It appears difficult to identify these people before they perform stereoscopic perceptual experiments, watch 3-D movies or play 3-D games, since it requires optometric screening to evaluate their binocular status objectively (Evans, 2007).

Our research provides more detailed insight in determining which optometric indicators (subjective, objective and performance based) are appropriate to predict which persons might experience visual discomfort when viewing stereoscopic displays. As a relatively simple indicator, the ratio of performance of the WRRT between 2-D and 3-D in crossed disparity, detects people who are susceptible to visual fatigue when viewing stereoscopic displays. Even though the specific ratio depends on the virtual viewing distance, it is consistent with susceptibility: participants that are susceptible have poorer reading performance in 3-D than in 2-D compared to people with normal binocular vision (i.e., without any associated visual discomfort of non-strabismic binocular anomalies in normal viewing situations). Such a relatively simple tool has potential to serve as a BIS test in perceptual research, possibly in addition to stereo acuity, visual acuity, and vergence facility tests, and in consumer applications to set individual norms for comfortable screen disparities based on viewers' binocular status.

5.3. Experiment 8: Thresholds in screen disparity for comfort

ABSTRACT

Previous research revealed that among people with normal vision, the susceptibility to visual discomfort resulting from stereoscopic content is related to individual binocular visual characteristics, i.e., binocular status (BIS). It was hypothesized that BIS also affects the threshold of comfortable viewing as a result of screen disparity.

We performed an experiment in which participants were asked to 1) perform the Wilkins Rate of Reading Test (WRRT) at three disparities (-1.5, 0 and 1.5 degrees) preceded and followed by fusion measurements and self-reports and 2) scale the screen disparity of stimuli to a threshold of comfort. The BIS of participants was categorized based on the ratio of the WRRT between 2-D and 3-D (WRRT-ratio) and validated with objective signs in the fusion range and subjective symptoms in self-report. The WRRT-ratio distinguished two subgroups: moderate BIS (MBIS) (N=6) and good BIS (GBIS) (N=27).

Our results reveal that only participants with an MBIS show trends in fusion range indicating visual fatigue, report significantly more visual discomfort in stereoscopic conditions and have lower thresholds in screen disparity for comfortable viewing than participants with a GBIS. Hence, combining fusion range data with self-report is appropriate to evaluate visual discomfort resulting from stereoscopic content. The WRRT-ratio has potential as a BIS test to be used in consumer applications to set individual norms for comfortable screen disparities.

5.3.1. Introduction

Experiments 6 and 7 showed that visual discomfort is not experienced equally by all people because of differences in the binocular visual characteristics of individuals. The development of guidelines and norms that enable comfortable viewing for stereoscopic displays should therefore take the binocular status (BIS) of people into account.

If viewers do not have some form of a binocular anomaly, the tolerances in fusion and AC systems allow adaptation to accommodation and vergence mismatches within one degree of screen disparity. Hence, to guarantee a comfortable 3-D viewing experience it is recommended to adhere to the 'one degree of screen disparity' rule of thumb, which determines a zone of comfortable viewing for the general population. For most application purposes, e.g., for stereoscopic movies and games on 3-D-TVs or mobile phones, a 'one degree of screen disparity' allows for satisfactory depth rendering, while keeping the amount and severity of visual discomfort to a minimum. The binocular visual system of people who have been characterized as normal (i.e., without any associated visual discomfort of non-strabismic binocular anomalies in normal viewing situations), however, can still show individual variation (see section 6.4). There is a paucity of epidemiological data on the prevalence of binocular vision anomalies, but these anomalies have been found to affect approximately 20% of patients consulting community optometrists. Somewhat deviating percentages have been reported as well: higher percentages have been revealed, especially for people who often perform near work (Ciuffreda, 2002; Evans, 2007; Karania and Evans, 2006; Richman and Laudon, 2002; Scheiman, 1996; Stidwill, 1997). A distinction needs to be made between binocular anomalies that typically prevent stereopsis (principally, strabismus and amblyopia) and non-strabismic binocular anomalies that permit stereopsis but predispose the patient to visual discomfort (asthenopia). Consequently, the perceptual impact of stereoscopic content is not similar for all people. Experiments 6 and 7 demonstrated that a natural variation in susceptibility to visual discomfort associated with stereoscopic displays exists among people with normal binocular vision. Since the categorization of people with lower binocular abilities is generally established via extensive optometric screening, a performance test was constructed to approach this categorization. This test consists of the ratio of reading performance between 2-D and 3-D (number of words read in 3-D relative to number of words read in 2-D).

What is needed are general and easily applicable indicators of visual discomfort as a consequence of viewing stereoscopic content. In that respect, the aim of this research is to relate the individual's binocular visual characteristics to comfortable screen disparity thresholds. The categorization in BIS will be based on the reading performance test and validated by objective and subjective indicators. More specifically, we hypothesize that people with moderate BIS have a lower threshold of comfortable viewing as a result of screen disparity than people with good BIS, as illustrated with a hypothetical example in Figure 46. If this is the case, the reading performance test to categorize people's BIS can be used in consumer applications to warn viewers for potential adverse visual effects or to ensure a comfortable viewing experience by adjusting the screen disparity.

5.3.2. Background

In the literature, "visual discomfort" is used interchangeably with "visual fatigue", yet here they are distinguishable. Visual fatigue refers to a decrease in the performance of the binocular visual system as a consequence of physiological strain or stress resulting from excessive exertion (Chapter 3), and it can be objectively measured.

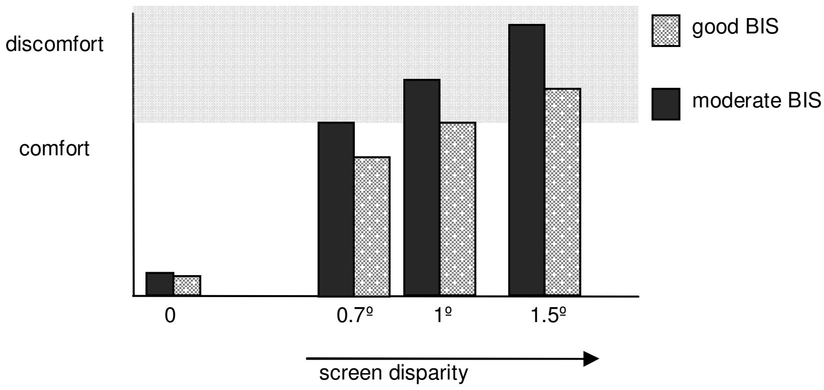


Figure 46. The assumed relationship between BIS, screen disparity and visual comfort.

Visual discomfort refers to its subjective counterpart. When formulated in this way, visual discomfort determined via subjective measurements, is expected to provide an indication of the objectively measurable visual fatigue. Note that a change in the binocular system representing visual fatigue can also be regarded as a healthy characteristic of our biological system that adapts to altered visual environments due to its high degree of plasticity (think of, e.g., prism adaptation). Only physiological changes that are accompanied by negative psychological effects in performance or comfort should be critically examined for their magnitude and subjective impact (Peli, 1998). The occurrence of visual discomfort alone, however, can be sufficient cause for concern. Consumers will be reluctant to purchase a display that induces visual discomfort, even if the visual discomfort is harmless in terms of visual fatigue. More importantly, absence of visual fatigue related to short-term viewing might still compromise the binocular visual system for longer viewing durations (e.g., two hours) during extended periods (e.g., weeks, months or years). Therefore, multiple types of measurements e.g., objective, subjective and performance measurements, need to be combined in order to determine the degree of visual fatigue and visual discomfort in a sensitive, accurate, reliable and valid way for both short- and long-term viewing.

Performance measure

It is widely reported that visual deficits (e.g., visual acuity, refractive errors, fixation disparity, fusion, accommodation and vergence) can be evaluated with reading performance, since reading utilizes dynamic accommodative and binocular oculomotor control (Wilkins et al., 2004; Jeanes et al., 1997; Buzzelli, 1991; O'Leary and Evans, 2006; Legge, 2007). Reading impairments can be properly assessed with the Wilkins Rate of Reading Test (WRRT; Wilkins et al., 1996; Wilkins, 2002; Kriss and Evans, 2005; O'Leary and Evans, 2006. This WRRT is discussed in more detail in Chapter 5.3.4. Experimental set-up.

In Experiments 6 and 7 the performance of the WRRT was related to people's BIS. Based on the optometric algorithm constructed by Evans (1997), participants with a good BIS were distinguished from those with a moderate BIS. Evaluation of the BIS using Evans' algorithm consists of the computation of a score that is the cumulative value of ten single scores that each relate to the result of a single optometric indicator of decompensated heterophoria or binocular instability. Based on the Evans' algorithm, in Experiments 6 and

7, 18% of the participants were categorized with a moderate BIS, in line with the 10%-20% described by Evans (2007). These people with a moderate BIS performed differently from people with a good BIS when reading 2-D and stressful 3-D passages of the WRRT. The ratio of performance of the WRRT between 2-D and 3-D at crossed disparity (WRRT-ratio) was demonstrated to be significantly higher for people with a moderate BIS than for people with a good BIS. The specific WRRT-ratio distinguishing people according to their BIS depended on the specific viewing conditions, such as the virtual viewing distance, but under all conditions participants with a moderate BIS had a higher WRRT-ratio than participants with a good BIS, i.e., their decrease in reading performance as a consequence of increased screen disparity was larger.

Objective measures

Precise coordinated alignment of the two eyes is required for good stereoscopic depth perception (Cole and Boisvert, 1974; Ukwade et al., 2003). Visual discomfort as a consequence of stereoscopic content can be caused by multiple binocular deficits in e.g. fixation disparity, fusion range and heterophoria. People with fixation disparity and/or heterophoria, have less tolerance for disparity changes in either direction (Evans, 2007; Steinman et al., 2000). Similarly, people with small fusion ranges are less able to compensate adequately for heterophoria and fixation disparity (Evans , 2007).

In Experiment 6 multiple optometric indicators were evaluated that each addressed a different aspect of the binocular system to determine their appropriateness to measure visual fatigue associated with stereoscopic content. Participants with a moderate BIS (as determined with Evans' algorithm (1997)) revealed objective signs of visual fatigue by a shift in the fusion range towards the direction of the stimulus, most clearly indicated by the convergent fusion range. Since these physiological changes in the binocular system were accompanied by the subjective experience of visual discomfort and a decrease in performance on the WRRT, the fusion range appeared to be sensitive and specific enough to indicate visual fatigue resulting from stereoscopic displays. This observation was in line with previous research by Emoto et al. (2004), who found significant changes in fusion amplitude after viewing a longer stereoscopic movie, yet only in participants who were unable to free-fuse stereoscopic stimuli.

Subjective measures

Visual discomfort can be experienced via many different perceptions that can be properly evaluated with questionnaires (Sheedy et al., 2003; Emoto et al., 2004 and 2005; Kuze and Ukai, 2008; Pölönen et al., 2009). Sheedy et al. (2003) developed a general questionnaire to evaluate the degree of visual discomfort. Based on this questionnaire we believe that any questionnaire evaluating visual discomfort associated with stereoscopic content should at least incorporate general visual discomfort items such as *sleepiness*, *discomfort*, *headache*, *double vision*, *irritated eyes*, *strain in the eyes* and *sharpness*. It might be useful to include additional items relating to a specific application or requesting background information such as previous experience with similar applications. In our specific case, when the reading performance is measured at different screen disparity levels, the Convergence Insufficiency Symptom Survey (CISS; Borsting et al., 2003) is appropriate to use since it includes items that specifically address visual discomfort perceived whilst reading.

To determine the threshold of comfortable viewing as a result of screen disparity, the method of limits (also known as the tuning method) is used (Gescheider, 1997; Shy et al., 2003). In this method, participants gradually and systematically change a specific characteristic of a stimulus until a threshold in a perceived characteristic is attained. This method can be used to let participants increase the screen disparity until the threshold of visual discomfort is reached, i.e., the screen disparity that is perceived as being just (un)comfortable. Possible drawbacks of the method are that participants either tend to repeat previous responses (error of habituation), or that they anticipate the threshold before it has even occurred (error of expectation) (Gescheider, 1997). In ascending order the former error affects the data by falsely increasing the threshold, whereas the latter error falsely decreases the threshold. In descending order the errors have the reverse effect on the threshold values.

5.3.3. The current study

The present research aims to provide a sensitive, accurate, reliable and valid measurement of visual discomfort associated with stereoscopic content by combining objective, subjective and performance indicators. In specific terms a first objective is to investigate whether the WRRT-ratio used to categorize viewers' BIS is a good measure to predict objective signs of visual fatigue in the fusion range and subjective symptoms of visual discomfort. We hypothesize that the combination of fusion range measurement and self-report is appropriate to indicate visual fatigue and visual discomfort of those people with a moderate BIS. Secondly, we expect that the threshold of comfortable viewing as a result of screen disparity directly relates to the BIS of people, with higher disparities generating more visual fatigue and visual discomfort for people with a moderate BIS than for people with a good BIS, as illustrated in Figure 46. The results of this research should enable guideline definition for individual settings in screen disparity that should result in comfortable viewing for consumers.

5.3.4. Experimental set-up

Design

The experimental design consisted of two parts. In the first part, the WRRT was performed under three different screen disparity conditions (Disparity): -1.5, 0 (2-D) and 1.5 degrees, where '-' refers to crossed screen disparity. A different WRRT page was assigned randomly to each screen disparity level, and the order of the screen disparity levels was randomised across participants. Each WRRT was preceded and followed by an objective fusion range measurement and a subjective questionnaire. Based on the ratio of WRRT between 2-D and crossed 3-D (WRRT-ratio) participants were assigned to two subgroups: moderate BIS (MBIS, N=6) and good BIS (GBIS, N=27). Note that neither the participants, nor the experimenter knew to which subgroup a participant belonged to during the experiment. In the second part (hereafter referred to as the tuning experiment), the screen disparity of five stereoscopic stills (Image) was tuned to a threshold of comfort for both crossed and uncrossed screen disparity.

Participants

Thirty-three participants, employees as well as graduate students working in a research environment, participated. Eight were male and twenty-five were female, with a mean age of 23 years (range between 19 and 34 years). All had a good visual acuity of ≥ 1 (tested with the Landolt C-test) and a good stereo acuity of ≤ 30 arc seconds (tested with the RANDOT stereo test).

Measurement methods

The participants' BIS was categorized based on the WRRT-ratio, i.e., the number of words read in 2-D divided by the number of words read in the crossed 3-D condition (i.e, -1.5 degrees).

The objective impact of the stereoscopic stimuli on the binocular visual system was evaluated with pre-post-measurements of the fusion range, for which the measurement principle is depicted in Figure 47 (Evans, 2007). The fusion range was characterized with prism loads, and measured for three different conditions: (1) at the point where accommodative relaxation reached its maximum and binocular single vision became blurred (blur point), (2) at the point where fusion vergence reached its maximum and single vision was lost (break point), and (3) at the point where single vision was recovered again (recovery point). We only measured the fusion range in the convergent direction, since (1) too many fusion range measurements would stress the binocular visual system, and (2) the convergent direction appeared more appropriate than the divergent direction to indicate physiological changes associated with stereoscopic stimuli (Experiment 6). It is known that large differences in fusion range can be obtained solely by mental effort expended by participants; some force fusion during the measurement, whereas others simply gaze at the target (Evans, 2007). In an attempt to limit this variation, all participants received similar instructions concerning the fusion range measurement: "look at the target normally, but continue to concentrate on it throughout the test".

The subjective impact of the stereoscopic stimuli was evaluated with the CISS questionnaire that consisted of 15 items including general asthenopic items as well as specific items concerning reading. The items were: *discomfort*, *loss of concentration*, *double vision*, *sleepiness*, *sharpness*, *exhaustion*, *appearance of moving words*, *slower reading*, *losing position in text*, *trouble remembering words*, *re-reading words*, *headache*, *pain in the eyes*, *strain in the eyes* and *irritated eyes* (Borsting et al., 2003). Participants were asked to assess the items on a scale labelled with the adjective terms: never / infrequently / sometimes / fairly often / always. The resulting scores were transformed into numerical values ranging from zero to four.

Equipment

For measuring the fusion range, Risley prisms with a range of 0 to 30 prism diopters in both the convergent and the divergent direction were used (note that only measurements were performed in the convergent direction). A Risley prism is depicted in Figure 47. The prism loads through which participants have to perceive an object can be easily increased or decreased by turning a knob. When measuring the fusion range, the change in prism load was around one prism diopter per second and the target. The object of focus was the a vertical row of letters of about 0.5 degrees in horizontal direction and about 6 degrees in vertical direction. The object had enough detail to induce accommodation, but was resolvable by the eye with the worst acuity (which was in line with recommendations made in Evans, 2007).

The stimuli were displayed on a 22” Planar SD2020 Stereoscopic Monitor, which directed the left- and right-eye images to the appropriate eye via a polarization technique. It consisted of two active matrix LCDs with a resolution of 1600 x 1200 positioned under a 110º angle. A passive, polarized mirror was positioned in between the two LCDs, and acted as a beam splitter. Viewers were able to perceive stereoscopic content with a pair of polarized glasses.

The toolbox Psychtoolbox in Matlab was used to control the Stereoscopic Monitor, to position the stimuli properly on the two displays and to tune the screen disparity of the stimuli.

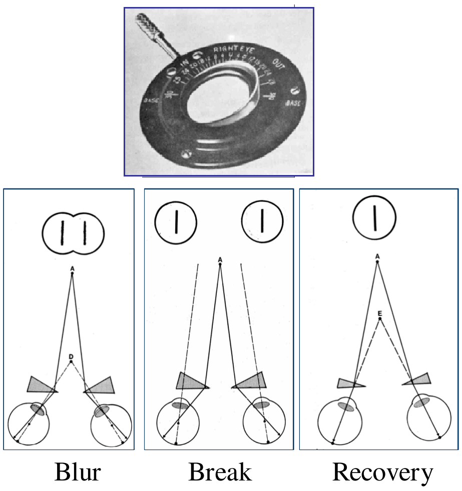


Figure 47. A photograph of a Risley prism and the principle to measure the fusion range. By adjusting the prism load of the Risley prism, characteristic fusion values, i.e., the blur, break and recovery point, can be easily obtained.

Stimuli

The stimuli in the first part of the experiment were three different passages of the Wilkins Rate of Reading test (WRRT; Wilkins et al., 1996) that were randomly assigned to conditions. For a complete description of the WRRT see Chapter 5.1.4. Experimental set-up. Since in Experiment 6 a modified version was used, Figure 42 depicts a screen shot of a stimulus as it was used in the current experiment. The visual angle of the text, inner and outer frame were 6.45, 10.08 and 11.53 degrees in horizontal direction and 3.17, 3.69 and 5.67 degrees in vertical direction respectively.

Five stimuli were used in the tuning part of the experiment: a version of the WRRT and four still images depicted in Figure 49. The horizontal and vertical visual angles of all four still images were 6.93 degrees. The text in the in the upper part of the image *Search* had a vertical visual angle of 1.61 degrees. Two of these images, *Search* and *Labyrinth*, facilitated a task that required participants to keep the stimuli fused and in focus for a requested minimum period.

The image *Search* was presented in two versions: one in which the image plus the text could be tuned in terms of screen disparity and one in which only the objects could be tuned in screen disparity. The latter version forced participants to switch continuously between perception of the 2-D text and the 3-D objects. The *Medical* image was incorporated since it contained a high level of detail. The *Bureau* image contained relative stereoscopic depth, i.e., stereoscopic depth within the image. This stereoscopic depth was created with a stereoscopic studio camera in a toed-in configuration with a convergence distance of the cameras of 1.30 m and a base distance of 80 mm. All other stimuli contained no relative stereoscopic depth, i.e., the left- and right eye image were equal and rendered with a fixed translation (i.e., disparity) with respect to each other.

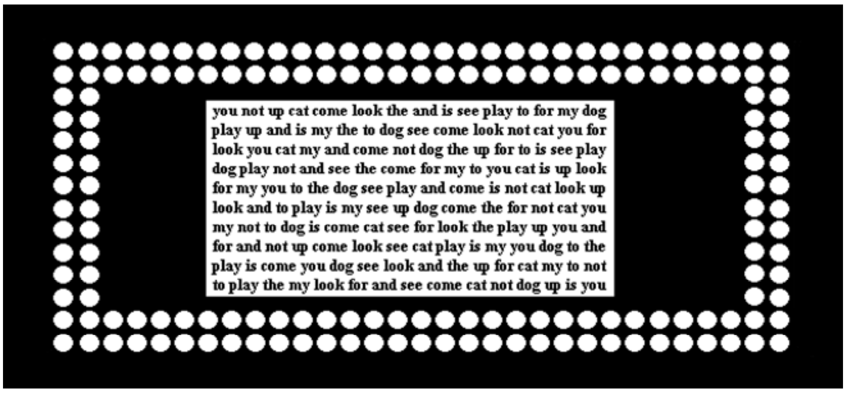


Figure 48. Screen shot of a stimulus; a passage of the Wilkins Rate of Reading test (Wilkins et al. (1996).

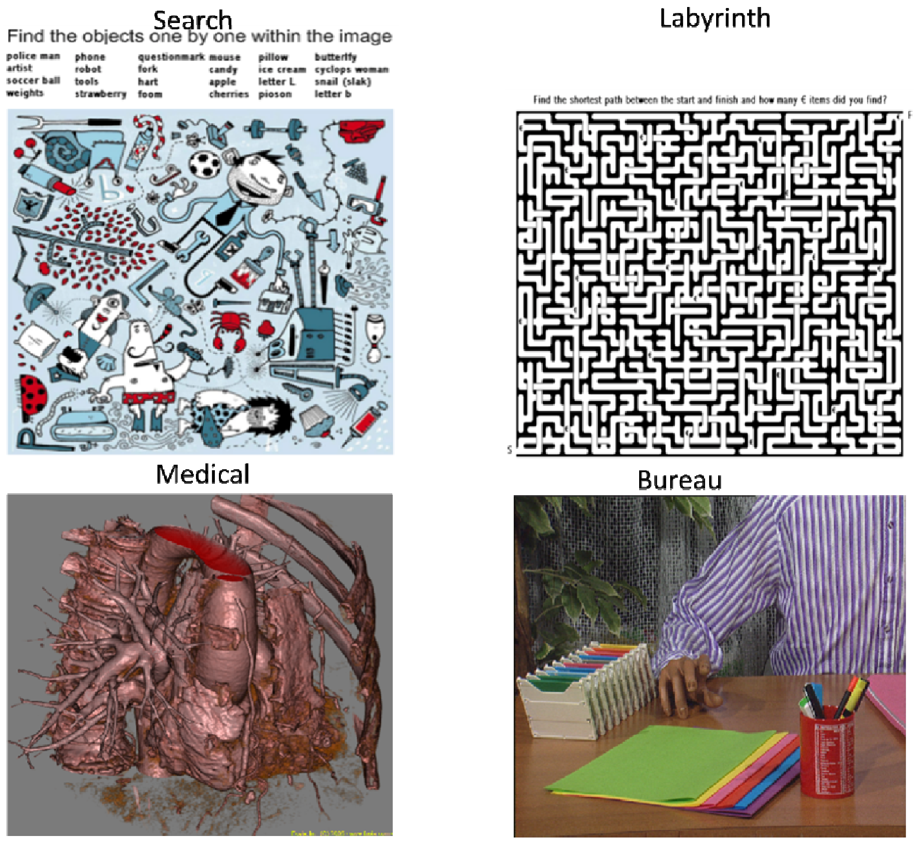


Figure 49. Images used in the second part of the Experiment 8, i.e., the tuning part.

Procedure

Participants were provided with an informed consent statement containing information about the experiment and explaining possible occurrence of visual discomfort. This statement also included the WRRT on paper to familiarize participants with the specific reading task. All questions concerning the procedure of the experiment were answered. After signing the statement, they proceeded with the experiment. Prior to the experiment, participants performed an optometric screening that included measurements of visual acuity, stereo acuity, pre-fusion range and the CISS-questionnaire. For the fusion range measurements a chin-rest with the Risley prisms attached to it was placed next to the Planar Stereoscopic Monitor. Participants were required to complete the Landolt-C test (visual acuity of ≥ 1 for both monocular and binocular vision) and the RANDOT stereo test (stereo acuity of ≤ 30 seconds of arc) before they were allowed to participate in the experiment.

Participants were provided with polarized glasses and had to rest their heads in a chin-rest positioned in front of the Planar Stereoscopic Monitor to keep the viewing distance equal. In both the first and second parts of the experiment (WRRT and tuning), the initial setting of the stimulus was always with zero screen disparity, yet participants were able to increase the screen disparity stepwise by pressing keypad number 8 (up arrow) in steps of 0.10 degrees. Depending on the condition of the stimulus the WRRT changed in stereoscopic depth: either in crossed or uncrossed disparity. In case of the 2-D condition (zero disparity), no screen disparity could be added. If participants had trouble fusing the stimuli, keypad number 2 (down arrow) could be used to decrease the screen disparity. This stepwise adaptation of screen disparity was included since Experiment 6 revealed that not all participants were able to fuse large steps in screen disparity, e.g., 1.5 degrees in one step.

In the first part of the experiment the screen disparity had to be set to any of the three screen disparities (-1.5, 0 and 1.5 degrees) according to the previously mentioned procedure. Once the proper screen disparity was set, the participants were asked to read the words in each stimulus 'out loud' as rapidly as possible for 60 seconds. It was emphasized that the text needed to be in focus (sharp) and single. Each WRRT was immediately followed by the fusion range measurement, the CISS-questionnaire and a brief rest period of approximately three minutes. The participants were instructed that each time the CISS questionnaire was carried out, their responses should relate to symptoms during the experimental condition that had just taken place. This entire procedure was performed three times (i.e., for three different disparity conditions) with the order randomised across participants.

In the second part of the experiment participants were requested to set the screen disparity of the stimuli to a threshold of comfort. Participants were asked to attend to the stereoscopic image for a minimum of 30 seconds while scanning through the entire range of screen disparities before setting the desired value. The entire experiment, i.e., optometric screening and the two experimental parts, was performed at a viewing distance of 0.70 meters and required about 30 minutes to complete.

5.3.5. Results

The WRRT-ratio between 0 and -1.5 degrees of screen disparity (crossed screen disparity) was used to categorize participants based on their BIS: MBIS (N=6) and GBIS (N=27). The cut-off for this categorization was chosen as the mean plus one STD of all participants, i.e., 1.20 + 0.25. In the following analyses of the questionnaire, fusion range and tuning results, BIS was incorporated as a binary nominal variable and all possible interactions were included.

Subjective measurements

Most of the CISS items yielded an averaged visual discomfort score around two, referring to moderate levels of visual discomfort. Not all items were affected equally by changes in Disparity or experienced equally by the two subgroups of participants. A principal component analysis (PCA) was performed on the 15 CISS items to find intrinsic correlations and to reveal if certain items shared similar underlying attributes of visual discomfort (Hair et al., 2006). The PCA was combined with a non-orthogonal rotation method (Oblimin). The resulting PCA revealed two underlying factors that respectively explained 34% and 21% of the variance of the data, which indicates that almost half of the variance remained unexplained. Factor 1 received meaningful factor loadings (>.50 (Hair et al., 2006)) of the items *discomfort*, *double vision*, *movement of words*, *slower reading*, *sharpness*, *losing position in text* and *rereading words*. Factor 2 consisted of the items *exhaustion*, *headache*, *pain in the eyes*, *discomfort* and *strain in the eyes*. These results are depicted in the left side of Figure 50. Reliability testing revealed Cronbach's alphas of 0.80 and 0.79 for factor 1 and 2, respectively (a Cronbach's alpha of > 0.70 is considered acceptable (Hair et al., 2006). The items *irritated eyes*, *concentration problems*, *sleepiness* and *remembering words* received a low factor loading on both factor 1 and factor 2. Based on the reliability analysis it was decided to exclude these items from further analysis. For each participant a factor score was calculated for each factor in each condition; i.e., the score given by the participant on each questionnaire item was weighted with the factor loading of that item, and then summed over all items of the factor. A MANOVA with BIS and Disparity as independent variables and the factor scores of factor 1 and 2 as dependent variables revealed that none of the independent variables had any significant effect on the factor scores of factor 2. Factor 1 was significantly affected by BIS ($F(1, 120) = 15.092$, $p < .001$) and Disparity ($F(3, 120) = 4.552$, $p < .005$), but their was no interaction present between BIS and Disparity (p = .187) The results of the factor scores of factor 1 are depicted in the right side of Figure 50. Note that an item yields a negative factor score when the item is negatively related to that factor.

The results show that the seven items that load on factor 1 indicate no increase in visual discomfort in any of the disparity conditions for participants that are categorized as GBIS. Participants that are categorized as MBIS, however, indicate more visual discomfort in the 3-D conditions, both in the crossed and uncrossed direction.

Objective measurements

Figure 51 depicts the averaged fusion range in convergent direction measured at the blur point (upper left graph), break point (upper right graph) and recovery point (bottom graph), including the 95% confidence interval. The four conditions, i.e., pre-, 2-D-, 3-D uncrossed-, and 3-D crossed are mentioned along the x-axes, and the coloured bars represent the two groups in BIS.

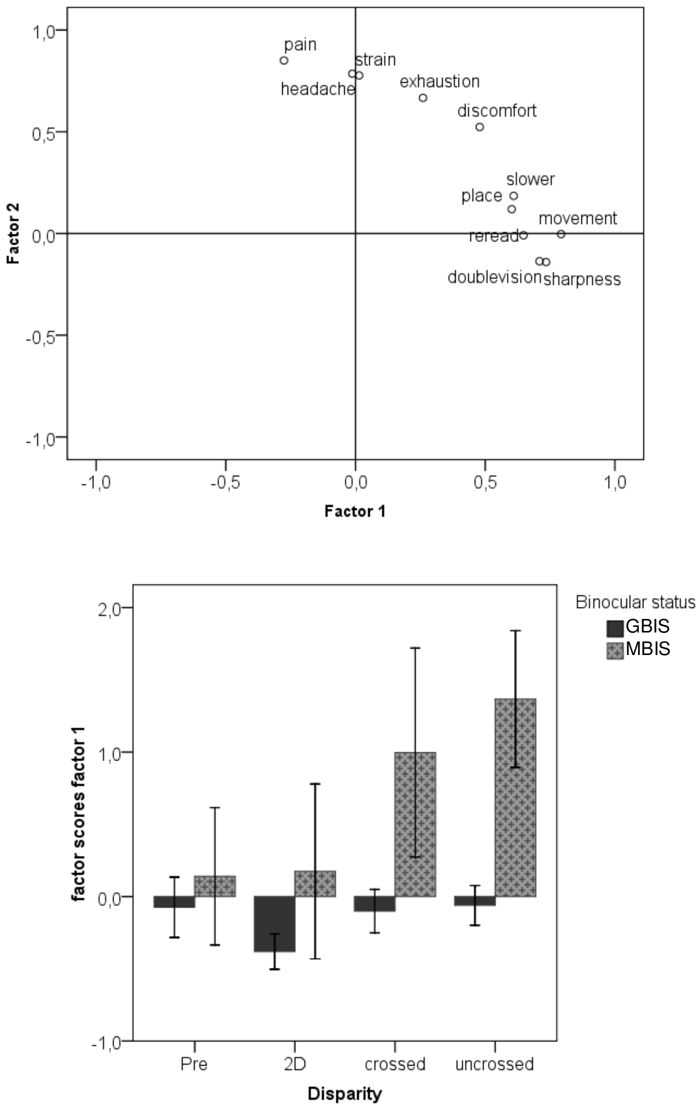


Figure 50. Results of the principal component analysis. Upper panel: the items with their factor loadings in a factor plot. Lower panel: the factor scores of factor 1 with 95% confidence intervals as a function of two independent variables. The x-axis represents the Disparity, the y-axis the standardized factor scores and the different columns represent the categorization in BIS.

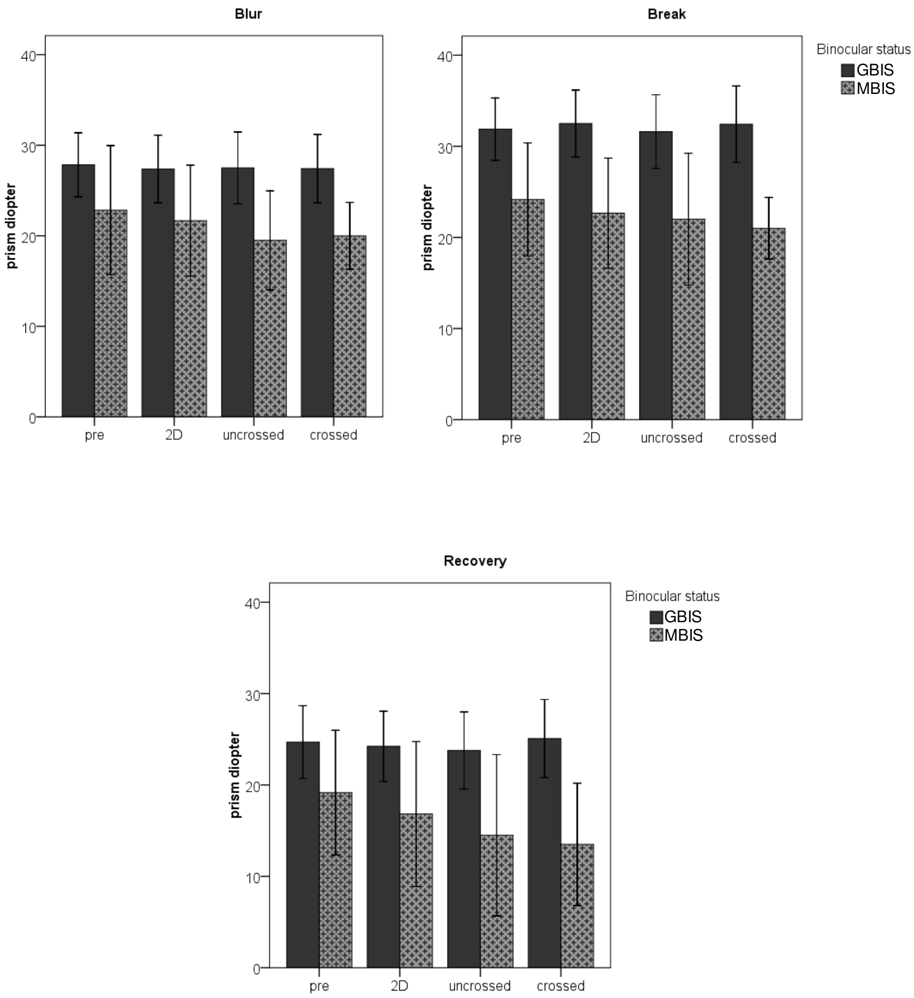


Figure 51. The mean convergent fusion range with 95% confidence interval as a function of BIS and Disparity. The different graphs refer to different fusion characteristics (blur, break and recovery in the upper left, upper right and bottom graph, respectively). The x-axes represent the variation in Disparity, the y-axes the fusion range values in prism diopters and the colours the BIS categorization.

An MANOVA was performed with these three fusion range values as dependent variables and BIS and Disparity as independent variables. The results confirm a significant effect of the categorization in BIS for the blur ($F(1, 120) = 10.524$, $p < .001$), break ($F(1, 120) = 22.060$, $p < .001$) and recovery ($F(1, 120) = 14.395$, $p < .001$) fusion range values. There is no significant difference between Disparity conditions nor does Disparity interact with BIS, though for participants with a MBIS the fusion range seems to be considerably lower in the 3-D conditions. This trend is especially visible in the recovery values.

Thresholds tunings

Figure 52 depicts result of the second part of the experiment; the averaged threshold of comfortable viewing as a result of screen disparity with the corresponding 95% confidence interval. For uncrossed screen disparities participants with a GBIS had a slightly higher threshold than participants with a MBIS (0.46 vs. 0.40 degrees), whereas the difference was considerably larger for crossed disparities (1.24 vs. 0.72 degrees).

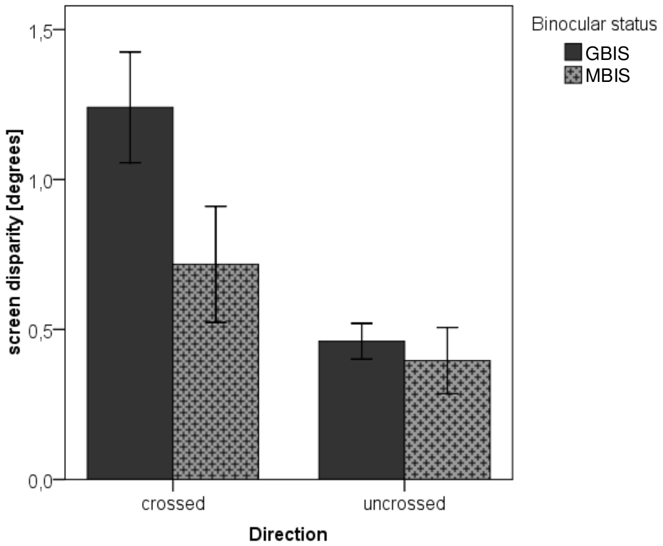


Figure 52. The mean threshold of comfortable viewing as a result of screen disparity with 95% confidence intervals. The x-axis represents the two Disparity conditions, i.e., uncrossed and crossed, the y-axis represents the averaged tuned thresholds and the coloured bars represent the two groups in BIS.

An ANOVA with Image, Disparity and BIS as independent variables and the tuned threshold as dependent variable confirmed that the threshold was significantly lower for uncrossed disparities than for crossed disparities ($F(1, 295) = 30.137$, $p < .001$), and was significantly lower for participants with a MBIS than for those with a GBIS ($F(1, 295) = 8.682$, $p < .005$). In addition, participants with a MBIS had a lower threshold in the crossed disparity condition (($F(1, 295) = 5.463$, $p < .05$), but not in the uncrossed disparity condition. Image did not significantly affect the threshold.

5.3.6. Discussion

Some people report visual discomfort as a result of watching 3-D displays. We performed an experiment to evaluate the visual discomfort associated with stereoscopic viewing by combining multiple types of measurements; e.g., fusion range, questionnaire, tuning preference and reading performance. Two subgroups (GBIS and MBIS) were asked to, first, perform the WRRT at three disparities (-1.5, 0 and 1.5 degrees), and second, adjust the screen disparity of stimuli to the threshold of comfortable viewing. The categorization in

BIS was based on the WRRT-ratio. Results reveal that only those participants with a MBIS indicate trends of visual fatigue in the fusion range, report significant visual discomfort in stereoscopic viewing conditions, and set the screen disparity to a lower value for comfortable viewing. We conclude that the WRRT-ratio is a proper measure for BIS and thereby appropriate to predict visual fatigue and visual discomfort resulting from stereoscopic viewing. Hence, it has potential as a tool to be used in consumer applications for the purpose of setting individual norms for comfortable screen disparities.

Though the WRRT-ratio by itself is sufficient to predict the adverse side-effects of stereoscopic viewing, this needs to be interpreted with caution for two reasons. First, no single test to evaluate the BIS is 100% effective, which is also the case for the WRRT-ratio. And second, the WRRT-ratio only provides relative information, i.e., who is more prone to visual discomfort. Subjective and objective measurement methods still provide complementary information, e.g., absolute values or indications of other determinants than BIS of visual discomfort. For example, objective measurements are particularly useful when evaluating potential physical effects of long-term viewing. Subjective questionnaires can provide detailed information about the evaluation of specific video characteristics (e.g., subtitles) or display technologies (autostereoscopic vs. stereoscopic with glasses). Screen disparity preferences can be used to establish standards for different 3-D display settings. Nonetheless, the WRRT-ratio is still of use in predicting amongst people with normal binocular vision every day life, yet who are susceptible to visual discomfort resulting from stereoscopic viewing and who are not.

Sample size

It must be mentioned that although our findings are in line with those of Experiment 6 and 7 and are robust, the sample sizes used in these studies are still relatively small. All participants in our research had normal binocular vision without any large abnormalities (indicated by normal fusion ranges). More specifically, even amongst people with unimpaired binocular vision, some people are more susceptible to visual discomfort based on small differences in the BIS. Larger scale studies would be required to be confident that these results generalize across the entire population. In particular, children were not tested in the current studies, which is an important limitation that should be addressed in future studies of this kind.

First hypothesis

With respect to the first hypothesis, the combination of fusion range measurement and self-report is appropriate to evaluate visual discomfort of those people with a moderate BIS. The fusion range for this subgroup of people is lower than that for participants with GBIS. In addition, the fusion range characteristics of participants with MBIS decrease considerably after viewing an excessive screen disparity, in contrast to the fusion characteristics of participants with a GBIS. This is best visible in the recovery value, which is approximately 30% (six prism diopter) lower compared to the pre-measurement, and approximately 20% (four prism diopter) lower compared to the 2-D-measurement. Although these differences are barely significant or meaningful from a clinical point of view (Peli, 1998), the effects were obtained with only short-term viewing. Moreover, all participants had normal binocular vision (see section 6.4). Future work could reveal if longer or more stressful stimuli have a more profound impact on the visual system, though this raises ethical issues since long-term visual discomfort and headaches might be induced.

The recovery value is indicative of the unprompted readiness of the fusion system to adapt and regain single binocular vision (Ciuffreda et al., 2006). Since blur and recovery values revealed a small yet similar trend, the signs in recovery value can be indicative of visual fatigue. The indication of visual fatigue in fusion amplitude is in line with previous research. Emoto et al. (2004) found a significant decrease in fusion amplitude viewing after a long-term stereoscopic movie, yet only in participants who were unable to free-fuse stereoscopic stimuli. The decrease they found in fusion range was similar to the decrease we found in fusion range (± 30%), yet was revealed in the break values. In Experiment 6 it was found that both convergent break and recovery values of participants with MBIS increased as a consequence of viewing stereoscopic stimuli. This increase in convergent fusion range was accompanied by a decrease in divergent range, and thus, could be indicative of visual training (Peli, 1999). More specifically, as a consequence of five preceding stereoscopic stimuli with only crossed disparity, the change could be interpreted as an accumulative change or adaptation of the fusion range in the direction of the stimulus. In the current experiment, an adaptation in the direction of the stimulus did not occur since stimuli with different disparities, i.e., crossed and uncrossed, were randomly assigned to the participants. The common finding in all three experiments is that the fusion system of participants with MBIS is more susceptible to changes as a consequence of stereoscopic stimuli than the fusion system of participants with GBIS. Since these changes are accompanied by visual discomfort and a decrease in performance, it is referred to as visual fatigue.

The two underlying factors in visual discomfort, i.e., referring to Disparity and BIS, explained 55% of the variance in the data set, indicating that additional factors or noise might be present. At this point, however, we are mostly interested in how these two underlying factors relate to visual discomfort. Since these factors are similar to those found in Experiments 6 and 7, a meta-factor analysis across experiments can validate whether specific items relate to specific underlying attributes of visual discomfort. Since the disparity settings differed between experiments, the different disparities were simplified into 2-D or 3-D..

The resulting PCA revealed two underlying factors that respectively explained 49% and 19% of the variance of the data. These results are depicted in Figure 53. Factor 1 received meaningful factor loadings of the items *discomfort*, *double vision*, *movement of words*, *slower reading*, *sharpness*, *losing position in text* and *rereading words*. Factor 2 consisted of the items *headache*, *irritated eyes*, *pain in the eyes*, *discomfort* and *strain in the eyes*. Based on an ANOVA with BIS and Disparity as independent variable and factor scores as dependent variables, factor 1 was significantly affected by Disparity ($F(1, 330) = 4.473$, $p < .05$) and factor 2 was significantly affected by BIS ($F(1, 330) = 3.910$, $p < .05$). The items *concentration problems*, *exhaustion, sleepiness* and *remembering words* did not relate to any of the two factors. This does not indicate that they are useless, as we believe they are more likely to relate to long-term visual discomfort.

Hence, in line with our previous statements in Chapter 3, visual discomfort questionnaires associated with stereoscopic displays should at least incorporate this subset of asthenopic items (discomfort, headache, strain, sharpness, double vision, irritation, pain) extended with items that relate to experimental settings or stimuli.

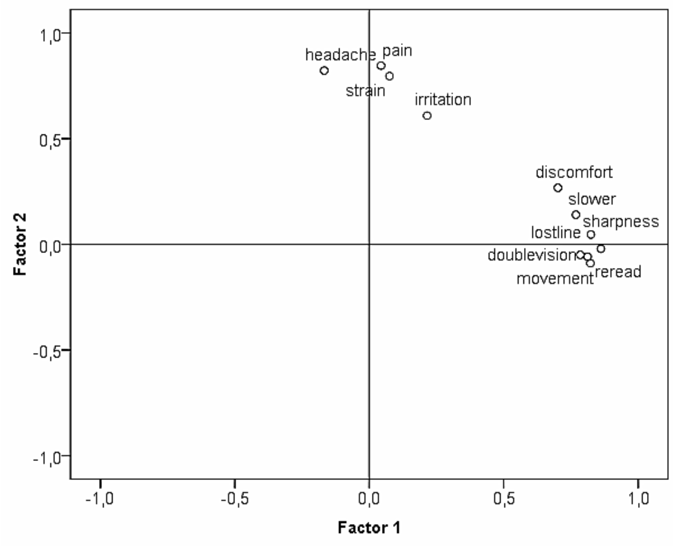


Figure 53. Results of the meta-factor-analysis across experiments. The graph shows the factor loadings of all visual discomfort items on the two factors. The x-axis represents factor 1 and the y-axis represents factor 2.

Second hypothesis

With respect to the second hypothesis, the t threshold of comfortable viewing as a result of screen disparity relates to the BIS of people; people with MBIS have a lower threshold. This lower threshold of people with MBIS was significant only for crossed disparities. This can be attributed to lower plasticity in the divergent fusion range than in the convergent fusion range (Evans, 2007), which makes it more difficult for the human visual system to diverge than converge independent of small differences in BIS. The relative difference in thresholds between participants with MBIS and GBIS is of more importance than their absolute threshold value. First, it is plausible that by increasing disparity stepwise at participants' own pace, adaptation to higher screen disparities is facilitated (error of habituation) or lower screen disparities can be the consequence when participants already are experiencing visual discomfort due to preceding stimuli (error of expectation). Second, when real-life dynamic content is used, thresholds presumably become lower since such content often includes video characteristics (e.g., motion and changing screen disparity) that induce visual discomfort as revealed in Experiments 4 and 5. Hence, visual discomfort is scene-dependent; adjusting screen disparity limits should relate to the content as well as to the viewer's BIS. The WRRT-ratio as a BIS test can still be of added value in consumer applications. Ideally, the BIS test should be used to decrease screen disparity automatically to allow comfortable viewing, which is possible with popular acknowledged formats for stereoscopic content (Redert, 2007). The test can be implemented in, e.g., the set-up menu and be used to warn the viewers who are more susceptible to experiencing visual discomfort.

5.3.7. Conclusion

Since some people report visual discomfort while or after viewing stereoscopic content, it is important to construct measurement protocols to evaluate their impact. Our research indicates that it is important to incorporate differences in binocular status of people in such protocols, since these partly determine the perceptual impact of stereoscopic content. A specific performance measure (WRRT-ratio, i.e., the number of words read in 2-D divided by the number of words read in the crossed 3-D condition) can be used to identify people with a moderate binocular status amongst people with a normal binocular status (i.e., without any associated visual discomfort of non-strabismic binocular anomalies in normal viewing situations). People with a moderate binocular status are more susceptible to visual discomfort resulting from watching stereoscopic content, as is shown by the objective signs of visual fatigue in the fusion range and on subjective symptoms of visual discomfort as a consequence of short-term stressful stereoscopic reading tasks. The WRRT-ratio test is very promising, since it can be implemented in 3-D commercial displays as a screening tool, allowing display manufacturers to warn consumers who might be more susceptible to visual discomfort yet not aware of their own moderate binocular status. They can then be advised to lower the screen disparity range, or it can be done automatically.

-Chapter 6-

Discussion

"A painting, though conducted with the greatest art, and finished to the last perfection, with regard to its contours, its lights, its shadows, and its colours, can never show a relief equal to that of the natural objects unless these be viewed at a distance and with a single eye."

Leonardo da Vinci (1584) in Wheatstone (1838; p. 2)

6.1. General discussion

Optimizing image quality (e.g., by creating displays with higher resolution, contrast and brightness or a more realistic colour rendering using a wider colour gamut) remains one of the main goals in the development of imaging systems. Yet, enriching the overall viewing experience is also gaining attention in research on innovative, next-generation displays such as three-dimensional television (3-D TV; Heynderickx, 2006).

3-D TV goes beyond optimizing image quality by displaying stereoscopic depth, i.e., part of the image content is rendered such that it is projected behind or in front of the display screen. In most cases this can only be achieved at the expense of spatial and/or temporal resolution (i.e., at the expense of important image quality aspects), whereas some people can in addition experience visual discomfort when watching 3-D content. Hence, the total visual experience of a 3-D display is expected to be a combination of image quality (including 2-D and 3-D aspects), the added value of having stereoscopic depth and the possible annoyance of visual discomfort as depicted in Figure 54. The present thesis is aimed at understanding, measuring and eventually, modelling and predicting any added value of stereoscopic depth and the accompanying visual discomfort associated with 3-D TV.

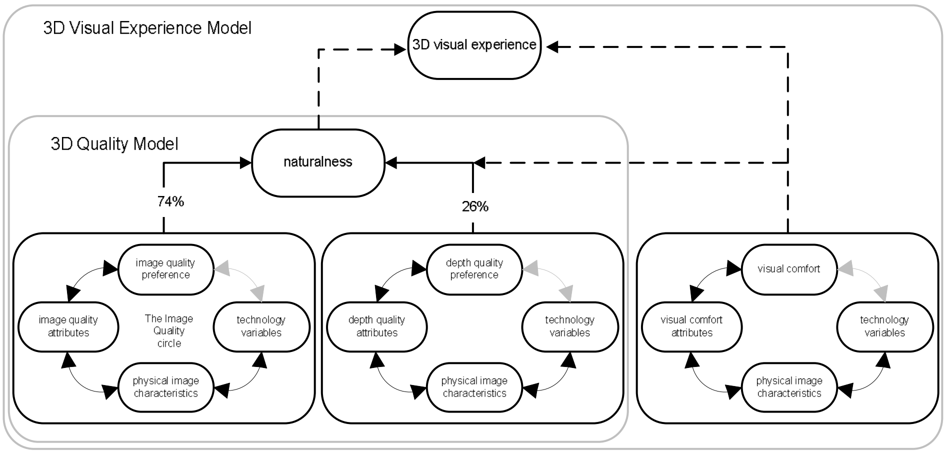


Figure 54. 3-D Visual Experience Model, which describes the overall 3-D visual experience as a weighted combination of image quality, depth and visual comfort. The 3-D Quality Model, which describes naturalness as a weighted combination of image quality and depth,is part of the 3-D Visual Experience Model.The solid lines between blocks are confirmed in this thesis, whereas the dotted lines require more research.

6.2. The added value of stereoscopic depth

A first step in constructing the 3-D Visual Experience Model, is to evaluate the balance between the added value of stereoscopic depth and the image quality. Since display manufacturers are in general interested in the relationship between the technology parameters of the imaging system and accompanying image quality and perceived depth, the framework of Engeldrum's Image Quality Circle appeared appropriate for both concepts. In Chapter 2, we therefore propose the 3-D Quality Model as an extension of Engeldrum's Image Quality Circle, since this Image Quality Circle does not incorporate the added value of depth. The 3-D Quality Model describes a higher-level evaluation metric as a weighted sum of image quality and depth and is depicted in Figure 54. The results show that perceived image quality and perceived depth are not independent in their relationship to physical image characteristics, but are at the perceptual level. Based on naturalness as a higher level evaluation metric, the model is validly applicable to stereoscopic stills, and the value of naturalness is determined for approximately 74% by image quality and for approximately 26% by the added value of stereoscopic depth as depicted in Figure 54. An important aspect is that these results are consistent over a range of different 3-D displays, content generation methods (thereby reflecting different depth percepts) and image quality attributes. Future research should reveal if naturalness is also appropriate to evaluate stereoscopic movies, and if so, whether similar weights apply.

6.3. Describing visual discomfort

A second step in constructing the 3-D Visual Experience Model is to conceptualise visual discomfort in order to understand the underlying perceptual mechanisms and their relationships with specific technology variables, i.e., the relationships that are modelled in the right part of the Figure 54.

In Chapter 3 it is recommended to adhere to the 'one degree of screen disparity' rule of thumb, which defines a zone of comfortable viewing. Though this zone often serves as a rule-of-thumb, we recommend it here as a limit for a zone of comfortable viewing similar to auditory limits; i.e., it is not recommended to set the volume to the maximum level for extended periods. For most application purposes, e.g., 3-D movies and games and 3-D mobile phones, this limit allows satisfactory depth rendering, whereas the most frequently mentioned determinants for visual discomfort, i.e., excessive binocular parallax and the AC conflict, do not induce any visual discomfort within this limit. The description of the zone of comfortable viewing in this thesis is currently dichotomous: a region that is comfortable and everything outside of that region is uncomfortable. I acknowledge that this description is certainly an over-simplification. Conflicts that are just outside the comfort zone are surely less than conflicts that are far outside the zone. Furthermore, some conflicts within the nominal zone are surely less comfortable than others. Hence, it makes more sense to describe the zone of comfort as continuous.

However, even within the one-degree limit, there are certain factors that can still contribute to visual discomfort. To date, we have identified four such factors: (1) fast motion in depth, (2) 3-D artifacts resulting from insufficient depth information in the incoming data signal, yielding spatial and temporal inconsistencies, (3) unnatural amounts of blur, and (4) subtitles. The perceptual impact of these four factors is directly related to certain video characteristics (technology variables); the amount of screen disparity or the amount of motion. A larger screen disparity puts more strain on the accommodation-vergence linkage, makes 3-D artifacts more visible, generates more blur, and increases cue conflicts between 3-D content and subtitles. Higher speeds of in-depth motion also increase stress in the accommodation-vergence linkage more, whereas higher horizontal speeds make the viewer more sensitive to 3-D artifacts. See the next paragraph for a more elaborative discussion. An additional exception to the one-degree limit, and perhaps the most important one, is the differences in binocular vision between people. Even amongst people with normal binocular vision, some people are more susceptible to visual discomfort based on small differences in their binocular status. See section 6.4 for a more elaborative discussion.

In Chapter 4 the aim was to reveal if these video characteristics could be used to predict the perceived visual comfort, i.e., to predict visual discomfort along the dark arrows in the visual comfort block in Figure 54.

In Experiments 4 and 5, three stereoscopic movies of a maximum length of 30 minutes were continuously assessed by participants. The moment-to-moment value of the video characteristics can be extracted from the 3-D movies with motion and depth estimation algorithms, which allows these values to be correlated directly to the continuous assessment scores of visual discomfort. In Experiment 4 it is demonstrated that the visual comfort of stereoscopic scenes can be predicted as a linear combination of screen disparity range and offset, changing screen disparity and lateral motion. Their specific contributions, however, depend on the activity of the scene. The 3-D movie in Experiment 4, however, comprised relatively simple screen disparity (i.e., objects of interest were rendered on the foreground layered over a solid background that was positioned statically behind the display plane), and was displayed on an autostereoscopic display that induced high amounts unnatural blur. In contrast, the stereoscopic content used in Experiment 5 had high spatial and temporal complexity and was rendered on a high resolution two-view stereoscopic TV, i.e., a configuration that is likely to be present in home theatre sets in the near future. Via extensive questionnaires participants indicated that they incorporated two main video characteristics in the continuous assessment; one that was motion related (*fast camera*, *speed*, and *scene switches*) and the other that was disparity related (*depth motion*, and *depth*). In line with Experiment 4, these video characteristics, e.g., (change in) screen disparity offset and range and lateral motion, provided a good approximation for the experienced visual comfort using a simple linear combination for a single scene, yet not for an entire movie. I think it is plausible that the variation in visual characteristics of the 3-D content mainly affected the instantaneous visual comfort and less the accumulative effect on visual comfort. For a pre-emptive modification of video characteristics in order to maintain a comfortable viewing experience more research is required. Note that similar problems arise when predicting image quality with objective image characteristics. More specifically, it is known which predictors are most important in estimating image quality (i.e., brightness, contrast, colour and sharpness), yet how these predictors need to be combined in such a way that they describe image quality for all sorts of content, remains elusive.

In addition, subtitles can generate some visual discomfort as a consequence of depth inconsistencies. Subtitles are often rendered at the display plane and induce depth cue conflicts by occluding 3-D objects that are rendered in front of the display plane. The impact of subtitles is also explored in Experiment 5. Results showed that the reading of subtitles required more effort, which drew the attention away from the content and negatively affected the visual comfort. In case of dynamic content (e.g., fast camera and object motion, scene switches, changes of camera perspective, changes in depth) this makes it difficult to follow the movie, resulting in more double and/or blurred images. In case of more passive it was not comfortable to perceive slow, steady background motion at a different depth layer when the eyes are fixated on the subtitles.

A limitation of our approach (correlating the CA with different video characteristics) is that CA can be limited to the assessment of fairly short sequences of e.g. ± 30 minutes. Concentration loss and mental fatigue as a consequence of the assessment task can start to have an impact when judging feature-length movies of e.g., 2 hours. In the literature, evaluations of longer movies are therefore often performed with assessment of short sequences extracted from the movie or with retrospective assessment, both requiring less effort. Since it cannot be assumed that they provide good indications of the visual comfort of the entire movie, in Experiment 4 the results of all three evaluation methods are compared with each other. Results reveal low consistency between the evaluation methods, implying the two methods might not be appropriate as a substitution. This does not mean that the visual comfort experienced after a short sequence or after the entire 3-D movie is a poor reflection of visual comfort at that moment, but it has limited value to generalize conclusions drawn from the assessment of such short sequences to the entire movie.

6.4. Measuring visual discomfort

In order to design new 3-D displays with the right balance between stereoscopic depth at the expense of a reduced image quality and tolerable (if any) visual discomfort, it is very relevant to extend the 3-D Quality Model by incorporating visual comfort as indicated in the 3-D Visual Experience Model in Figure 54. As the dotted lines in Figure 54 imply, however, the weights within this balance are left for future research. Priority is given to establish a protocol for the measurement of visual discomfort.

In Chapter 3 visual fatigue is defined as physiological strain or stress resulting from excessive exertion of the visual system, which can be objectively measured. Visual discomfort is defined as its subjective counterpart. To adequately characterize and understand both concepts, multiple types of measurements, both objective and subjective, are needed. Visual discomfort can be evaluated with validated questionnaires or other self-report measures. For visual fatigue clinical measurement methods are best suited since they are relatively cheap, concise, quantitative with a high sensitivity and specificity and applicable to a large group of participants. Subsequently, in Chapter 5 a measurement protocol is constructed for the evaluation of objective signs of visual fatigue and subjective symptoms of visual comfort associated with stereoscopic displays.

In Experiment 6 the objective is to compare different measurement methods for their appropriateness to evaluate visual discomfort associated with stereoscopic displays. Results reveal that a combination of fusion range measurements and self-report is appropriate for evaluating visual discomfort of stereoscopic stills. Future research should reveal if this combination is also appropriate for stereoscopic movies, though previous research does strongly suggest this. Another appropriate method to evaluate visual discomfort caused by stereoscopic movies might be vergence facility. It addresses the adaptivity of the binocular system that is affected by stereoscopic movies with high amounts of varying screen disparity. The first results, however, were not very promising (Experiment 5).

An important aspect of such a measurement protocol is to incorporate differences in binocular status of people. Even among people with normal binocular vision there is a natural variation in the performance of the binocular system that can determine their susceptibility to visual discomfort. In Experiments 6 and 7 the categorization of participants' binocular status is accomplished based on an extensive optometric screening (i.e., an optometric algorithm) resulting in one subgroup with a good binocular status and one with a moderate binocular status. Results confirm that only participants with a moderate binocular status experience visual discomfort associated with stereoscopic content based on objective signs of visual fatigue in the fusion range and subjective symptoms of visual discomfort.

In addition, the results of Experiments 6 and 7 reveal a specific performance measure that is indicative of the binocular status of people; the WRRT-ratio, which describes the number of words read in 3-D relative to number of words read in 2-D. In Experiment 8 the WRRT-ratio is used to categorize participants' binocular status, and the evaluation of visual discomfort is performed by a combination of fusion range and self-report. In addition, the categorization of the WRRT-ratio is related to thresholds of comfort in terms of screen disparity. The results show that only participants with a moderate binocular status show trends of visual fatigue in fusion range. They also suffer significantly more visual discomfort in stereoscopic conditions and have lower thresholds of visual comfort than participants with a good binocular status. However, no single test to evaluate the BIS is 100% effective, which is also the case for the WRRT-ratio. The WRRT-ratio has a high sensitivity index of 2.4, indicating a high discriminating power. The identification of participants who experience visual discomfort and/or reveal changes in visual functions can be predicted by the WRRT-ratio to a similar extent as by other optometric indicators.

Hence, even amongst people with normal binocular vision, some people (approximately 15% to 20%) are more susceptible to visual discomfort based on small differences in their binocular status. It is important to emphasize that our results show that the largest part of people with normal binocular vision (> 80 %) do not experience any significant visual discomfort and viewing 3-D movies should be comfortable. One of the restrictions of these results is that they cannot be generalized to people with large refractive errors, amblyopia (lazy eye), stereo blindness, children, older people or people with visual impairment. Another restriction is that although our results are robust, our research has thus far included fairly modest sample sizes, and a single valid ratio for a close viewing distance and a large viewing distance has not been established yet. Larger scale studies would be required to confirm the sensitivity and specificity of the test and to be confident that these results generalize across the entire population.

Idiosyncrasy of binocular status

In this thesis the concepts 'normal binocular vision' or 'normal vision' have often been used, implying that everyone who has normal vision are likely to have the same binocular status without considering individual differences. The definition used throughout for 'normal vision' in this thesis is 'vision without any associated visual discomfort of non-strabismic binocular anomalies in normal viewing situations'. It is important to mention that the "person with normal binocular status" is a concept, not a person nor a group. For example, in section 3.6 it is stated that people without some form of a binocular anomaly should have tolerances in their fusion and accommodation-vergence systems to adapt to conflicts within one degree of screen disparity. The difficulty with this statement is that it implies that there is an absolute cut-off between people whose ocular motor status is "normal" and "abnormal", which is not the case. It seems quite likely that some people whose visual status is within the normal range may nonetheless experience visual discomfort even within the 1 degree limit. Some "normal" people have 8 prism diopter exophoria at near, and others have 4 prism diopter esophoria at near. These people would all be described as normal if they could adequately overcome their heterophoria under normal viewing conditions, but these are the people who might have trouble when viewing 3-D displays. By using Evans' algorithm an attempt was made to identify these people. The classification of the binocular status of people with normal vision by Evans' algorithm in this thesis, however, is currently dichotomous; a binocular status that is either 'good' or 'moderate', which is certainly a gross over-simplification. The reason that this issue is important is display manufacturers or movie producers who might not fully understand the diverse nature of ocular motor status within the population might assume that this means that such 3-D content cannot cause problems. This underlines the attention that needs to be given to idiosyncratic effects and thus also the important role for eye care practitioners who characterize binocular functions of viewers.

6.5. Applicability of findings

The 3-D Visual Experience Model can contribute to a lower cost design cycle for 3-D TV by taking into account the perceptual costs and benefits of the complete system. The technological parameters can be fine-tuned to optimize specific perceptual experiences e.g. depth, image quality, naturalness and visual comfort, without degrading others. The model aims at presenting maximally 'pleasing' images and movies, which can result in different implementations depending on personal preferences and situational circumstances. In current imaging systems SmartImage provides an analogue for the applicability of the model. SmartImage is an exclusive leading edge Philips technology that analyzes the content displayed on the consumer's screen and produces optimised display performance. A user-friendly interface allows the viewer to select various modes like *Office*, *Image*, *Entertainment*, *Economy* etc., to fit the application in use. Based on this selection, SmartImage dynamically optimises the contrast, colour saturation and sharpness of images and videos for maximal display performance. Also for sound settings similar options are incorporated in the display menu (e.g., *rock*, *jazz*, *live* and *pop* settings). With respect to 3-D settings, image quality, depth and visual comfort attributes can be optimized to fit specific viewing experiences, e.g., one that is most natural or realistic, one that is comfortable or one that increases the viewing experience. This development of specific 3-D viewing experiences is left for future research

The 3-D Quality Model describes a part of the 3-D Visual Experience Model in Figure 54 and can be relevant in multiple applications. The model provides insight in how people perceive 3-D quality of images and moview. It shows that people do appreciate the added value of stereoscopic depth, yet for natural content or content with a high viewing experience, image quality remains the most important aspect. In this ways the model can contribute to the construction of objective and subjective quality metrics for stereoscopic content. Subjective evaluations are still performed with 2-D quality metrics that do not incorporate recent enrichments such as 3-D. The model implies that evaluation recommendations such as the ITU-R for stereoscopic content, should improve by incorporating evaluation metrics that reflect the full extent of the user experience such as naturalness. For objective evaluation there is still a long way to go for metrics to become widely applicable and universally recognized, yet it has been acknowledged that it is essential to incorporate the perceptual impact of stereoscopic content on the human vision system when constructing objective metrics that assess 3-D quality (Benoit et al., 2008; Sazzad et. al, 2010; You et al., 2010). In most cases 2-D image quality is linearly combined with disparity information and temporal information in order to objectively predict the 3-D quality. The 3-D Quality Model provides additional weights for image quality and stereoscopic depth in these linear combinations to correct for their difference in perceptual impact, and therefore, assists in attempts to develop a perceptually based objective 3-D quality metric. A similar analogue is present in the construction of 2-D-to-3-D or other video processing algorithms that optimize the 3-D quality. The best performance of such algorithms can only be achieved when disparity information and the original image are weighted appropriately. The 3-D Quality Model provides the combination to optimize the performance of such algorithms in terms of naturalness (and viewing experience).

With respect to providing a comfortable 3-D viewing experience, we propose a relatively simple test that is indicative of binocular status: the WRRT-ratio, which describes the number of words read in 3-D relative to number of words read in 2-D. This test has potential in perceptual research to explain inter-subject differences in research results, and as such, should supplement stereo- and visual acuity tests during screening. This test can be easily implemented in 3-D consumer applications as a warning for possible visual discomfort or advice to pre-emptively decrease the range in which screen disparity can vary (or advice people to do it manually). Such a warning could be part of future international standardization of guidelines on image safety (van Nes, 2009).

One can comment that it would be more to the point to let people adjust screen disparity settings themselves when the viewing experience becomes uncomfortable. Of course, letting people adjst the screen disparity by themselves should always be an option, yet this has two drawbacks.

First, even though the viewing experience can be comfortable on the short-term, the adverse effects of too much screen disparity can increase in strength after prolonged viewing or can even induce after-effects. The results of Experiment 4 indicate that visual discomfort experienced during short-term stereoscopic viewing poorly predicts visual discomfort experienced after longer viewing periods. The WRRT-ratio relates visual discomfort associated with the binocular status, which in principle is independent of the content or period of viewing. More specifically, it is highly plausible that such visual discomfort does not deteriorate after long-term viewing, but rather increase in severity, since the binocular system will be fatigued and requires more effort to properly adjust to dynamic video characteristics. Second, despite the possibility that consumers will be given the opportunity to adapt a display to their own preferences, consumers avoid such adjustments (Rajae-Joordens and Heynderickx, 2008). Viewers encounter difficulties in multi-dimensional parameter spaces spanned by various display parameters, and as indicated in the 3-D Visual Experience Model, setting the preferred 3-D viewing experience requires adjustments in image quality, visual comfort and depth parameters.

A drawback of reducing screen disparity for the entire movie, however, is that it reduces the viewing experience of 3-D content. Fortunately, decreasing screen disparity is not necessary for the entire 3-D movie. We revealed that visual discomfort is built up by different video characteristics from which the contribution is scene-dependent. It can be sufficient to adjust the screen disparity only in scenes that have high degrees of 2-D and 3-D motion or to reduce screen disparity in the area where subtitles are present. Ideally, this adjustment should be related to the results of the WRRT-ratio, i.e., when viewers are categorized with a moderate binocular status via the WRRT-ratio, the adjustment in screen disparity should be larger. There are popular acknowledged formats for stereoscopic content (RGB + depth) that allow such processing. These results can also be used to guide 3-D production companies to create more comfortable videos, to limit depth when e.g. 2-D motion is high or to make viewers aware of the causes of potential uncomfortable effects.

BIBLIOGRAPHY

Aberson, D. H. A., and Bouwhuis, D. G. (1997). Silent Reading as Determined by Age and Visual Acuity. *Journal of Research in Reading*, 20:184-204.

Aldridge, R. P., Davidoff, J., Ghanbari, M., Hands, D. S., and Pearson, D. E. (1995). Recency effect in the subjective assessment of digitally-coded television pictures. In *Proceedings of the MOSAIC Workshop for Advanced Methods for the Evaluation of Television Picture Quality*, 49-56.

Aldridge, R. P., Hands, D. S., Pearson, D. E., and Lodge, N. K. (1998). Continuous quality assessment of digitally coded television pictures. *IEE Proceedings Vision Image and Signal Processing 145*, 2:116-123.

Allen, P. M., Radhakrishnan, H., and O'Leary, D. J. (2003). Repeatability and Validity of the PowerRefractor and the Nidek AR600-A in an Adualt Population with Healthy Eyes. *Optometry and Vision Science*, 80:245-251.

Barenbrug, B. (2006). 3-D Throughout the Video Chain. In *International Congress of Imaging Science'06*, 366-369.

Barry, S. R. (2009). *Fixing my gaze: a scientist's journey into seeing in three dimensions*, Basic Books: New York, USA.

Battiato, S., Curti, S., La Cascia, M., Tortora, M., and Scordato, E. (2004). Depth-map generation by image classification. In *Proceedings of the SPIE*, 5302: 95-104.

Beech, J. R., and Singleton, C. H. *The psychological assessment of reading*, First Ed. (Routledge, London, UK, 1997).

Benoit, A., Le Callet, P., Campisi, P., and Cousseau, R. (2008). Quality Assessment of Stereoscopic Images, In *EURASIP Journal of Image and Video Processing*, 2008:1-14.

Biocca, F., David, P., and West, M. 1994. Continuous response measurement (CRM): A computerized tool for research on the cognitive processing of communication messages. In Lang, A., editors, *Measuring psychological responses to media messages*, pages 15-64. Lawrence Erlbaum Associates, Inc., NJ.

Blake, R., and Logothesis, N. K. 2002. Visual competition. *Nature Reviews Neuroscience*, 3:13-21.

Blehm, C., Vishnu, S., Khattak, A., Mitra, S., and Yee, R. W. (2005). Computer Vision Syndrome: A Review. *Survey of Ophthalmology*, 50:253-262.

Blohm, W., Beldie, I. P., Schenke, K., Fazel, K., and Pastoor, S. (1997). Stereoscopic image representation with synthetic depth of field. *Journal of the Society for Information Display*, 5:307-313.

Borsting, E. J., Rouse, M. W., Mitchell, G. L., Scheiman, M., Cotter, S. A., Cooper, J., Kulp, M. T., and London, R. (2003). The Convergence Insufficiency Treatment Group, Validity and reliability of the revised convergence insufficiency symptom survey in children aged 9 to 18 years. *Optometry Vision Science*, 80:832-838.

Brautaset, R. L., and Jennings, J. A. M. (2006). Measurements of objective and subjective fixation disparity with and without a central fusion stimulus. *Medical Science Monitor*, 12:MT1-4.

Burge, J., Peterson, M. A., and Palmer, S. E. (2005). Ordinal configural cues combine with metric disparity in depth perception. *Journal of Vision*, 5:534-542.

Buzzelli, A. R. (1991). Stereopsis, Accommodative and vergence facility: Do they relate to dyslexia?. *Optometry and Vision Science*, 68:842-846.

Ciuffreda, K.J. (2002). The scientific basis for and efficacy of optometric vision therapy in nonstrabismic accommodative and vergence disorders. *Optometry*, 73:735-762.

Ciuffreda, M. A., Ciuffreda, K.J., and Wang, B. (2006). Repeatability and variability of near vergence ranges. *Journal of Behavioral Optometry*, 17:39-46.

Cole, R. G., and Boisvert, R. P. (1974). Effect of fixation disparity on stereo-acuity. *American Journal of Optometry and Physiological Optics*, 51:206-213.

Cooper, J. S., Burns, C. R., Cotter, S. A., Daum, K. M., Griffin, J. M., and Scheiman, M. M. (2001). Optometric clinical practice guideline care of the patient with accommodative and vergence dysfunction. St. Louis, America, American Optometric Association.

Cumming, G., and DeAngelis, G. C., (2001). The Physiology of Stereopsis. *Annual Reviews of Neuroscience*, 24:203-238.

Cutting, J., and Vishton, P. (1995). Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In Epstein, W. and Rogers, S., editors. *Perception of Space and Motion*, pages 69-117, Academic Press, San Diego, CA.

Dale, M. and Halgren, E. (2001). Spatiotemporal mapping of brain activity by integrationg of multiple imaging modalities. *Current opinion in Neurobiology*, 11:202-208.

de Haan, G. (2006). *Digital Video Post Processing*, University Press Eindhoven, Eindhoven.

de Ridder, H. (1996). Naturalness and Image Quality: Saturation and lightness variation in colour images. *Journal of Imaging Science and Technology*. 40:487-493.

de Ridder, H., Blommaert, F., and Fedorovskaya, E. (1995). Naturalness and image quality: Chroma and hue variations in colour images of natural scenes. *Proceedings of the SPIE*. 2411:51-56.

Dillon, T. W. and Emurian, H. H. (1996). Some factors affecting reports of visual fatigue resulting from use of a VDU. *Computers in Human Behavior*, 12:49-59.

Dodgson, N. A. (2005). Autostereoscopic 3-D display. *IEEE Computer*, 38:31-36.

Dodgson, N. A. (2004). Variation and extrema of human interpupillary distance. *Proceedings of the SPIE*, 5291:36-46.

Donovan, J. L., Brookes, S. T., Laidlaw, D. A. H., Hopper, C. D., Sparrow, J. M., and Peters, T. J. (2003). The development and validation of a questionnaire to assess visual symptoms/dysfunction and impact on quality of life in cataract patients: the Visual Symptoms and Quality of life (VSQ) Questionnaire. *Opthalmic Epidemiology*, 10:49-65.

Eadie, S., Gray, L. S., Carlin, P., and Mon-Williams, M., (2000). Modelling adaptation effects in vergence and accommodation after exposure to a simulated Virtual Reality stimulus. *Ophthalmic and Physiological Optics*, 20:242-251,

Ebenholtz, S. M. (2001). *Oculomotor Systems and Perception,* Cambridge University Press.

Edgar, G. K., (2007). Accommodation, cognition and virtual displays: A review of literature. *Displays*, 28:45-59.

Emoto, M., Niida, T., and Okana, F. (2005). Repeated Vergence Adaptation Causes the Decline of Visual Functions in Watching Stereoscopic Television. *Journal of Display Technology*, 1:328-340.

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

Engeldrum, P. G. (1999). Image Quality Modelling: Where Are We?. *IS&T's 1999 PICS Conference Proceedings*, 251-255.

Engeldrum, P. G. (2004). A theory of image quality: the image quality circle. *Journal of Imaging Science and Technology*, 48:447-457.

Evans, B. J. W. (2008). Optometric prescribing for decompensated heterophoria. *Optometry in Practise*, 9:63-78.

Evans, B. J. W. (2007). *Pickwell's Binocular Vision Anomalies*, Fifth edition, Oxford: Elsevier, UK.

Evans, B. J. W. (1997). *Pickwell's Binocular Vision Anomalies*, Third edition, Butterworth-Heinemann, Oxford.

Evans, B. J. W., Drasdo, N., and Richards, I. L. (1996). Dyslexia: the link with visual deficits. *Opthalmic and Physiological Optics*, 16:3-10.

Favalora, G. E., (2005). Volumetric 3-D Displays and Application Infrastructure. *IEEE Computer*, 38:37-44.

Fehn, C. (2004). Depth-Image-Based Rendering (DIBR), Compression and Transmission for a New Approach on 3-D-TV. *Proceedings of the SPIE*, 5291:93-104.

Fortuin, M. F., Lambooij, M. T. M., IJsselsteijn, W. A., Heynderickx, I., Edgar, D. F., and Evans, B. J. W. (2010). An exploration of the initial effects of stereoscopic displays on optometric parameters. *Ophthalmic and Physiological Optics*, 31:33-44.

Freeman, J., and Avons, S. E., (2000). Focus group exploration of presence through advanced broadcast services. *Proceedings of the SPIE*, 3959:530-539.

Gall, R., Wick, B., and Bedell, H. (1998). Vergence facility and target type. *Optometry and Vision Science*, 75:727–730.

Garcia, A., Cacho, P., Lara, F., and Megias, R. (2000). The relationship between accommodation facility and general binocular dysfunction. *Ophthalmic and Physiological Optics*, 20:98-104.

Gescheider, G. (1997). *Psychophysics: the Fundamentals*, Third edition, Lawrence Erlbaum Associates, London.

Goss, D. A., and Huifang, Z. (1994). Clinical and laboratory investigations of the relationship of accommodation and convergence function with refractive error. A literature review. *Documenta Ophthalmologica*, 86:349-380.

Hagura, H., Nakajima, M., Owaki, T., and Takeda, T., (2006). Study of asthenopia caused by the viewing of stereoscopic images: measurement by MEG and other devices. *Proceedings of the SPIE*, 6057:192-202.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (2006). *Multivariate Data Analysis*, Upper Saddle River, NJ: Pearson Prentice Hall, New York.

Häkkinen, J.(2004). Invited Paper: A Virtual Display for Mobile Use. *SID Symposium Digest of Technical Papers*, 35:1586-1589.

Häkkinen, J., Pölönen, M., Takatalo, J., and Nyman, G. (2006). Simulator Sickness in Virtual Display Gaming: A Comparison of Stereoscopic and Non-Stereoscopic Situations. *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, 159:227-230.

Häkkinen, J., Kawai, T., Takatalo, J., Leisti, T., Radun, J., Hirsaho, A., and Nyman, G. (2008). Measuring Stereoscopic Image Quality Experience with Interpretation Based Quality Methodology. *Proceedings of the SPIE*, 6808: 6808B.

Hamberg, R., and de Ridder, H. (1995). Continuous assessment of perceptual image quality. *Journal of the Optical Society of America*, 12:573–2577.

Hewage, C. T. E. R., Worral, S., Dogan, S., Kodikara Arachchi, H., and Kondoz, A. M. (2007). Stereoscopic TV over IP. *IEEE Proceedings of the IET European Conference on Visual Media Production*, 1-7.

Heynderickx, I. (2006). Performance evaluation of 2-D and 3-D-TV systems. *Proceedings of the Interbational Congress of Imaging Science*, 374-377.

Hillis, J. M., Ernst, M. O., Banks, M. S., and Landy, M. S. (2002). Combining Sensory Information: Mandatory Fusion Within, but Not Between Senses. *Science*, 298:1627-1630.

Hillis, J. M., Watt, S. J., Landy, M. S., and Banks, M.S. (2004). Slant from texture and disparity cues: Optimal cue combination. *Journal of Vision*, 4:967-992.

Hiruma, N., and Fukuda, T., (1993). Accommodation response to binocular stereoscopic TV images and their viewing conditions. *SMPTE journal*, 102:1137-1144.

Hoffman, D. M., Girshick, A. R., Akeley, K., and Banks, M. S., (2008). Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision*, 8:1-30.

Holliman, N. (2004). Mapping perceived depth to regions of interest in stereoscopic images. *Proceedings of the SPIE*, 5291:117-128.

Howard, P. (2002). *Seeing in Depth: Basic Mechanisms, vol 1*. I Porteous Publishing, Toronto.

Howard, P. and Rogers, B. J. (2002). *Seeing in Depth: Depth Perception vol. 2*. I Porteous Publishing, Toronto.

Howarth, P. A., (1996). Empirical studies of accommodation, convergence, and HMD use. *Proceedings of the Hoso-Bunka Foundation Symposium: The Human Factors in 3-D Imaging*.

Hung, G. K., (2001*). Models of Oculomotor Control*. World Scientific Publishing Co. Pte. Ltd, Singapore.

Hunt, O. A., Wolfssohn, J. S. and Gilmartin, B., (2002). Evaluation of the measurement of refractive error by the PowerRefractor: a remote, continuous and binocular measurement system of oculomotor funtion. *British Journal of Ophthalmology*, 87:1504-1508.

IJsselsteijn, W. A. (2004). *Presence in Depth*. Ph.D. dissertation, Eindhoven University of Technology, Eindhoven, The Netherlands.

IJsselsteijn, W. A., de Ridder., H., Avons, S. E., and Bouwhuis., D. (2001). Effects of stereoscopic presentation, image motion and screen size on subjective and objective corroborative measures of presence. *Presence*. 10:298-311.

IJsselsteijn, W., Bouwhuis, D., Freeman, F., and de Ridder, H. (2002). Presence as an experiential metric for 3-D display evaluation. *Proceedings of the SID*. 33:252-255.

IJsselsteijn, W. A., de Ridder, H., Hamberg, R., Bouwhuis, D., and Freeman., J. (1998a). Perceived depth and the feeling of presence in 3-DTV. *Displays*. 18:207-214.

IJsselsteijn, W. A., de Ridder, H., and Hamberg, R. (1998b). Perceptual factors in stereoscopic displays: The effect of stereoscopic filming parameters on perceived depth and reported eye-strain. *Proceedings of the SPIE*. 3299:282-291.

IJsselsteijn, W. A., Seuntiëns, P. H. J., and Meesters, L. M. J. (2005). Human Factors of 3-D Displays. In Scheer, O., Kauff, P., and Sikora. T., editors. *3-D Videocommunication - Algorithms, concepts and real-time systems in human-centred communication*, pages 219-234. John Wiley and Sons, Ltd.

Inoue, T., and Ohzu, H. (1997). Accommodative responses to stereoscopic three-dimensional display. *Applied Optics*, 36:4509-4515.

Iribarren, R., Fornaciari, A., and Hung, G. K., (2001). Effect of cumulative near work on accommodation facility and asthenopia, *International Ophthalmology*. 24:205-212.

ITU, (2000). Subjective Assessment for Stereoscopic Television Pictures, Recommendation. *Recommendation BT.1438.*

ITU-R, (2002). Methodology for the subjective assessment of the quality of television pictures. *Recommendation ITU-R BT.500-11.*

Iwasaki, T., Kubota, T., and Tawara, A., (2009). The tolerance range of binocular disparity on a 3-D display based on the physiological characteristics of ocular accommodation. *Displays*. 30:44-48.

Jeanes, R., Busby, A., Martin, J., Lewis, E., Stevenson, N., Pointon, D., and Wilkins, A.J. (1997). Prolonged use of coloured overlays for classroom reading. *The British Journal of Psychology* 88:531-548.

Jenkins, T. C. A., Pickwell., L. D., and Yekta, A. A. (1989). Criteria for decompensation in binocular vision. *Ophthalmic and Physiololical Optics*. 9:121-125.

Jones, G., Lee, D., Holliman, N., and Ezra, D. (2001). Controlling perceived depth in stereoscopic images. *Proceedings of the SPIE*. 4297:42-53.

Jones, B. L., and McManus, P. R. (1986). Graphic Scaling of Qualitative terms. *SMPTE Journal.* 95:1166-1171.

Jones, R., and Stephens, G. L. (1989). Horizontal fusional amplitude. *Investigative Ophthalmology and Visual Science*. 30:1638-1642.

Kanidodou, E. (2011). Amblyopia: a mini review of literature. *International Ophthalmology*. 31: 249-256.

Kaptein, R., and Heynderickx, I. (2007). Effect of Crosstalk in Multi-View Autostereoscopic 3-D Displays on Perceived Image Quality. *Journal of the SID*. 38:1220-1223.

Karania, R., and Evans, B. J. W. (2006). The Mallett Fixation Disparity Test: influence of test instructions & relationship with symptoms. *Ophthalmic and Physiological Optics*. 26: 507-52.

Kooi, F. L., and Toet, A. (2004). Visual comfort of binocular and 3-D displays. *Displays*. 25:99-108.

Kriss, I., and Evans, B. J. W. (2005). The relationship between dyslexia and the Meares-Irlen Syndrome, *Journal of Research in Reading*. 28:350-364.

Kozulin, P., Ames, S. L., and McBrien, N. A., (2009). Effects of a head-mounted display on the oculomotor system of children. *Optometry and Vision Science*. 86:845-856.

Kuze, J., and Ukai, K. (2008). Subjective evaluation of visual fatigue caused by motion images. *Displays*. 29:159-166.

Lambooij, M., Fortuin, M., IJsselsteijn, W., and Heynderickx, I., (2009). Measuring Visual Discomfort associated with 3-D displays. *Proceedings of the SPIE*. 7237:21.

Lambooij, M., IJsselsteijn, W., Bouwhuis, D., and Heyndericks, I. (2010). Evaluation of stereoscopic stills: Beyond 2-D quality. *IEEE Transactions on Broadcasting*. 57: 432-444.

Lambooij, M., IJsselsteijn, W. A., Fortuin, M., Evans, B. J. W., and Heynderickx, I. (2010). Measuring visual discomfort associated with 3-D displays. *Journal of Society of Information Displays*. 18:931-943.

Lambooij, M., IJsselsteijn, W.A., Fortuin, M., Evans, B. J. W., and Heynderickx, I. (2010). Susceptibility to visual discomfort of 3-D displays by visual performance measures, *IEEE Transactions on Circuits and Systems for Video Technology*: 21.

Lambooij, M., IJsselsteijn, W. A., Fortuin, M., and Heynderickx, I. (2009). Visual Discomfort and Visual Fatigue of Stereoscopic Displays: a Review. *Journal of Imaging Technology and Science.* 53:1-14.

Lambooij, M., IJsselsteijn, W. A., Fortuin, M., and Heynderickx, I., (2011). Impact of binocular vision characteristics on visual complaints resulting from watching stereoscopic content, *Displays: special issues 3-D environment*. submitted for publication.

Lambooij, M., IJsselsteijn, W. A., and Heynderickx, I. (2011). Visual discomfort of 3-D TV assessment methods and modelling. *Display: special issue image safety*. 32: 209-218.

Lambooij, M., Murdoch, M., IJsselsteijn, W. A., and Heynderickx, I. (2011). The impact of video characteristics and subtitles on the visual comfort of 3-D TV. *Displays*. submitted for publication.

Legge, G. E., Cheung, S., Yu, D., Chung, S. T. L., Lee, H., and Owens, D. P. (2007). The case for the visual span as a sensory bottleneck in reading. *Journal of Vision*. 7:1-15.

Lessiter, J., Freeman, J., Keogh, E., and Davidoff, J. (2001). Development of a new cross-media presence questionnaire: The ITC-sense of presence inventory. *Presence Teleoperators and Virtual Environments*. 10:282-297.

Li, H., Seo, J., Kham, K., and Lee, S. (2008). Measurement of 3-D visual fatigue using event-related potential (ERP): 3-D oddball paradigm. In *3-DTV Conference: The True Vision - Capture, Transmission and Display of 3-D Video.*pages 213-216.

Mangione, C., Lee, P. P., Gutierrez, P. R., Spritzer, K., Berry, S., and Hays, R. D., (2001). The National Eye Institute Visual Function Questionnaire Field Test Investigators, Development of the 25-item National Eye Institute Visual Function Questionnaire. *Archives of Ophthalmology*. 119:1050-1058.

Mark, W. R., McMillan, L., and Bishop, G. (1997). Post-rendering 3-D warping. *Proceedings 1997 Symposium on Interactive 3-D Graphics*, pages 7-16.

Mather, G., and Smith, D. R. R., (2000). Depth cue integration: stereopsis and image blur. *Vision Research*. 40:3501-3506.

Matthews, M. L., Lovasik, J. V., and Mertins, K. (1989). Visual performance and subjective discomfort in prolonged viewing of chromatic displays. *Human Factors*. 31:259-71.

McKay, H. C. (1953). *Three-Dimensional Photography: Principles of Stereoscopy*. American Photography Book Department.

Meesters, L. M. J., IJsselsteijn, W. A., and Seuntiëns, P. J. H. (2004). A survey of perceptual evaluations and requirements of three-dimensional TV. *IEEE Transactions on Circuits and Systems for Video Technology*. 14:381-391.

Melville, A. C., and Firth, A. Y. (2002). Is there a relationship between prism fusion range and vergence facility?. *British Orthoptic Journal*. 59:38-44.

Mendiburu, B. (1997). *3-D Movie Making: Stereoscopic Digital Cinema from Script to Screen*, First edition. Focal Press, Burlington, USA.

Mon-Williams, M., Wann, J. P., and Rushton, S. (1993). Binocular vision in a virtual world: Visual deficits following the wearing of a head-mounted display. *Ophthalmic and Physiological Optics*. 13:387-391.

Montes-Mico, R. (2001). Prevalence of general dysfunctions in binocular vision. *Annals of Ophthalmology*. 33:205-208.

Murata, A., Uetake, A., Otsuka, W., and Takasawa, Y. (2001). Proposal of an index to evaluate visual fatigue induced during visual display terminal tasks. *International Journal of Human Computer Interaction*. 13:305-321.

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

Nichols, M. J., and Newsome, W. T., (1999). The neurobiology of cognition. *Nature*. 402:C35-C38.

Nojiri, Y., Yamanoue, H., Hanazato, A., and Okana, F. (2003). Measurement of parallax distribution, and its application to the analysis of visual comfort for stereoscopic HDTV. *Proceedings of the SPIE*. 5006:195-205.

Nojiri, Y., Yamanoue, H., Ide, S., Yano, S., and Okana, F. (2006). Parallax distribution and visual comfort on stereoscopic HDTV. *Proceedings of the IBC*. pages 373-380.

Okada, Y., Ukai, K., Wolffsohn, J. S., Gilmartin, B., Iijima, A., and Bando, T. (2006). Target spatial frequency determines the response to conflicting defocus- and convergence-driven accommodative stimuli. *Vision Research*. 46:475-484.

O'Leary, C. I., and Evans, B. J. W. (2006). Double-masked randomised placebo controlled trial of the effect of prismatic corrections on the rate of reading and the relationship with symptoms. *Ophthalmic and Physiological Optics*. 26:555-565.

Ostrin, L. A., and Glasser, A., (2004). Accommodation measurements in a prepresbyopic and presbyopic population. *Journal of Cataract and Refractive Surgery*. 30:1435-1444.

Owens, D. A., and Wolf-Kelly, K. (1987). Near work, visual fatigue and variations of oculomotor tonus. *Investigative Ophthalmology and Visual Science*. 28:743-749.

Panel on Impact of Video Viewing on Vision Workers. (1983). *Video Displays, Work and Vision. Committee on Vision*, National Research Council, Washington, DC: National Academy Press.

Parker, J. (2007). Binocular depth perception and the cerebral cortex. *Nature reviews*. 8:379-391.

Sazzad, P., Yamanaka, S., and Horita, Y. (2010). Spatio-temporal segmentation based continuous no-reference stereoscopic video quality prediction. *Proceedings of Quality of Multimedia Experience*. pages 106-111.

Pastoor, S. (1993). Human factors of 3-D displays in advanced image communications. *Displays*. 14:150-157.

Pastoor, S., and Wöpking, M. (1997). 3-D displays: A review of current technologies. *Displays*. 17:100-110.

Patterson, R., and Fox, R. (1984). The effect of testing method on stereoanomaly. *Vision Research*. 24:403-408.

Patterson, R. and Martin, W. L. (1992). Human Stereopsis. *Human Factors*. 34:669-692.

Patterson, R. (2009). Review Paper: Human factors of stereo displays: An update. *Journal of Society of Information Displays*. 17:987-996.

Peli, E. (1998). The visual effects of head-mounted display (HMD) are not distinguishable from those of desk-top computer display. *Vision Research*. 38:2053-2066.

Peli, E. (1999). Optometric and perceptual issues with head-mounted display. In Mouroulis, P., editors. *Optical Design for Visual Instrumentation*. pages 205-276. McGraw-Hill, New York.

Pickwell, L. D., Kaye, N. A., and Jenkins. T. C. A. (1991). Distance and near readings of associated heterophoria taken on 500 patients. *Ophthalmic and Physiological Optics*. 11:291-296.

Pockett, L., Pölönen, M., Salmimaa, M. (2009). Subtitles in short 3-D movies. *Society of Information Display*. 40:38-41.

Pölönen, M., Salmimaa, M., Alltonen, V., Häkkinen, J. and Takatalo, J. (2009). Subjective measures of presence and discomfort in viewers of colour-separation-based stereoscopic cinema. *Journal of the Society of Information Displays*. 17:459-466.

Pouratian, N., Sheth, S. A., Martin, N. A., and Toga, A. W. (2003). Shedding light on brain mapping: advances in human optical imaging. *Trends in Neuroscience*. 26:277-282.

Rajae-Joordens, R. J. E., and Heynderickx, I. (2009). Multi-Dimensional Optimization of Visual Content. In *Visual Perception: New Research*. Nova Science Publishers, Inc, New York, pages 117-155,

Redert, A., Berretty, R. P., Varekamp, C., van Geest, B., and Bruijns, J. (2007). Challenges in 3-DTV image processing. *Proceedings of the SPIE*. 6490:14.

Reichelt, S., Häussler, R., Fütterer, G., and Leister, N., (2010). Depth cues in human visual perception and their realization in 3-D displays. *Proceedings of the SPIE*. 7690:76900B.

Richards, W. (1970). Stereopsis and Stereoblindness. *Experimental Brian Research*. 10:380-388.

Richman, J. E., and Laudon, R. C. (2002). A survey of the prevalence of binocular % accommodative dysfunctions in a sample of optometry students. *Journal of Behavioral Optometry*. 13:31-33.

Rotello, C. M., Masson, M. E. J., and Verde M. F. (2008). Type I Error Rates and Power Analyses for Single-Point Sensitivity Measures. *Attention, Perception & Psychophysics*. 70:389-401.

Rushton, S. K., and Riddell, P. M. (1999). Developing visual systems and exposure to virtual reality and stereo displays: some concerns and speculations about the demands on accommodation and vergence. *Applied Ergonomics*. 30:69-78.

Sammons, E. (1934). *The World of 3-D Movies*. Van Nostrand Reinhold, New York.

Scheiman, M. (1996). Accommodative and binocular vision disorders associated with video display terminals: diagnosis and management issues. *Journal of the American Optometric Association*. 67:531-539.

Semmlow, J. J., and Heerema, D. (1979). The role of accommodative convergence at the limit of fusional vergence. *Investigative Ophthalmology and Visual Science*. 18:970-976.

Scheiman, M., and Wick, B. (1994). *Clincal Management of Binocular Vision, Heterophoric, Accommodative, and Eye Movement disorders*. J.B. Lippincott Company, Pennsylvania.

Schiffman, R. M., Christianson, M. D., Jacobsen, G., Hirsch, J. D., Reis, B. L. (2000). Reliability and validity of the ocular surface disease index. *Archives of Ophthalmology*. 118:615-621.

Schor, C. M., and Kotulak, J. C. (1986). Dynamic interactions between accommodation and convergence are velocity sensitive. *Vision Research*. 26:927-942.

Schor, C., Wood, I., and Ogawa, J. (1984). Binocular sensory fusion is limited by spatial resolution. *Vision Research*. 24:661-665.

Seuntiëns, P. J. H., Meesters, L. M. J., and IJsselsteijn, W. A. (2005). Perceptual attributes of crosstalk in 3-D images. *Displays*. 26:177-183.

Seuntiëns, P. J. H., Meesters, L. M. J., and IJsselsteijn. W. A. (2006). Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation. *ACM Transactions on Applied Perception*. 3:95-109.

Seuntiëns, P. J. H., Heynderickx, I., and IJsselsteijn, W. A. (2008). Capturing the added value of 3-D-TV: Viewing experience and naturalness of stereoscopic images. *Journal of Imaging Science and Technology*. 52:1-5.

Sexton, I., and Surman, P. (1999). Stereoscopic and autostereoscopic displays systems. *IEEE Signal Processing Magazine*. 16:85–99.

Shade, J., Gortler, S., He, L., and Szeliski, R. (1998). Layered depth images. *Proceedings of SIGGRAPH 1998*. pages 231-242.

Sheard, C. (1934). The prescription of prism. *American Journal of Optometry*. 11:364-378.

Sheedy, J. E., Hayes, J., and Engle, J. (2003). Is all asthenopia the same?. *Optometry and vision science*. 80:732-739.

Sheedy, J. E., and Parsons, S. D. (1990). The video display terminal eye clinic: clinical report. *Optometry and Vision Science*. 67:622-626.

Sheedy, J. E., and Saladin, J. J. (1977). Phoria, vergence, and fixation disparity in oculomotor problems. *Am J Optom Physiol Opt*. 54:474-478.

Sheedy, J. E., and Saladin, J. J. (1978). Association of symptoms with measures of oculomotor deficiencies, *Am J Optom Physiol Opt*. 55:670-676.

Sheedy, J. E., and Saladin, J. J. (1983). Validity of diagnostic criteria and case analysis in binocular vision disorders. In: Schor CM, Ciuffreda KJ, editors, *Vergence Eye Movements: Basic and Clinical Aspects*. pages 517–40. Boston: Butterworth.

Shy, M. E., Frohman, E. M., So, Y. T., Arezzo, J. C., Cornblath, D. R., Giuliani, M. J., Kincaid, J. C., Ochoa, J. L., Parry, G. J., and Weimer, L. H. (2003). Quantitative sensory testing: report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. *Neurology*. 60:898-904.

Siegel, M. (2001). Perceptions of crosstalk and the possibility of a zoneless autostereoscopic display. *Proceedings of the SPIE*. 4297:34-41.

Smith, G., and Atchinson, D. A. (1997). *The Eye and Visual Optical Instruments*. Cambridge University Press.

Speranza, F., Tam, W. J., Renaud, R., and Hur, N. (2006). Effect of disparity and motion on visual comfort of stereoscopic images. *Proceedings of the SPIE*. 6055:60550B.

Strohmeier, D., and Tech, G. (2010). "Sharp, bright, three-dimensional" – open profiling of quality for mobile 3-DTV coding methods. *Proceedings of the SPIE*. 7542:75420T.

Steinman, S. B., Steinman, B. A., and Garzia, R. P. (2000). *Foundations of binocular vision*, First edition. McGraw-Hill Companies, New York, USA.

Stidwill, D. (1990). *Orthoptic assessment and management*, Blackwell Scientific Publications, Oxford.

Stidwill, D. (1997). Clinical survey: epidemiology of strabismus. *Ophthalmic and Physiological Optics*. 17:536-539.

Sullivan, J. M. (2008). *Visual fatigue and the driver*. Report No. UMTRI-2008-50, University of Michigan Transportation Research Institute.

Surman, P., Sexton, I., Bates, R., Lee, W. K., Craven, M., and Yow, K C. (2003). Beyond 3-D television: The multi-modal, multi-viewer TV system of the future. *Proceedings SID/SPIE FLOWERS 2003*. pages 208-210.

Suryakumar, R., and Bobier, W. R. (2004). Gain and Movement Time of Convergence-Accommodation in Preschool Children. *Optometry and Vision Science*. 81:E835.

Suryakumar, R. Meyers, J. P., Erving, E. L., and Bobier, W. R. (2007). Application of video-based technology for the simultaneous measurement of accommodation and vergence. *Vision Research*. 47:260-268.

Suryakumar, R., Meyers, J. P., Erving, E. L., and Bobier, W. R. (2007). Vergence accommodation and monocular closed loop blur accommodation have similar dynamic charactreristics. *Vision Research*. 47:327-337.

Suyame, S., Ishigurw, Y., Takada, H. Nakazawa, K., Hosohata, J., Takao, Y., and Fujikado, T. (2005). Evaluation of Visual Fatigue in Viewing a Depth-fused 3-D Display in Comparison with a 2-D Display. *NTT Techical Review*. 3: 82-89.

Takeda, T., Hashimoto, K., Hiruma, N., and Fukui, Y. (1999). Characteristics of accommodation toward apparent depth. *Vision Research*. 39:2087-2097.

Talmi, K., and Liu, J. (1999). Eye and gaze tracking for visually controlled interactive stereoscopic displays. *Signal Processing: Image Communication*. 14:799-810.

Tam, W., Stelmach, L., and Corriveau, P. (1998). Psychovisual aspects of viewing stereoscopic video sequences. *Proceedings of the SPIE*. 3295:226-235.

Tam, W. A., Vázquez, C., and Speranza, F. (2007). Surrogate depth maps for stereoscopic imagaing: different edges typs. *Proceedings of the SPIE*. 6490:6490C.

Torgerson, W. (1967). *Theory and methods of Scaling*. New York: John Wiley & Sons.

Tsukuda, S., and Murai, Y., (1988). A case report of manifest esotropia after viewing anaglyph stereoscopic movie. *Japanese Orthoptic Journal*. 18:69–72.

Ukai, K., and Howarth, P. A. (2008). Visual fatigue caused by viewing sterescopic motion images: Background, theories and observations. *Displays*. 29:106-116.

Ukai, K., and Kato, Y. (2002). The use of video refraction to measure the dynamics properties of the near triad in observers of a 3-D display. *Opthalmic and Physiological Optics*. 22:385- 388.

Ukwade, M.T, (2000). Effects of nonius line and fusion lock parameters on fixation disparity. *Optometry and Vision Science*. **77**: 309-320.

Ukwade, M. T., Bedell, H. E., and Harwerth, R. S. (2003). Stereopsis is perturbed by vergence error. *Vision Research*. 43:181-193.

van Berkel, C., and Clarke, J. (1997). Characterization and optimization of 3-D-LCD module design. *Proceedings of the SPIE*. 3012:179-187.

van Nes, F. L. (2009). Visually representing reality; aesthetics and accessibility aspects. *Proceedings of the SPIE*. 7240:7240L.

Wandell, A., and Dougherty, R. F. (2006). Computational neuroimaging: maps and tracks in the human brain. *Proceedings of the SPIE*. 6057:605701.

Wang, B. W., and Ciuffreda, K. J. (2006). Depth-of-focus of the Human Eye: Theory and Clinical Implications. *Survey of Ophthalmology*. 51:75-85.

Wang, L., Tu, Y., Chen, L., Zhang, P., Teunissen, K., and Heynderickx, I. (2010). Cross-talk acceptability in natural still images for different (auto) stereoscopic display technologies. *Journal of the SID*. 18:405-414.

Wann, J. P., and Mon-Williams, M. (1997). Health issues with virtual reality displays: what we do know and what we don't. *Computer graphics*. 31:53-57.

Wann, J. P., Rushton, S., and Mon-Williams, M. (1995). Natural problems for stereoscopic depth perception in virtual environments. *Vision Research*, 35:2731-2736.

Ware, C., Gobrecht, C., and Paton, M. (1998). Dynamic adjustment of stereo display parameters. *IEEE transactions on Systems, Man and Cybernetics*. 7:225-240.

Watten, R. G. (1994). Reinvention of visual fatigue: accumulation of scientific knowledge or neglect of scientific history?. *Ophthalmic and Physiological Optics*. 14:428-432.

Westheimer, G. (1994). The Ferrier Lecture, 1992. Seeing depth with two eyes: stereopsis. *Proceedings of Royaal Society of London: Biological Science*. 257:205-219.

Wheatstone, C. (1838). On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London*. 128:371-394.

Wilkins, A. J. (2002). Coloured overlays and their effects on reading speed: a review. *Ophthalmic and Physiological Optics*. 22:448-454.

Wilkins, A. J, Huang, J., and Cao, Y. (2004). Visual stress theory and its application to reading and reading tests. *Journal of Research in Reading*. 27:152-162.

Wilkins, A. J., Jeanes, R. J., Pumfrey, P. D., and Laskier, M. (1996). Rate of Reading Test: its reliability, and its validity in the assessment of the effects of coloured overlays. *Ophthalmic and Physiological Optics*. 16:491-497.

Williams, C., Northstone, K., Howard, M., Harvey, I., Harrad, R. A., and Sparrow, J. M. (2008). Prevalence and risk factors for common vision problems in children: data from the ALSPAC study. *British Journal of Ophthalmology*. 92:959-964.

Witmer, B. and Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments*. 7:225-240.

Woods, A., Docherty, T., and Koch, R. (1993). Image Distortions in Stereoscopic Video Systems. *Proceedings of the SPIE*. 1915:36-49.

Wopking, M. (1995). 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*. 3:101-103.

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

Yano, S., Ide, S., Mitsuhashi, T., and Thwaites, H. (2002). A study of visual fatigue and visual comfort for 3-D HDTV/HDTV images. *Displays*. 23:191-201.

Yeh, Y. Y., and Silverstein, L. D. (1990). Limits of fusion and depth judgement in stereoscopic colour displays. *Human Factors*. 32:45-60.

Yekta, A. A., Jenkins, T., and Pickwell, D. (1987). The clinical assessment of binocular vision before and after a working day. *Ophthal Physiol Opt*. 7: 349-352.

Yendrikhovskij, S. N. (1998). *Colour reproduction and the naturalness constraint*. Ph.D. dissertation, Eindhoven University of Technology, The Netherlands.

You, J., Xing, L., Perkins, A., and Wang, X. (2010). Perceptual quality assessment for stereoscopic images based on 2-D image quality metrics and disparity analysis, *Proceedings of the VPQM*. 9.

Zellers, J. A., Alpert, T. L., and Rouse, M. W. (1984). A review of the literature and a normative study of accommodation facility. *Journal of the American Optometric Association*. 55:31-74.

Summary

The embracing of 3-D movies by Hollywood and fast LCD panels finally enable the home consumer market to start successful campaigns to get 3-D movies and games in the comfort of the living room. By introducing three-dimensional television (3-D TV) and its desktop-counterpart for gaming and internet applications on the public consumer market, viewers will be provided with a whole new experience. The difference between 3-D TV and its predecessor is the introduction of binocular disparity, i.e., the fact that the left and the right eye receive a slightly horizontally shifted perspective of the same scene, from which the brain extracts depth information. As a consequence, the viewer perceives the image as if its content is positioned in three-dimensional space, i.e., both in front of and behind the television screen. Central to these developments are be the viewer's experiences which will signify the success or failure of proposed innovative imaging technology, i.e., both perceived image quality and viewing comfort should be at least comparable to conventional television. The aim of this thesis is therefore to understand, measure and eventually, model and predict the added value of stereoscopic depth as well as the accompanying visual discomfort associated with 3-D TV.

With respect to the added value of stereoscopic depth, a 3-D Quality Model as an extension of Engeldrum's Image Quality Circle is proposed in Chapter 2, since many of our results confirm that the added value of stereoscopic depth is not captured with the Image Quality Circle. The 3-D Quality Model describes higher level evaluation metrics as a weighted sum of image quality and depth. Two higher level evaluation metrics are investigated; naturalness and viewing experience. In experiments 1, 2 and 3 image quality (levels of noise and blur) and stereoscopic depth (different camera base distances and levels of screen disparities) are varied and evaluated in terms of image quality, depth, viewing experience and naturalness. The results show that perceived image quality and perceived depth are not independent in their relationship to physical image characteristics, but are on the perceptual level. Variations in image quality are reflected by viewing experience and naturalness to a similar extent, yet the added value of stereoscopic depth is more incorporated in naturalness. Naturalness is most appropriate to evaluate the added value of 3-D quality of stereoscopic stills, since it weights stereoscopic depth most in addition to image quality. An important aspect is that these results are consistent over a range of different 3-D displays and content generation methods (thereby reflecting different depth percepts) and image quality attributes. Hence, we have shown that the 3-D Quality Model based on naturalness as evaluation metric is validly applicable to stereoscopic stills and that its value is determined for approximately 75% by image quality and for approximately 25% by the added value of stereoscopic depth.

With respect the accompanying visual discomfort associated with 3-D TV a first step is to arrive at a full understanding of visual discomfort, its determinants and contributing factors, and the measurable effects it has on viewers' visual functioning and subjective experience. In line with this, a theoretical frame work is presented in Chapter 3 on which our research concerning the visual discomfort is centered and that reduces the ambiguity of the visual discomfort associated with stereoscopic display technology and image generation.

Our general recommendation is to adhere to a 'one degree of disparity' limit. This limit allows for sufficient depth rendering for most application purposes and should guarantee a comfortable viewing in stereoscopic television. This range is based on human vision characteristics; screen disparities can be fused and the buffer in our accommodation-vergence system accounts for mismatches. Within this zone of comfortable viewing, visual discomfort can still occur to an extent, however, which is likely to be caused by one or more of the following three factors: excessive demand of accommodation-convergence linkage; 3-D artifacts; and unnatural amounts of blur.

Chapter 3 also describes potential measurement methods to evaluate the visual discomfort associated with stereoscopic viewing. We define visual fatigue as physiological strain or stress resulting from excessive exertion of the visual system, which can be objectively measured and visual discomfort as its subjective counterpart. To adequately characterize and determine the degree of visual fatigue and visual discomfort in a sensitive, accurate, reliable and valid way, multiple indicators for both components can be relevant. Visual discomfort can be evaluated with validated questionnaires or other self-report measures. For visual fatigue clinical measurement methods are best suited since they are relatively cheap, concise, quantitative with a high sensitivity and specificity and applicable to a large group of participants.

In Chapter 4 knowledge is gained on how visual discomfort is built up whilst watching stereoscopic content within the 'one degree of disparity' limit. The aim was to determine which video characteristics, e.g., lateral object and camera motion and (changes in) disparity, induce visual discomfort. The values of such video characteristics can be extracted from stereoscopic movies with motion and depth estimation algorithms and directly related to a continuous assessment (CA) in terms of visual discomfort. Hence, a CA of a long-term stereoscopic movie in terms of visual comfort can provide valuable moment-to-moment information concerning the perceptual impact of specific video characteristics.

Two experiments (experiments 4 and 5) were designed with the primary objective of relating the impact of these video characteristics to the assessment of visual discomfort. In addition, experiment 4 compares a continuous assessment method with other assessment methods of visual comfort, and experiment 5 investigates the impact of subtitles on visual comfort. Three 3-D movies were assessed in terms of visual comfort via a continuous assessment. The continuous assessment scores were directly compared to video characteristics that were derived from the 3-D movies. Additional assessment methods in experiment 4 included the assessment of six 10-second sequences captured from the 3-D movie and a single retrospective assessment of the entire 3-D movie. The two 3-D movies in experiment 5 were shown with or without subtitles.

Results show that the visual comfort of stereoscopic scenes can be predicted as a linear combination of screen disparity range and offset, changing screen disparity, and lateral motion. The specific contributions of these characteristics depend on the scene, yet more complex models are required to extend the comfort prediction to entire movies, incorporating different scenes. In addition, the results of experiment 5 reveal that subtitles required additional effort to keep vision comfortable and the results of experiment 4 show that the correlation between the assessment of the 10-second sequences captured from the 3-D movie and their corresponding parts within the continuous assessment is low, whereas the correlation between the retrospective assessment and the mean of the continuous assessment score over scene parts with a high screen disparity is higher.

The aim of Chapter 5 is to construct a measurement protocol for the evaluation of objective signs of visual fatigue and subjective symptoms of visual comfort. It is known that some people have a sensitive binocular vision and their visual system is challenged too much under unnatural viewing conditions. Consequently, they may experience visual discomfort regardless the absence of visual discomfort factors in the 3-D content. To adequately characterize the degree of objective visual fatigue and subjective visual discomfort, a measurement protocol was constructed that also incorporated differences in performance of the binocular system.

In experiment 6 the categorization of participants' binocular status is accomplished based on an optometric algorithm that reflects the values of ten single optometric tests; one subgroup with a good binocular status and one with a moderate binocular status. The main objective of experiment 6 is to identify the most appropriate optometric indicators for visual discomfort and visual fatigue. Objective and subjective optometric tests were applied before and after short-term stressful stereoscopic reading tasks both in 2-D and 3-D. The reading tasks were different passages of the Wilkins Rate of Reading test (an optometric test). Results reveal that participants with a moderate binocular status are more susceptible to visual discomfort associated with stereoscopic content based on objective signs of visual fatigue in the fusion range and the subjective symptoms of visual discomfort. In addition, the results reveal the number of words read in 3-D relative to number of words read in 2-D, referred to as WRRT-ratio, is indicative of the binocular status of people.

In experiment 7 the aim is to confirm that the WRRT-ratio is a proper indicator of the binocular status of people and thus of their susceptibility to visual discomfort. Similar as in experiment 1, the categorization of participants' binocular status is accomplished based on the optometric algorithm. The results reveal that even though the specific WRRT-ratio depends on viewing conditions, it is consistent with susceptibility: participants that have a moderate binocular status have poorer reading performance in 3-D than in 2-D and experienced more visual discomfort compared to people with normal binocular vision.

In experiment 8 the objective is to determine whether the threshold in screen disparity for visual discomfort is related to the binocular status of people. The categorization of participants' binocular status is accomplished based on the WRRT-ratio and objective and subjective evaluation is performed by the fusion range and questionnaires respectively. The results show that only participants with a moderate binocular status reveal a tendency in changed fusion range. These participants also indicate significantly more visual discomfort in stereoscopic conditions and set lower thresholds in screen disparity for visual discomfort than participants with a good binocular status. Hence, a combination of fusion range measurements and self-report is appropriate for evaluating visual complaints. The results also show that the WRRT-ratio is indicative of the binocular vision of people. This WRRT-ratio is very promising, since 1) it can be used to explain inter-subject differences in results, 2) it provides additional information to stereo- and visual acuity tests, and 3) it can be implemented in 3D commercial displays to advise people with moderate binocular vision to lower the screen disparity range in order to avoid visual complaints.

Chapter 5 briefly look back at the previous chapters by summarizing and discussing the most important findings and presents the applicability of the main findings as well as a future view and research that is left for future research to perform.

ACKNOWLEDGEMENTS

The journey towards the Ph. D. degree sometimes felt like walking the yellow brick road, of which the end would have been an illusion at most, if it were not for the help of many people.

First of all, I would like to thank Ingrid Heynderickx for being my first promotor and my daily supervisor. Ingrid, I could not have asked for a better supervisor in many ways. I can honestly say that with your encouragement and positive attitude the yellow brick road was much easier to follow (and occasionally easier to find back). I hope I inherited your way of approaching and tackling perceptual-technical oriented problems. Wijnand IJsselsteijn, my copromotor, I would like to thank for giving me direction with your interesting and innovative ideas. I do hope that your ability of writing poetic journal papers is at least partly passed on to me. I would like to thank Don Bouwhuis, my second promotor, for his fresh and new view on my research. Though we mainly worked together in last year, your hand is clearly visible in my work. In addition I would like to thank Marten Fortuin for a pleasant cooperation. A large part of this research could not have been accomplished without your contribution. But not only did Ingrid, Wijnand, Don and Marten made many contributions to this work, they were also very pleasant to work with, which is at least equally important in my opinion.

Furthermore, I would like to thank my roommate Maurice van Beurden and other colleagues. Not in the last place for their willingness to help and support me, but also for making the labour ambiance a pleasant one and making me look forward to another day at the job (a Ph.D. project includes more than just work). I do ask myself how many pages my thesis would count if it was not for the HTI Dreamteam!?

And last but certainly not least, special thanks goes out to my mother for her unconditional support during my ‘career’ so far. Even though I had to make my own lunch in the morning, without her, none of my accomplishments would have had been possible. Looking back at my childhood, I think no one in those years believed that that little boy, unmanageable as I was, would be where I am today. Except for one: Mom, thank you for always keeping the trust in me. It is to you I dedicate this book.

CURRICULUM VITAE

Marc T. M. Lambooij was born in Eindhoven, The Netherlands, on January 13, 1981. He attended the Van Maerlantlyceum in Eindhoven from 1993 to 1998, where he obtained a HAVO diploma in 1998. From 1998 to 2002, he studied Technical Engineering at the Physical Science department at the Fontys Hogescholen. During his graduation project of his bachelor in Medical Science he developed an algorithm to cluster epileptic brain activity based on specific brain signals. From 2002 to 2005 he completed the master in Human Technology Interaction at the faculty Industrial Engineering and Innovation Sciences at the Eindhoven University of Technology. After a well deserved world trip in which he visited South-East Asia, Australia, New Zealand, Fiji and the United States, he started his PhD. This was a collaborative research project between the Human-Technology Interaction Group of the Eindhoven University of Technology and Philips Research, the Netherlands. His PhD research focused on positive and negative visual perceptions associated with 3-D-display systems from a human- and a technological perspective.

Till this moment, he is very active in the field of 3-D research, has edited various journal issues related to this topic and his research reached the national and international newspapers. He is continuously interested in approaching the impact of technological innovations on our perception from a multidisciplinary viewpoint. His research is focused on performing and supervising application- and fundamental oriented research, in which the measurement and modelling of the impact of innovative media technologies on human experiences are central. At this moment he works for Philips Consumer Lifestyle as research scientist in the Viewing Experience Innovation Group.

LIST OF PUBLICATIONS

Published

1. Lambooij, M., IJsselsteijn, W.A., Fortuin, M., Evans, B. J. W., and Heynderickx, I. (2010). Susceptibility to visual discomfort of 3-D displays by visual performance measures. *IEEE Transactions on Circuits and Systems for Video Technology*. 21.
2. Lambooij, M., IJsselsteijn, W., Bouwhuis, D., and Heyndericks, I. (2010). Evaluation of stereoscopic stills: Beyond 2-D quality. *IEEE Transactions on Broadcasting*. 57: 432-444.
3. Lambooij, M., IJsselsteijn, W. A., and Heynderickx, I. (2011). Visual discomfort of 3-D TV assessment methods and modelling. *Display: special issue image safety*. 32: 209-218.
4. Fortuin, M., Lambooij, M., IJsselsteijn, W., Heynderickx, I., Edgar, D., and Evans, B. J. W. (2011). An exploration of the effects of stereoscopic displays on optometric parameters and asthenopic symptoms. *Journal of Ophthal. Physiol.* Opt. 31: 33-44.
5. Lambooij, M., Fortuin, M., IJsselsteijn, W. A., Evans, B. J. W., and Heynderickx, I. (2010). Measuring visual discomfort associated with 3-D displays. *Journal of Society of Information Displays.* 18:931-943.
6. Lambooij, M., Fortuin, M., IJsselsteijn, W. A., and Heynderickx, I. (2010). Visual discomfort associated with 3-D display. *VPQM*. 44.
7. Lambooij, M., IJsselsteijn, W. A., Fortuin, M., and Heynderickx, I. (2009). Visual Discomfort and Visual Fatigue of Stereoscopic Displays: a Review. *Journal of Imaging Technology and Science.* 53:1-14.
8. Kuijsters, A., IJsselsteijn, W., Lambooij, M., and Heynderickx, I. (2009). Influence of chroma variations on Naturalness and Image Quality of Stereoscopic Images. *Proceedings of the SPIE.* 7240:85.
9. Lambooij, M., Fortuin, M., IJsselsteijn, W., and Heynderickx, I. (2009). Measuring Visual Discomfort associated with 3-D displays. *Proceedings of the SPIE.* 7237:21.
10. Lambooij, M., Fortuin, M., IJsselsteijn, W., and Heynderickx, I. (2009). Measuring Visual Discomfort associated with 3-D displays. *Illumination & Displays: SPIE Newsroom.*
11. Lambooij, M., IJsselsteijn, W., and Heynderickx, I. (2009). Continuous assessment of visual discomfort in 3-DTV: towards a model prediction based on spatial image parameters. presented at the VIMS 2009.
12. IJsselsteijn, W., Lambooij, M., and Heynderickx, I. (2009). Assessing 3-D Displays: Naturalness as a Visual Quality Metric. Presented at 3-D Media Workshop, Berlin, 15 - 16 October.
13. Kaptein, R., Kuijsters, A., Lambooij, M., IJsselsteijn, W., and Heynderickx, I. (2008). Performance evaluation of 3-D-TV systems. *Proceedings of SPIE.* 6808:19.
14. Lambooij, M., Heynderickx, I., and IJsselsteijn, W. (2008). Evaluation of stereoscopic stills: beyond the 2-D quality model. *Perception.* 37:101.

15. Lambooij, M., IJsselsteijn, W., and. Heynderickx, I. (2007). Visual discomfort in stereoscopic displays: a review. *Proceedings of SPIE*. 6490:I.
16. Lambooij, M., IJsselsteijn, W., and Heynderickx, I. (2007). Stereoscopic displays and visual comfort: a review. *Illumination & Displays: SPIE Newsroom*.
17. Seuntiëns, P., Heynderickx, I., IJsselsteijn, W., van den Avoort, P., Berentsen, J., Dalm, I., Lambooij, M., and Oosting, W. (2005). Viewing experience and naturalness of 3-D images. *Proceedings of SPIE*. 6016.

Under submission:

1. Lambooij, M., IJsselsteijn, W. A., Fortuin, M., and Heynderickx, I., (2011). Reading performance as screening tool for visual complaints from stereoscopic content. *Displays: special issues 3-D environment*. submitted for publication.
2. Lambooij, M., Murdoch, M., IJsselsteijn, W., and Heynderickx, I. (2011). The impact of video characteristics and subtitles on the visual comfort of 3-D TV. *Displays*. submitted for publication.
3. IJsselsteijn, W., Lambooij, M., Kuijsters, A., van Vugt, F., and Heynderickx, I. (2011). Assessing 3-D Displays: Naturalness as a Visual Experience Metric. *Journal of Experimental Psychology*.