Visual experience of 3D TV

DOI:
10.6100/IR609714

Document status and date:
Published: 01/01/2006

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

Visual Experience of 3D TV

Pieter J.H. Seuntiëns

The work described in this thesis has been carried out at
the Eindhoven University of Technology and Philips Research
Eindhoven, the Netherlands.

Printing: Eindhoven University Press Facilities
Omslagontwerp: Oranje Vormgevers
Foto: Xmoix

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Visual Experience of 3D TV / by Petrus Johannes Hendrikus Seuntiëns -
Eindhoven: Technische Universiteit Eindhoven, 2006. - Proefschrift.-
ISBN 90-386-0705-9
ISBN 978-90-386-0705-4
NUR 778
Keywords: 3D television / visual experience / image quality

Visual Experience of 3D TV

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 dinsdag 13 juni 2006 om 16.00 uur

door

Petrus Johannes Hendrikus Seuntiëns

geboren te Eindhoven

Dit proefschrift is goedgekeurd door de promotoren:

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

Copromotor:
dr. W.A. IJsselsteijn

Contents

Chapter 1

Introduction

1.1 Aim of this thesis

Since the introduction of television, much has been done to improve the overall experience of viewers. Improvements in color, picture quality, sound quality, and increasing involvement based on larger screen sizes have contributed to a better overall viewing experience. A logical next step is the introduction of three-dimensional television enabling people to watch their content in three dimensions. Proponents of 3D-TV have argued that it will bring the viewer a whole new experience, a fundamental change in the character of the image, not just an enhancement of quality (IJsselsteijn, 2004; Smith and Dumbreck, 1988).

Comparisons between television sets are done quite regularly on perceptual and/or technical aspects to determine where to put future investments. The performance of a 3D television system is often evaluated using 2D image quality models as proposed by Engeldrum (2000). Earlier research in this area defined some dominant perceptual factors affecting 2D image quality, for instance, blur, brightness, color, blockiness, or noise. Psychophysical scaling experiments are used to quantify the strengths of these artefacts. People use perceptual rules to combine the measured strengths into a prediction of the overall image quality (de Ridder, 1992). This thesis investigates whether 2D image quality models are sufficiently adequate to measure 3D quality because typical stereoscopic distortions and the depth reproduction are not incorporated in 2D image quality models.

The aim of this thesis is to understand, measure and eventually, model and predict the 3D 'Visual Experience'.

1.2 Human Vision

One of the major functions of our visual system is to construct a 3D representation of the world surrounding us. According to Marr (1982), "vision is the process of discovering from images what is present in the world, and where it is". The images on our retina are patterns of reflected light from our environment, and to discover what is present, the visual system relies on internal representations. Since our world is three-dimensional we have to perceive all three dimensions in order to acquire a full representation of these relationships. The problem is however that external space is projected onto the retina of both eyes as two-dimensional images. So the question arises, how is this transformation from two retinal two-dimensional images with a slightly different perspective to a three-dimensional representation of our environment achieved? The mechanism we use to reconstruct the three-dimensional world is referred to as stereopsis.

The sources of depth information can be divided in four categories (Palmer, 1999): ocular information (accommodation and convergence), stereoscopic information (binocular disparity), dynamic information (motion parallax) and pictorial information (occlusion, relative size, etc.)

Ocular information

Two cues for depth perception are the convergence and accommodation of the eye. Convergence is related to the fixation of the eyes (binocular information). If we look at an object nearby the eyes converge more than they do if we look at an object far away. The accommodation of the lens is the process of focusing on an object (monocular information). The muscles of the lens are relaxed when focusing on objects far away and contracted when focusing on objects near by. Muscle tension of the eye in combination with the visual input are essential for depth perception. Although accommodation and convergence are not very strong sources of depth information, they are important at close distances for specifying

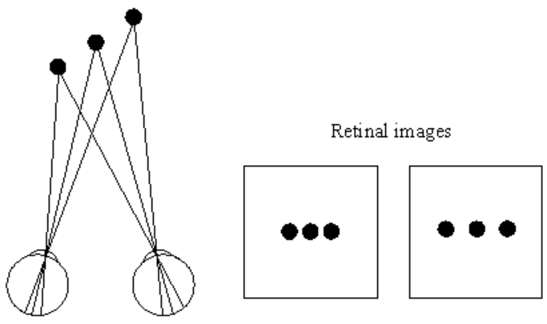


Figure 1.1: Horizontal separation of the eyes causes an interocular difference in the relative projections of monocular images onto the left and right retina

the absolute distance of objects. The absolute distance is the perceived distance from observer to objects.

Stereoscopic information

Perhaps the richest source of depth information comes from stereopsis. Due to the fact that our eyes are separated by 6.3 cm on average (Dodgson, 2004), each eye receives a slightly different perspective of the same scene (Figure 1.1). The brain fuses these two images, and because each image is slightly displaced with respect to the other - a phenomenon known as retinal disparity - the relative (perceived distance between objects) and absolute depths (perceived distance from observer to objects) of objects in space are perceived. The ability of the brain to perform these computations is referred to as stereopsis.

Dynamic information

Dynamic information occurs with a change in visual structure over time because of image motion or optic flow. Depth information about a scene becomes available when observers move with respect to the scene (motion parallax). The motion parallax cue provides depth information because image points at different distances from the observer move at different retinal velocities as the observer moves. Objects that are closer will move faster than objects that are further away.

Pictorial information

Although ocular, stereoscopic, and dynamic information produce a compelling sense of depth, pictorial information is very powerful because it provides a good depth perception of static and monocularly viewed pictures. If you close one eye and keep your head still, the world still looks three-dimensional.

The most powerful monocular cue is occlusion (Figure 1.2, panel a). Occlusion occurs when an object is partly hidden by another object and tells us that the hidden object is further away. Relative size refers to the fact that objects of similar size produce smaller retinal images when placed further away (Figure 1.2, panel b). The height in the visual field cue refers to the fact that objects below the horizon, appear closer to the observer as they are positioned lower in the visual field. Objects above the horizon appear to be closer as they are positioned higher in the visual field (Figure 1.2, panel b). The shading cue provides information about the shape of an object and occurs because not all parts of an object reflect the same amount of light (Figure 1.2, panel c).

The aerial perspective cue arises because the air contains microscopic particles of dust and moisture that make distant objects look less saturated and less sharp. The more atmospheric particles between the viewer and a distant object the more light that is scattered (Figure 1.3, panel a). Linear perspective refers to the fact that parallel lines, such as railroad tracks, appear to converge with distance, eventually reaching a vanishing point at the horizon. The more the lines converge, the farther away they appear (Figure 1.3, panel b).

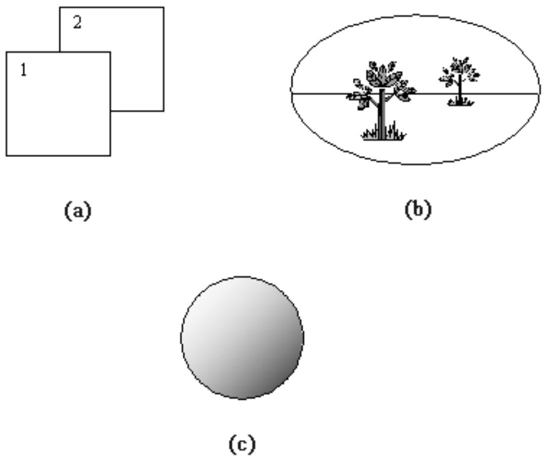


Figure 1.2: Monocular cues providing depth information of objects in a scene: occlusion (a), relative size, height in the visual field (b), and shading (c).

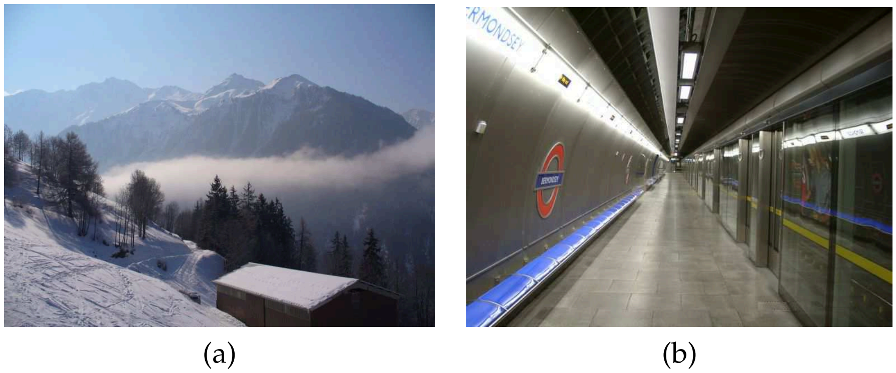


Figure 1.3: Monocular cues providing depth information of objects in a scene: aerial perspective (a) and linear perspective (b).

So, it may be clear that depth perception is a result of many contributing processes, which vary in their degree of cognitive complexity. Apart from the stereopsis mechanism, which can be considered as a relatively low-level computation producing basic surface layout information, there are many higher cognitive functions involved in the interpretations of the various monocular cues, which also affect the way we perceive depth. Even such high cognitive functions as expectations, reasoning and memory or knowledge about objects and the world affect the way we interpret the depth of a visual scene (Mansson, 1998).

1.3 3D-TV broadcast system

The introduction of 3D-TV is becoming increasingly feasible because of recent technologies and breakthroughs in image processing, display design and camera development as well as an improved understanding of 3D human factors. For a successful implementation, the 3D technology should be backward compatible with existing conventional broadcast television to ensure a gradual transition from one system to the other. Figure 1.4 presents a 3D-TV broadcast chain (IST-ATTEST project approach) starting from content generation and coding schemes for efficient transmission to adequate displays presenting a high-quality 3D picture.

1.3.1 Content generation

By far the most 3D material has been shot using a dual-camera configuration. In general, two systems can be distinguished: 1) the parallel configuration and 2) the toed-in configuration. An important difference between both configurations is that for a parallel camera configuration, depth is conveyed exclusively by crossed disparities (objects appear closer to the viewer compared to the fixation point), because the zero-disparity point is located at infinity. Therefore, binocular disparities for objects near the camera (within 2 meters) can be very large and cause visual discomfort. For a toed-in configuration, the zero-disparity point is at a finite distance, so depth is conveyed by both crossed and uncrossed disparities (objects appear closer and further away compared to the fixation point). Consequently, the same depth range is distributed among crossed and un-

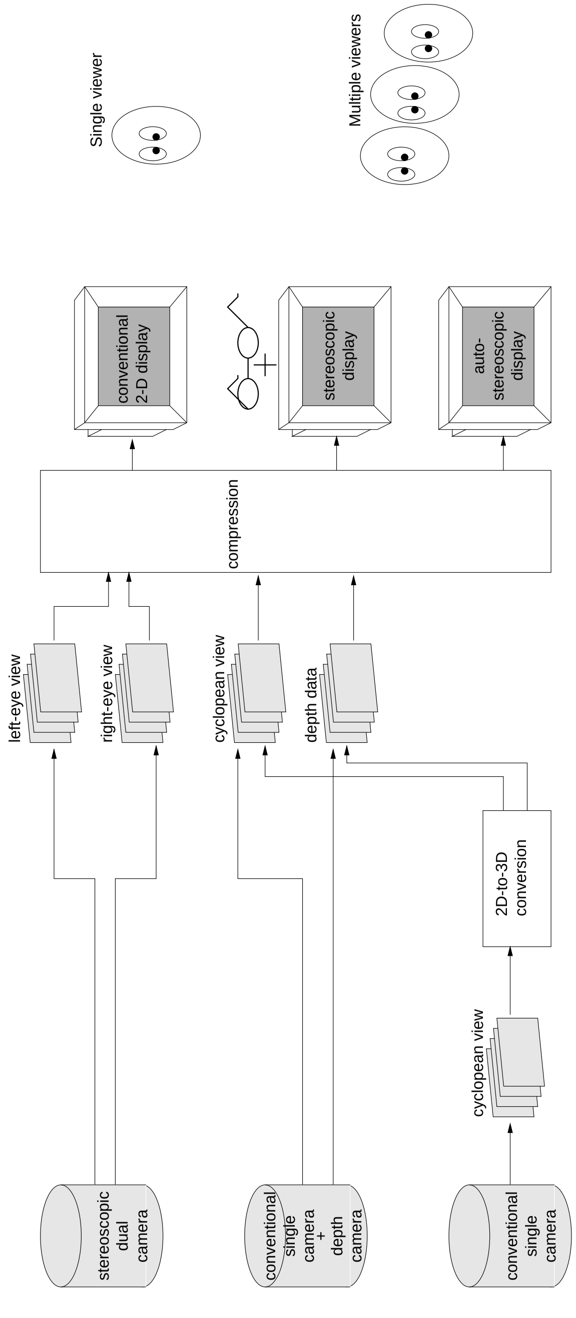


Figure 1.4: 3D-TV broadcast chain including content generation, coding, transmission, and 3D displays (Meesters et al., 2004).

crossed disparities for the toed-in configuration resulting in a smaller absolute disparity compared to the parallel configuration (Stelmach et al., 2003). However, converging cameras introduce keystone distortions of opposite sign resulting in vertical disparities which are greatest in the corners of the image. So, using a converging camera configuration involves a trade off between reduced binocular disparities for objects located near the camera on the one hand (less visual discomfort) and the introduction of vertical disparities on the other hand (more visual discomfort).

The short-term need for 3D-video content can only partially be satisfied with newly recorded material. Therefore, 2D-3D conversion algorithms are being developed to convert existing 2D-video material into 3D. Conversion of existing 2D video material is a challenging task, because of problems with pixel-accurate automatic video segmentation.

1.3.2 Compression and transmission

The storage and transmission of stereoscopic image material involves a large amount of data because one stereoscopic image consists of multiple views. Therefore, a considerable research effort is focused on realizing digital image compression (such as JPEG or MPEG coding) to obtain savings in bandwidth and storage capacity. This is of particular relevance in the case of stereoscopic HDTV, where a single uncompressed HDTV channel may cost up to one Gbit/s transmission bandwidth, or in the case of stereoscopic video transmission over low-bandwidth transmission channels, such as the Internet (Johanson, 2001). In terms of compatibility with current existing broadcast systems, a double bandwidth would be needed for transmitting the left- and right eye view of a dual camera. The use of a depth range camera (Axi-Vision by NHK, Z-cam by 3DV Systems) registering the RGB image and accompanying depth value per pixel overcomes this bandwidth problem. Although this is a promising camera technique, there are still some challenges to recover the left and right-eye view correctly from the RGB-depth video material. The goal is to achieve a video data format that is compatible with traditional coding standards (MPEG-2/4/7) and 2D TV-sets as well as suited for novel 3D TV applications.

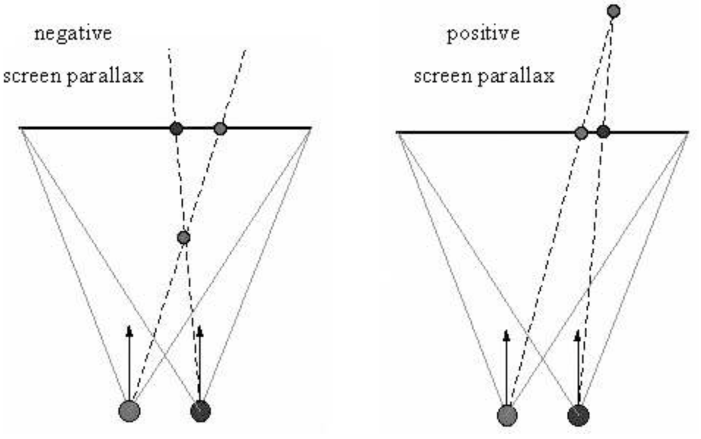


Figure 1.5: The left image shows a negative screen parallax. Objects appear in front of the display screen. The right image shows a positive screen parallax. Objects appear behind the display screen.

1.3.3 Stereoscopic and auto-stereoscopic displays

The principle of stereoscopic imaging systems is based on displaying two images with a slightly different perspective in such a way that the left view is only seen by the left eye and the right view is only seen by the right eye. The horizontal distance on the display screen between corresponding points in the left and right eye view is called the screen parallax. When the screen parallax for a certain point in the image is zero (no difference between left and right eye view), this point will be seen at the screen plane. Negative and positive screen parallaxes result in objects in front or behind the display screen (see Figure 1.5).

In general, there are three distinguishing features characterizing stereoscopic displays namely: 1) the separation technique for the left and right eye view, 2) whether or not motion parallax (multi-view) is supported, and 3) the number of observers that can watch 3D simultaneously. Many techniques can be used to realize left/right eye separation in a stereoscopic display. Usually a distinction is made between stereoscopic and auto-stereoscopic displays. Stereoscopic displays require the viewer to wear an optical device to direct the left and right eye images to the appropriate eye (e.g., polarized glasses, shutter glasses) while the separation technique used in auto-stereoscopic displays is integrated in the display screen. In both stereoscopic and auto-stereoscopic displays, perfect separation of the left and right eye view is one of the major challenges for display designers. An overview of the different types of stereo displays available can be found in Sexton and Surman (1999) and Schreer et al. (2005). In this thesis, three different types of stereoscopic imaging systems are used to carry out the experiments, which will be discussed next.

Philips multi-view auto-stereoscopic display

The first display system is the 20″ Philips multi-view auto-stereoscopic display using a lenticular lens to separate the left- and right-eye view (van Berkel and Clarke, 1997). The advantage of this display, besides 3D viewing without glasses, is the support of motion parallax enabling the viewers to look around objects by moving their head. Figure 1.6 shows the basic principle of the display system. Figure 1.6a shows an observer watching a set of objects. The left and right eye both receive a different view of the scene. By moving their head, observers receive other views of the scene enabling them to see a potentially infinite number of views. Figure 1.6b shows the same viewing window, but this time divided into a finite set of horizontal frames. Each eye receives a view from a single frame, thereby preserving the effect of motion parallax, but with a reduced number of views. Nine different views are recorded and integrated in the multi-view auto-stereoscopic display (Figure 1.6c and 1.6d).

A set of nine successive views is called a viewing zone and repetition of this viewing zone enables multiple viewers to watch 3D. Figure 1.7 shows three zones consisting of nine views each. The resolution of the two 3D multi-view displays available was 1600x1200 pixels and the optics were optimized for a viewing distance of 0.4 and 1.5 meters, respectively.

AEA-Technology polarized stereoscopic display

The second stereoscopic display used in one of the experiments was developed by AEA-Technology (AEAT). The AEAT system consists of two Barco CPM 2053FS CRT color monitors mounted perpendicular to each other (see Figure 1.8). The dual monitor system displayed the right and left image at the same time using a half see-through mirror and a polar-

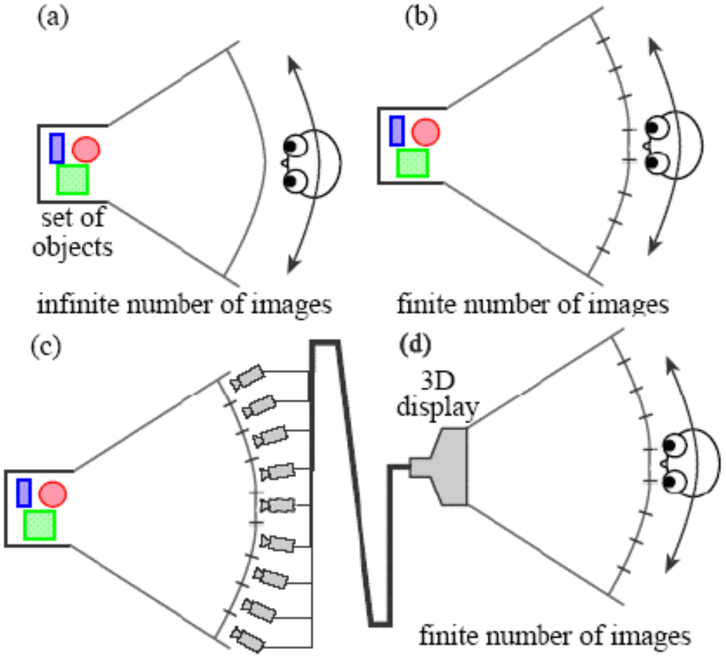


Figure 1.6: Basic principle of a multi-view auto-stereoscopic display. Panel (a) shows an observer watching a set of objects. The viewing window is divided in nine different perspective views in panel (b). The nine different views were recorded using nine different cameras as shown in panel (c). The screen displays the nine different views in a viewing zone in panel (d).

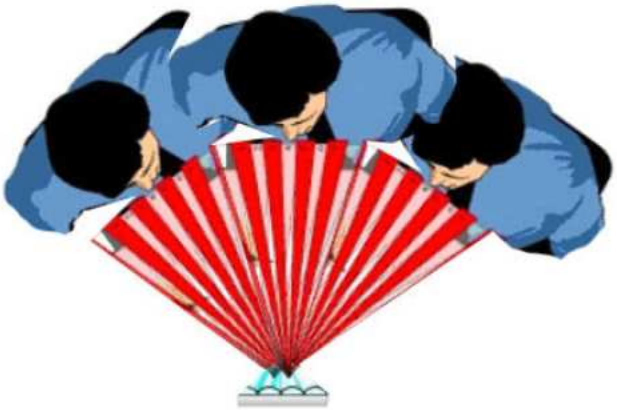


Figure 1.7: Three viewing zones consisting of nine different perspective views each. The repetition of viewing zones enables multiple viewing.

ization filter in front of each screen. The observers wore polarized glasses in order to provide left-right separation with very little crosstalk in the stereo pair. The linear polarized filters contained less than 0.1% crosstalk (Pastoor and Wöpking, 1997). A SUN ISP system provided the CRT monitors with a video signal. Custom built software was used to synchronize the output of the 2 codecs transferring the images.

Screenscope mirror stereoscope

The Screenscope^{TM} (mirror stereoscope) was used to direct the left- and right-eye image of a side-by-side displayed stereo pair to the appropriate eye. The Screenscope was attached to the computer screen as shown in Figure 1.9. This system is location multiplexed thus containing zero crosstalk, allowing us to have complete experimental control. This is not the case with systems based on, e.g., shutter glasses (time-multiplexed) or polarized glasses (polarization-multiplexed), where crosstalk is intrinsic to the system.

The principle of the Screenscope viewer (see Figure 1.10) is based on the Wheatstone stereoscope (Wheatstone, 1838). The only difference between

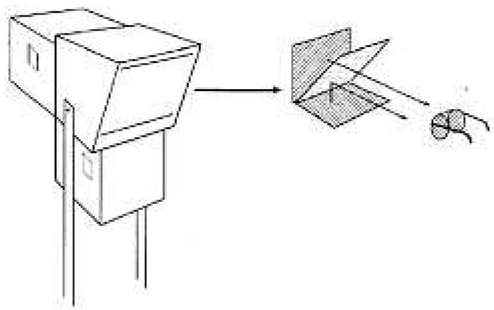


Figure 1.8: The AEAT system consisting of two Barco CRT monitors displaying the left and right eye images at the same time. The polarized glasses are used to separate the left and right views.

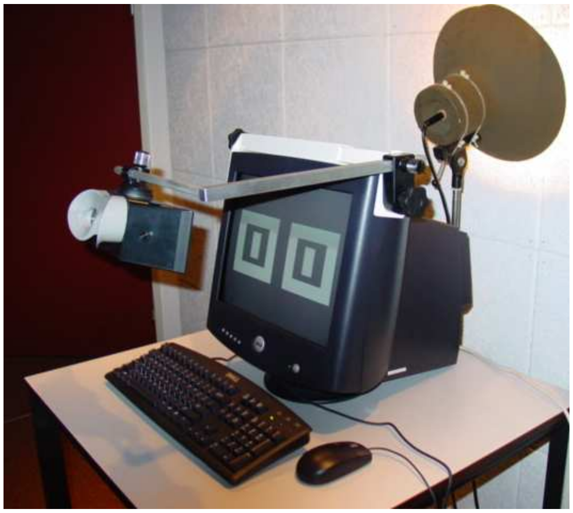


Figure 1.9: Screenscope^{TM}attached to a PC monitor with a light source in the background. The monitor shows a stereoscopic test pattern illustrating the approximate size of the natural stimuli that were employed.

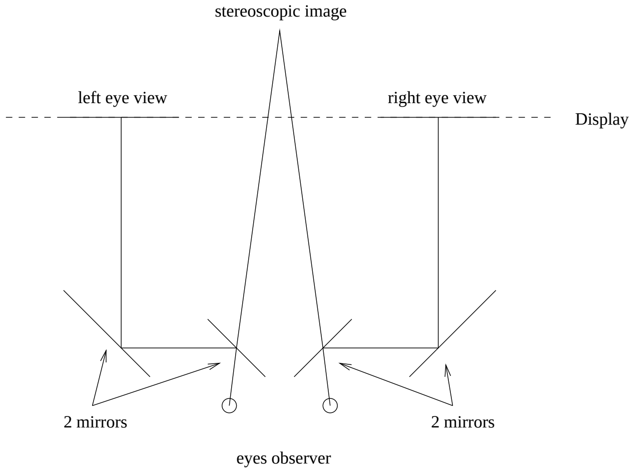


Figure 1.10: Principle of the Screenscope^{TM}stereo viewer based on the Wheatstone construction. The difference is that this system uses four mirrors instead of two.

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. The viewing distance from the Screenscope to the CRT screen was 30 cm (three times stimulus height).

1.4 Subjective assessment methods

Subjective assessment methods for evaluation of 2D and 3D television systems are a necessity to compare competitive systems and monitor applications. Standardized methods to quantify perceptual attributes such as perceived image quality, depth, and sharpness enable engineers to optimize their display systems. Subjective assessment methods use a human being as the measuring instrument to determine the quality of a display system. These methods are often viewed as inferior measurement methods compared to objective methods (physical measures). This may be true from a precision or accuracy point of view, but it misses the fundamental point that humans are the customers of the imaging systems, so their view on image quality is the correct one. Therefore, subjective assessment methods for perceptual evaluation of monoscopic and stereoscopic television pictures, such as described in recommendations of the ITU (2000a) and ITU (2000b), are widely accepted. Assessment methods used to evaluate new imaging systems like 3D TV can be divided in explorative studies and direct scaling paradigms.

1.4.1 Explorative studies

Explorative studies are used to explore viewers' unprimed attitude, feelings and reactions towards a new technology such as 3D TV. An example of an explorative study are focus groups, where naive viewers participate in small groups and discuss their experiences while viewing an imaging system. Freeman and Avons (2000) used focus group experiments to collect viewers reactions about novel 3D TV. The results showed that viewers report a sense of "being there" when watching 3D content. Furthermore, this feeling of "being there" was related to attributes such as realism, naturalness, and involvement. The focus group also identified program types suited for 3D TV. In general, observers preferred action movies and life events such as sports, theater, and concerts. Program types such as news, soap operas, documentaries, and talk shows were thought of as inappropriate for 3D TV. Moreover, observers indicated that they would like to decide on a program-by-program basis whether they wanted to watch it in 2D or 3D. In summary, focus groups can be used to (i) collect unbiased viewer's descriptions of the sensations evoked by a stereoscopic imaging system, (ii) investigate the added value of new imaging systems, without imposing predefined appreciation criteria such as image quality, and (iii) determine attributes underlying concepts such as image quality, naturalness and presence without directed questions.

1.4.2 Direct scaling paradigms

Several experimental paradigms can be used to measure and quantify image quality of images and sequences. Roufs (1992) differentiates between two types of perceptual image quality: performance-oriented and appreciation-oriented image quality. Performance-oriented image quality is applicable whenever the purpose of the images is to facilitate detection tasks, for instance, medical diagnosing. The purpose of those images is to give accurate information. In appreciation-oriented applications, such as 3D TV, the goal is to display 3D images as ”pleasing” as possible. For instance, excessive disparities result in visual discomfort and viewers experience this as unpleasant. The subjective assessment of appreciation-oriented applications, such as 3D TV, is described in the ITU-R BT.1438 recommendation for stereoscopic television pictures (ITU, 2000b). These assessment methods are adopted from the ITU-R BT500.10 recommendation for conventional 2D TV (ITU, 2000a). The proposed methods are used to measure overall image quality and overall image impairment of distorted still images and image sequences. The methods can also be applied to obtain ratings for attributes such as sharpness, depth, eye-strain, naturalness, or presence. In general three different experimental paradigms are proposed: the double-stimulus methods, single-stimulus methods and stimulus-comparison methods.

Double stimulus methods

In the double-stimulus-continuous-quality-scale approach (DSCQS), observers assess the overal image quality for a series of image pairs, each consisting of an undistorted image (reference) and a distorted image (test). Observers are asked to assess the overall image quality of both (reference and test) resulting eventually in difference scores between reference and test image. In DSCQS, a continuous graphical scale (labeled with verbal terms excellent - good - fair - poor - bad) is used to avoid forcing observers to answer within too coarse a category. In the double-stimulus-impairment scale method (DSIS), again a series of stereoscopic images are presented in time (reference + test), however, observers are asked to judge only the impairments in the test image taking in mind the reference. The scale used during impairment scaling is labeled with the verbal terms imperceptible - perceptible, but not annoying - slightly annoying - annoying - very annoying.

Single stimulus methods

In single-stimulus (SS) methods, the subject assesses each image in the stimulus set individually. In case of a sequence, the subject provides a score for the entire presentation. Also SS-methods can be applied on both quality scaling and impairment scaling using corresponding rating scales.

Stimulus comparison methods

Stimulus comparison methods assign a relation between two images or sequences. The comparison scale used during an experiment is labeled with the verbal terms much worse - worse - slightly worse - the same - slightly better - better - much better.

In the context of 3D-TV, an alternative assessment method (single-stimulus-continuous-quality-evaluation) was proposed to obtain continuous quality judgements of longer stereoscopic sequences moving a hand-held slider. IJsselsteijn et al. (1998b) used this method to continuously assess observer's sense of presence, depth and naturalness watching 3D-TV over a longer period of time. This method seems very appropriate because normally television is watched for longer periods and it mimics home viewing conditions.

1.5 Image Quality

Image quality can be regarded as one of the most important considerations of customers in purchasing an imaging or display product, along with purchase factors such as costs. Achieving good image quality requires extensive research in content generation, coding algorithms, transmission and display technology. Therefore, it is important to connect the preferences of customers to the technological parameters of the display system. Perceived 3D image quality is one of the criteria to assess the overall performance of new media such as 3D-TV. However, subjective testing is time-consuming and needs to be repeated for each new parameter setting. Therefore, quality models are needed to obtain a better understanding of the relationship between technical system parameters and perceived 3D image quality. For conventional imaging systems, image quality models have been proposed to predict 2D image quality. Nevertheless, a better understanding is needed of the relationship between system parameters and perceptual factors contributing to the overall perceived 3D image quality. The principles of modeling 2D image quality can be used to gain insight into the relationship between 3D-TV system parameters and 3D image quality.

1.5.1 Image Quality Modeling

Several approaches have been proposed to obtain a quantitative measure of image quality for conventional 2D images or sequences. In this paragraph, some quality models are discussed that are based on 1) a mathematical function to express the loss of information in a physical signal, 2) the transformations in the peripheral human visual pathways, 3) identifying and quantifying the impairment strengths, and 4) knowledge of human visual information processing.

Objective fidelity criterion models use a mathematical function of the original image and a processed version of it, to express the loss of information in an image. Often used functions are the root mean square error (RMSE) or the mean-square signal-to-noise ratio (SNR) (Gonzalez and Woods, 1992) The simple calculations needed to express the loss of image information have led to a large number of related measures (Eskicioglu and Fisher, 1995). Objective fidelity criteria are probably satisfactory within certain constraints but are not always suited as image quality measures. For instance the image quality of a particular scene processed at several levels with the same processing method can probably be quantified by these objective fidelity criteria. However, applied across scenes or different types of distortion their reliability is most questionable. Daly (1993) showed that differently impaired images with similar RMSE can be of different subjective quality.

The lack of taking the visual system into account is probably one of the serious limitations of the above mentioned measures. Instrumental image quality measures that include properties of the human visual system (HVS) are more likely to approximate subjective image quality. HVS-based quality measures model the path an image passes through the human visual system, including the optics of the eye, the retina, and the primary visual cortex. Several variations of implementing these stages of the visual system are possible (Ahumada, 1993; Watson, 1987; Daly, 1993; van den Branden Lambrecht, 1996; Winkler, 1999). A typical HVS measure is described in detail by Lubin (1993).

A different technique to model image quality is based on identifying the underlying attributes of image quality and quantifying the perceived strengths of each attribute. For this approach, descriptions of the subjective attributes, such as noise, blur or blockiness, as well as their technical characterization are needed (Karunasekera and Kingsbury, 1995; Kayargadde and Martens, 1996b; Libert and Fenimore, 1999). To relate the attribute strengths to overall image quality, different combination rules can be used (de Ridder, 1992). The attribute strengths can be quantified from the reference image, usually the original, and a processed version of it (Karunasekera and Kingsbury, 1995). At present, much effort is spent on developing single-ended measures, which quantify the degree of impairment directly from the processed image and do not require an original image. For example, estimation algorithms based on the Hermite transform were used to estimate the perceptual strength of blur and noise or blockiness directly from the processed image (Kayargadde and Martens, 1996a; Meesters, 2002).

Another approach is to consider image quality in terms of the adequacy of the image to enable humans to interact with their environment. In this sense image quality is related to terms like usefulness and naturalness, expressing the precision of the internal image representation and its match to the description stored in memory, respectively. To quantify the image quality attributes usefulness and naturalness, measures of discriminability and identifiability were used (Janssen and Blommaert, 2000).

1.5.2 Engeldrum's Image Quality Model

One of the criteria to evaluate the performance of an imaging system is to assess the perceived image quality. 2D image quality is considered to be a multidimensional construct and is affected by several technical parameters. Modeling 2D image quality starts with defining the most important attributes influencing image quality, for instance, blockiness, brightness, noise, color rendering, and blur. Adequate assessment methods to define such attributes are for instance focus groups. Subsequently, the strength of those attributes is measured with psychophysical scaling methods as defined by the ITU. People use perceptual combination rules to combine the strengths of these attributes and finally come to a prediction of the overall 2D image quality. This relation between technical

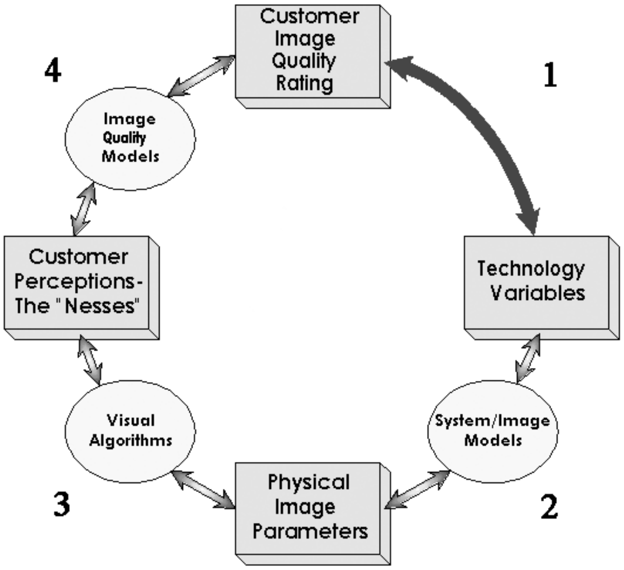

Customer Image Quality Rating
Technology Variables
System/Image Models
Physical Image Parameters
Visual Algorithms
Customer Perceptions- The "Nesses"
Image Quality Models
1
2
3
4

Figure 1.11: Image Quality Circle originally proposed by Engeldrum (2000, 2004).

system parameters and the customer's quality rating is described in the Image Quality Model (Figure 1.11) proposed by Engeldrum (2004), based on his earlier work in Engeldrum (2000).

The four elements in the Image Quality Circle break down the model in measurable and definable steps. Customer quality ratings reflect the customer's judgement about the overall image quality. The technological parameters are a set of elements that the imaging system designer manipulates to change the image quality. Physical image parameters are the measurable properties of the display that are normally ascribed to image quality, such as optical density, spectral reflectance or color. Customer perceptions such as, e.g., sharpness, darkness, and graininess form the basis of the quality rating or judgment by the customer. The direct link between technological parameters and the customer quality rating (arrow 1) is inefficient over time because customers have to judge the quality over and over again every time a technical parameter is changed.

2D image quality models are not adequate to measure 3D visual experience since depth reproduction, the most important factor in 3D-TV, and typical stereoscopic distortions (for instance crosstalk, or image ghosting) are not incorporated. So, a 3D visual experience model is required that is multidimensional, incorporating perceptual factors related to reproduced depth, 3D image impairments, and visual comfort.

1.5.3 3D Image Quality

No comprehensive 3D visual experience model has been formulated to date, yet it is likely that a diverse set of image attributes contributes to the overall perceived quality of 3D-TV images. Some attributes will have a positive contribution to the overall image quality (e.g., increased depth sensation, or increased sharpness), while others may have a limiting or negative effect (e.g., visual discomfort due to exaggerated disparities, or image distortions). An appropriate 3D visual experience model will account for both positive and negative factors, allowing for a weighting of the attributes based on perceptual importance, and for interactions that may occur as a consequence of (potentially asymmetric) binocular combinations. For example, a 3D distortion like crosstalk becomes more visible with increasing left-right image separation, a manipulation that also increases perceived depth. In such a case the perceptual benefit of increased depth can be nullified by the perceptual cost of increased crosstalk. The interactions between such positive and negative contributions, and their relative weighting deserve further study, in order to arrive at a more complete understanding of 3D visual experience. The 3D visual experience is a trade off between positive and negative factors and should therefore contain the attributes image quality, depth and visual comfort (see Figure 1.12). The added value of depth needs to be incorporated in a 3D visual experience model, especially when 2D picture quality is used as reference (Schreer et al., 2005). IJsselsteijn et al. (2000c) already demonstrated the added value of depth for uncompressed stereoscopic images. Other research, however, showed that when observers were asked to rate the perceived image quality of MPEG-2 and JPEG compressed images, the image quality results were mainly determined by the introduced impairments and not so much by depth (Tam et al., 1998). In this thesis new concepts are explored and it is investigated wether they respond sensitively to the added value of depth when 2D or 3D distortions are present in the image material. The new concepts are explained in the next section.

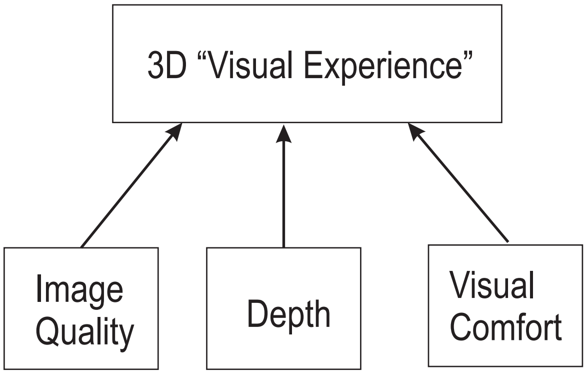


Figure 1.12: Proposed 3D ″Visual Experience″ model with underlying attributes image quality, depth and visual comfort.

1.6 New concepts

1.6.1 Presence

Witmer and Singer (1998) defined the concept presence as the subjective experience of being in one place or environment even when one is situated in another. Presence is also referred to as an unremarked sense of ″being there and reacting to″ in a mediated environment (IJsselsteijn et al., 2000a; Slater et al., 2002). Today, the construct of presence is of particular interest because it has potential relevance for the design and evaluation of interactive and non-interactive media. For instance, new broadcast and display developments such as 3D-TV give more sensory information to the viewer than conventional flat 2D-TV. The addition of binocular depth in 3D-TV gives people a higher sense of ″being there″ in a displayed 3D scene. As the sense of presence increases, people become more aware of the mediated environment, and less aware of the environment in which they are physically located. Freeman and Avons (2000) performed an explorative study, discussed earlier, using focus groups to explore viewers' reactions to conventional 2D-TV and novel 3D-TV. The results showed that non-expert viewers reported sensations of presence with respect to stereoscopic sequences. Furthermore, this sense of presence was related to attributes such as involvement, realism, and naturalness. IJsselsteijn et al. (1998b) first applied the concept of presence to 3D-TV research. They concluded that an increase in sensory information, through the addition of stereoscopic and motion parallax cues, may enhance the viewers' sense of presence. However, when image material became unnatural the sense of presence directly decreased. Other research revealed that moving sequences in contrast to still scenes had a large significant effect on presence ratings, however, the significant effect of dimension (2D/3D) on presence ratings was relatively small (IJsselsteijn et al., 2001). Therefore, it seems that presence may be a useful concept for measuring the added value of 3D stereoscopic moving sequences, at least for distortion-free image material.

1.6.2 Naturalness

Originally, the term naturalness was introduced to establish a criterion for determining the perceived quality of color reproduction, especially in color photography (Fedorovskaya et al., 1997). It has been proposed that images of good quality should at least be perceived as natural, implying a strong relationship between perceived naturalness and the quality of images of real-life scenes. As support, a high correlation between quality and naturalness judgments has been obtained (Fedorovskaya et al., 1997; de Ridder, 1996). This finding is consistent with other data (Laihanen et al., 1994) showing that the impression of naturalness of reproduced skin colors correlates positively with an improvement in image quality. IJsselsteijn et al. (2002) emphasize the relationship between perceived quality and naturalness in the context of color rendering. They state that perceived quality does not necessarily mean a realistic or truthful reproduction. The difference between naturalness and quality as a subjective evaluation concept lies in the fact that naturalness refers to what observers perceive as a truthful representation of reality (i.e., perceptual realism), whereas perceived quality refers to a subjective preference scale. Research on image quality in the color domain has shown that observers are able to differentiate between the two concepts in an experimental situation, and an interesting relation between image quality and naturalness has been demonstrated. For instance, de Ridder et al. (1995) and de Ridder (1996) found a small but systematic deviation between image quality and naturalness. This deviation was interpreted to reflect the observers preference for more colorful but, at the same time, somewhat unnatural images.

Results in the area of stereoscopic image evaluation suggested a similar relation between quality and naturalness. Observers preferred (i.e. judged of high quality) a reproduction of stereoscopic depth they also judged to be slightly unnatural (IJsselsteijn et al., 1998a, 2000c). In a study assessing viewers depth and naturalness ratings to stereoscopic video sequences, IJsselsteijn et al. (1998b) showed that depth and naturalness were related, yet could vary independently depending on the scene content and image parameters (stereo, motion parallax). Not all stereo images look realistic because different kinds of distortions can be introduced into a stereo image. The image may contain exaggerated depth or compression, and the apparent scale of an object may be enlarged or reduced. These effects are the result of variables associated with content generation, coding and displaying techniques. When a view does reproduce spatial realism faithfully, it is called an orthoscopic view. When shooting an orthoscopic view, the angular field of view of the camera must match the angular field of vision of the observer. The two recorded viewpoints by the camera must be separated by the same distance as the distance between a typical observers eyes. Yamanoue et al. (1998) showed in subjective tests that stereoscopic images shot under orthostereoscopic conditions duplicate the real space at a certain display size. Also 3D programs shot under the same conditions look more natural than those shot using the toed-in camera configuration at any display size. In sum, the naturalness concept seems to be a concept taking into account the added value of 3D and also image distortions due to content generation, coding or display techniques. In this thesis, the naturalness concept will be explored further as an evaluation criterion for 3D-TV.

1.6.3 Viewing experience

Stereoscopic displays are expected to enhance the user's viewing experience, however, to date little research has been carried out using viewing experience as an evaluation criterion. Viewing experience is, just like image quality, a complex, multidimensional concept reflecting users' general experience with a certain application. Previous experiences with comparable applications can affect the strength of viewing experience. The intensity of emotional sensations (linked with viewing experience) decreases when the interaction frequency with the application increases. A higher degree of imagination is expected to increase the viewing experience similarly as with presence, i.e., when watching a movie, we know we are not ’in’ the movie, but we nevertheless react in a physical and emotional sense to the story. But, when a movie is watched more than once, the intensity of our reactions decrease, both physically as well as emotionally, and therefore the viewing experience may decrease. In this thesis, the concept viewing experience as an attribute to measure the added value of depth in stereoscopic imaging systems is explored.

1.7 Overview of this thesis

The central research aim of this thesis is how to understand, measure and, eventually, model and predict the 3D ’Visual Experience’.

It is important to have a clear understanding of the potential added value (depth dimension) and drawbacks (eye-strain) of a 3D-TV broadcast service. A 3D visual experience model, incorporating perceptual factors related to reproduced depth, image quality, and visual comfort, could contribute to a more effective design circle for 3D-TV and the technological parameters can be optimized to the customer’s quality preference. The users’ experience is evaluated on a perceptual basis by subjective assessments methods. Choosing the appropriate criterion that incorporates depth, quality and visual comfort is of essential importance for measuring the overall 3D visual experience. This thesis explores several new concepts for the evaluation of 3D content and compares these concepts with traditional image quality criteria.

Chapter 2 addresses the relative importance of image quality and depth on naturalness, presence and viewing experience. The first experiment of this chapter presents a first exploration on assessment criteria for stereoscopic image material when using different 2D to 3D conversion algorithms. Each of these prototype 2D to 3D conversion algorithms generated some 3D (depth artefacts) distortions. In the second experiment, two assessment criteria with the most discriminating power (resulting from experiment 1) were used to investigate the added value of 3D over 2D stills. The experiment used ’perfect’ 3D content with no conversion or depth artefacts. Manually, several noise levels were added to the 2D and 3D images to degrade image quality. The goal was to investigate which evaluation term incorporates depth the most in the absence of depth artefacts. The best criterion was further used in Chapter 3 and 4 investigating different 2D (JPEG coding) and 3D (crosstalk) artefacts.

In Chapter 3, the presented work investigates the effect of JPEG coding (typical 2D distortion) in combination with a variation in camera separation (2D/3D) on perceived image quality, perceived sharpness, perceived depth and perceived eye-strain of stereoscopic images. The next experiment in this chapter investigates whether the added value of depth in JPEG-impaired images (2D distortion) can be measured using the naturalness criterion (resulting from Chapter 2).

Chapter 4 addresses the effect of crosstalk (typical 3D distortion) in combination with a variation in camera separation (2D/3D) on perceived image distortion, perceived depth and perceived visual strain. The second experiment investigates whether the added value of depth in crosstalk-impaired images (3D distortion) can be measured using the naturalness criterion (resulting from Chapter 2).

Chapter 5 describes an experiment combining both experiments from Chapter 2 measuring the effect of a reduction in image quality in combination with the added value of depth on image quality, depth, viewing experience, and naturalness. In this chapter the viewing experience and naturalness is predicted in terms of image quality and depth with a linear regression analysis.

In Chapter 6, we will briefly look back on the previous chapters and discuss the most important findings. At the end we redefine the 3D visual experience model as described in Chapter 1 and discuss the applicability of the model.

Chapter 2

Exploration of new evaluation concepts for 3D-TV

Abstract

The goal of this chapter is to explore and determine which evaluation criterion is most appropriate to assess the performance of 3D-display systems. It is assumed that these evaluation criteria take into account image quality as well as reproduced depth. The criterion that weighs depth most in addition to image quality is considered most appropriate. Experiment 1 explores the assessment criteria image quality, depth, naturalness, presence and viewing experience. It presents empirical work on these assessment criteria for stereoscopic image material when using different 2D to 3D conversion algorithms. Results show that viewing experience and naturalness have the most discriminating power between the various algorithms. Hence, the second experiment focuses on these criteria and uses 'perfect' 3D content with no conversion or depth artefacts. Several noise levels were added to the 2D and 3D images to degrade image quality. The goal is to investigate whether viewing experience or naturalness incorporates depth the most in the absence of depth artefacts. Results show that the noise distortion is weighted equally both with viewing experience and naturalness. Naturalness is more sensitive to depth than viewing experience.

[0] This chapter is based on Seuntiëns et al. (2005a) and Lambooij et al. (2005)

2.1 Introduction

The main objective of this chapter is to explore and determine which evaluation criterion is most appropriate to assess 3D quality. It is assumed that 3D evaluation criteria take into account image quality as well as reproduced depth. The criterion that gives depth the highest weighting in addition to image quality is considered most appropriate for 3D-TV research. Earlier research confirmed that image quality has a relationship with presence and naturalness (Fedorovskaya et al., 1997; de Ridder, 1996; IJsselsteijn et al., 2002) and most likely with viewing experience as well. From literature it is known that depth perception is related to presence (IJsselsteijn et al., 1998b) and it is assumed that this is also the case for naturalness and viewing experience. So far, the relation between image quality, depth, viewing experience, naturalness, and presence is not known. Therefore, an experiment was performed in which image quality and the depth percept were assumed to vary, and measured image quality and depth together with viewing experience, naturalness, and presence (see Chapter 1). For this experiment, (not yet optimal) 2D-3D conversion algorithms introducing depth artefacts and a realistic set of test scenes were used to explore the concepts image quality, depth, viewing experience, naturalness, and presence. The most appropriate assessment concepts resulting from the first experiment were used in the second experiment. The second experiment used recorded 3D content (nine cameras) with no conversion or depth artefacts. Manually, several noise levels were added to the 2D and 3D images to degrade image quality. The goal was to investigate which evaluation term incorporates depth the most in the absence of depth artefacts.

2.2 Experiment 1

This experiment explores the assessment criteria image quality, depth, naturalness, presence and viewing experience. It presents empirical work on assessment criteria for a realistic set of stereoscopic image material when using different 2D to 3D conversion algorithms.

2.2.1 Method

Design

The experiment had a within subjects design with Image (ten scenes) and Algorithm (four conversion algorithms) as independent variables and image quality, depth, naturalness, presence, and viewing experience as dependent variables.

Observers

Two female and eighteen male naive observers participated in the experiment. Three observers were employees in a research environment and seventeen observers were internal graduate students with a technical background. Their ages ranged from 24 to 32. Four observers had prior experience with viewing 3D material. All observers had good stereo vision <40 seconds of arc (as tested with the Randot stereo test).

Equipment

A 20″ Philips multi-view auto-stereoscopic display was used in this experiment as described in Chapter 1, section 1.3.3. The viewing distance of the observers was 0.4 meters. Nine different views were generated using 2D-3D conversion software (thus, not recorded by nine cameras) and these nine views were integrated in the multi-view auto-stereoscopic display. Custom-built software (PORT, Perceptie Onderzoek Research Tool) was used to conduct this psychophysical experiment. The custom-built software enables communication between three different hardware components. The first component is the PORT console, which is an ″ordinary″ PC that displays the user interface to the test leader, controls the experiment and gathers the test results. The second component, a notebook, is the observers' interface on which the assessment is executed. The third component is the video device, which displays the stimuli on the 3D display. All assessments took place on the laptop, except for the session with image quality and depth. PORT is not able to display two assessment scales simultaneously, so the assessment was made on paper.

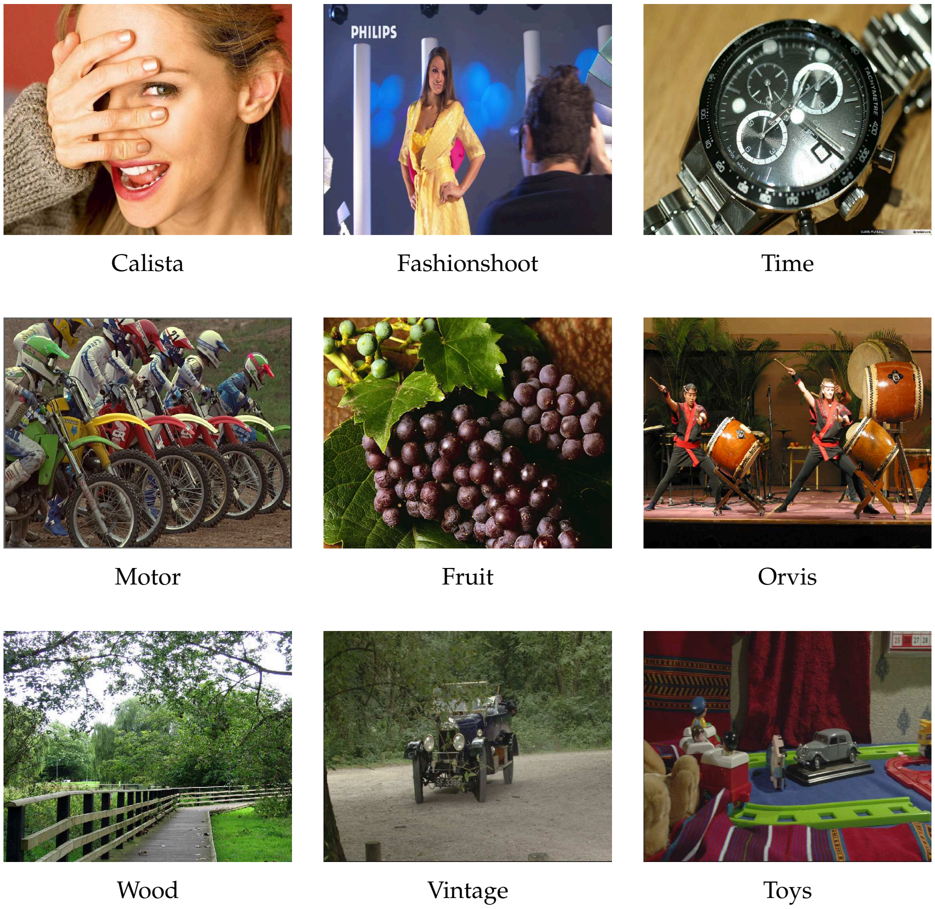
Stimuli

The image material used in this experiment consisted of ten original scenes, both moving (10 seconds) and static scenes. The originals contained objects, humans, and nature covering a broad range of image material and various kinds of distortions including depth artefacts due to imperfect 2D-3D conversion algorithms. Figure 2.1 shows the nine images as used in the experiment. The images *Fashionshoot*, *Toys*, and *Vintage* are single frames taken from the sequences, where the *Fashionshoot* image was also used as a static scene in the experiment. Thus, in total seven static scenes and three moving sequences were assessed by the observers.

The ten 2D originals were converted into 3D stimuli with four different conversion algorithms. All conversion algorithms use single 2D images as input for depth map estimation. The output image (1600x1200) contains all nine views as generated by the algorithm and can directly be displayed on the 20″ Philips multi-view autostereoscopic monitor. The ′focus′ algorithm estimates the amount of depth based on the assumption that blurring is caused by the limited focal depth of the camera. Objects in-focus are clearly rendered and objects at other distances are blurred. Depth from ′gravity′ relies on the fact that the bottom of most objects is connected with the object below it, e.g., the ground, a table or chair. The direction of the gravity is regularly downwards, i.e., with the bottom of the image closer to the viewer than the top of the image. The ′luminance′ algorithm assigns depth values based on their luminosity. Dark areas are considered to be far away and light areas closer by. The ′realtime′ algorithm is a combination of the ′focus′, ′gravity′, and ′luminance′ algorithms. For all algorithms, the view-offset was set to 1/3 which implies that 1/3 of the depth volume was displayed in front of the display screen (and 2/3 behind the display screen). The virtual camera distance was set to 0.01 meters for all images.

Procedure

The 3D content was evaluated by five different evaluation criteria, namely image quality, depth, naturalness, presence and viewing experience. The 5-point categorical scale for all evaluation criteria was labeled with the adjective terms [bad]-[poor]-[fair]-[good]-[excellent] according to the ITU



Calista

Fashionshoot

Time

Motor

Fruit

Orvis

Wood

Vintage

Toys

Figure 2.1: Original scenes used in the experiment. *Fashionshoot*, *Toys* and *Vintage* are frames taken from sequences, where *Fashionshoot* was also used as a still in the experiment, making a total of 10 stimuli.

(2000a) recommendation on subjective quality assessment. The criteria naturalness, presence and viewing experience were evaluated in three different sessions and image quality and depth were evaluated together in one session. Each session consisted of 80 conditions (4 algorithms x 10 scenes x 2 repetitions) randomized for each session to prevent order effects. Prior to the experiment, observers were given a brief introduction on paper about the experiment. Any remaining questions were answered and subsequently a short training session was conducted.

The training session allowed the observers to get used to the setting as well as the tasks. The training consisted of two parts. In the first part observers could scroll back and forth through twelve 3D images (three images converted with four algorithms) to get acquainted with viewing 3D image material and with the different kind of distortions. The image content used in this first part of the training session was not rated and not used in the actual experiment.

The second part of the training consisted of six 3D images that were also used in the rest of the experiment. This second part of the training was implemented to make the observers familiar with the assessment method and again, to make them acquainted with the different kind of distortions in the images. Subsequently, the actual experiment (80 stimuli) started and took approximately 45 minutes. The lighting conditions of the room were constant for all observers and the level of light in the room was 3 lux, measured perpendicular to the display in the direction of the viewer.

2.2.2 Results

Figure 2.2 shows the mean ratings of the five evaluation criteria per conversion algorithm on the y-axis. On the x-axis, four different conversion algorithms are presented. The lines in the figure represent the five evaluation criteria. Connecting lines between data points are used for a more convenient and quicker interpretation of the data, but do not indicate any relationship.

Looking at figure 2.2, some results catch the eye immediately. First of all, the averaged scores for the 3D images are relatively low for all assessment criteria and for all algorithms, indicating that on average observers perceived the content as 'fair'. The horizontal trend of the criterion 'depth'

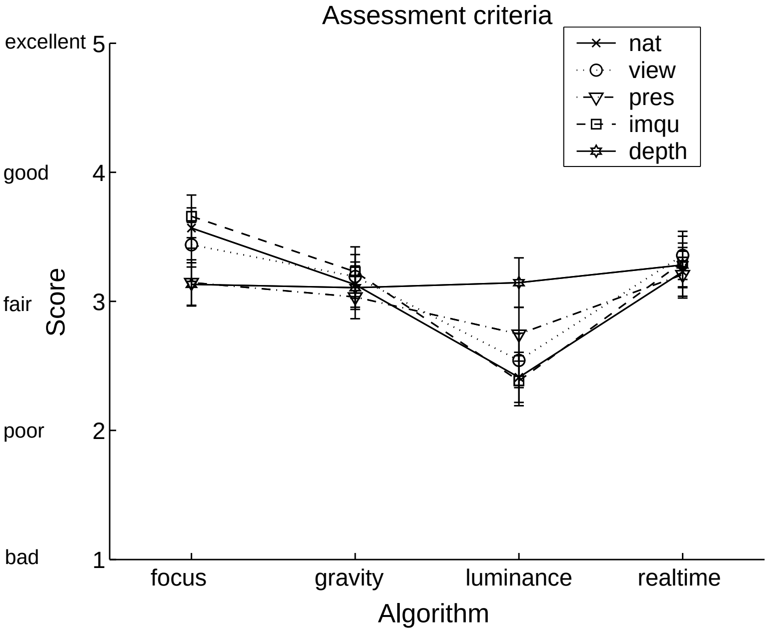


Figure 2.2: Effect of algorithm on the assessment criteria naturalness, viewing experience, presence, image quality and depth. The x-axis represents the four conversion algorithms and the y-axis represents the averaged ratings and the standard errors for all criteria. The lines in the graph represent the assessment criteria.

indicates that perceived depth is not affected by the different conversion algorithms. Moreover, the criteria naturalness, viewing experience, presence and image quality were assessed quite similarly for all algorithms. Furthermore, the algorithm luminance shows the lowest averaged scores for these criteria.

A MANOVA was performed with Algorithm and Image as independent variables and the five evaluation criteria as dependent variables. Results show that the evaluation criteria naturalness, image quality, viewing experience, and to a lesser degree presence made similar significant discriminations between the four algorithms. The perceived depth was the only criterion that was not affected by the different algorithms (F(3,734) = 2.279, p=0.078). Furthermore, there was a significant effect of Image for all evaluation criteria.

Next, the MANOVA was split up in moving and still scenes. Viewing experience and naturalness were assessed relatively similarly to image quality as a function of the algorithms for both sequences and stills. Presence, however, was assessed differently for stills than for sequences. For the stills, presence revealed a similar behaviour to image quality as a function of the conversion algorithms. Yet, this behaviour is not as strong as the relation between image quality, viewing experience and naturalness. For sequences, there was no significant effect of algorithm on the presence scores (F(3,212) = 0.237, p=0.870) and no significant effect on the depth scores (F(3,212) = 1.974, p=0.119). Thus, the presence ratings resembled more the depth scores. Following IJsselsteijn et al. (2001), presence seems to be a good measurement attribute for evaluating stereoscopic sequences, but maybe less appropriate for evaluating stereoscopic stills, since their results showed that the introduction of motion had a much higher impact on presence than the introduction of depth. Therefore, the evaluation criterion presence will not be investigated further.

2.3 Experiment 2

In this experiment, the added value of 3D over 2D stills was investigated for viewing experience and naturalness. The experiment contained 'perfect' 2D and 3D material with no conversion or depth artefacts giving us full control over the stimulus material (in contrast to Experiment 1).

The goal was to investigate which evaluation term (viewing experience or naturalness) is most sensitive to depth in the absence of conversion or depth artefacts. In addition, our experiment served to calibrate the sensitivity of these evaluation concepts in relation to each other in terms of their response pattern to increasing levels of noise introduced in the 2D and 3D images. Yano (1991) performed an experiment quantifying the difference in image quality, sensation of power, and sensation of depth between *undistorted* 2D and 3D images in terms of a change in image size. In this experiment, the sensitivity to depth of the concepts naturalness and viewing experience was calibrated trying to quantify the potential stereoscopic advantage in terms of dB noise-level (*distorted*).

2.3.1 Method

Design

The experiment had a mixed design with Image (4 images), Dimension (2D, 3D) and Noise (6 levels) as within subject factors, and the two different evaluation concepts (naturalness and viewing experience) tested between subjects.

Observers

Thirty observers from a research environment were invited to participate in the experiment. Twenty observers participated in the viewing experience experiment and ten observers participated in the naturalness experiment. All observers had a visual acuity of ≥ 1 (as tested with the Landolt-C test) and good stereo vision <30 seconds of arc (as tested with the Randot stereo test).

Equipment

The 2D and 3D images were shown on the 20″ Philips multi-view autostereoscopic display, described in Chapter 1, section 1.3.3. The viewing distance was 150 cm. Nine different views were recorded using nine different cameras. The 2D situation was simulated by implementing the middle view (view five) into all nine views. In this case, the observer always perceives the same image on both eyes, resulting in a 2D percept.

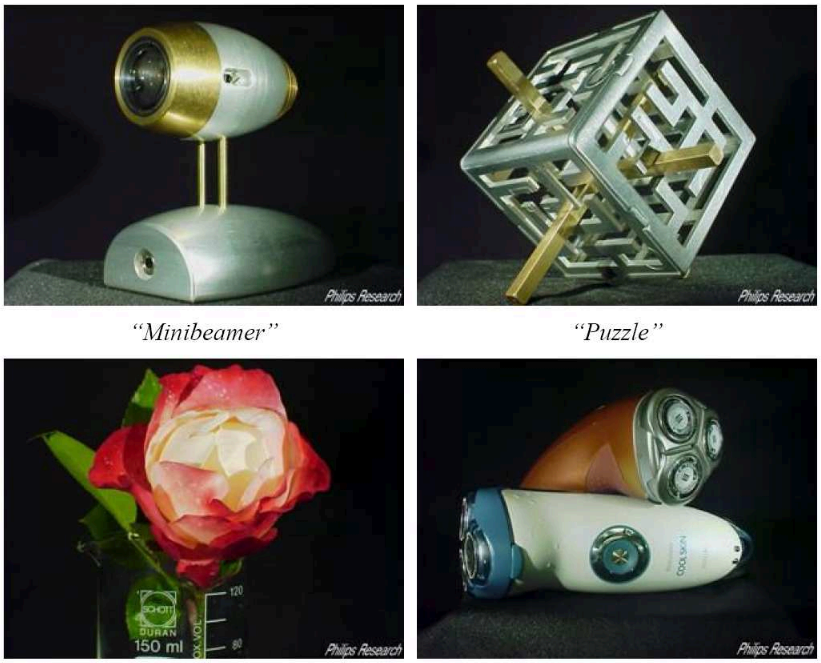
Stimuli

The image material used in this experiment consisted of four still scenes, *Minibeamer*, *Puzzle*, *Rose* and *Shaver*, recorded with a nine-camera set-up. The advantage of recording all the views with nine cameras instead of converting a 2D image into 9 views, is that all required information is available and no distortions due to limited depth information are introduced in the 3D material. Displaying the nine views on the multi-view auto-stereoscopic display resulted in a 3D percept of the image because each eye receives a different view with a different perspective. The 2D situation was simulated by implementing the middle view (view five) into all nine views. In this case, the observer always perceives the same image on both eyes, resulting in a 2D percept. The middle view (camera five) of each image is shown in Figure 2.3.

Since our main goal was to quantify the added value of depth through the concepts viewing experience and naturalness in terms of the affordable loss in quality, an appropriate image distortion had to be chosen. To avoid effects of image content, additive noise was chosen as the introduced artefact. The visibility of artefacts like for instance blurring, blocking and ringing depends on image content. Additive noise however seems to manifest itself in the same way over many different systems. Image independent noise can be described by an additive noise model, where the resulting image *f(i,j)* is the sum of the true image *s(i,j)* and the noise *n(i,j)*. The model is shown in Equation 2.1.

$$f(i,j) = s(i,j) + n(i,j) \tag{2.1}$$

The noise is modeled with a zero-mean ($\bar{x}$) Gaussian distribution described by its standard deviation (σ), or variance (σ^2). This means that each pixel in the noisy image is the sum of the true pixel value and a random, Gaussian-distributed noise value. The additive noise is evenly distributed over the frequency domain (i.e., white noise). The white Gaussian noise impairment was implemented using the Matlab image noise



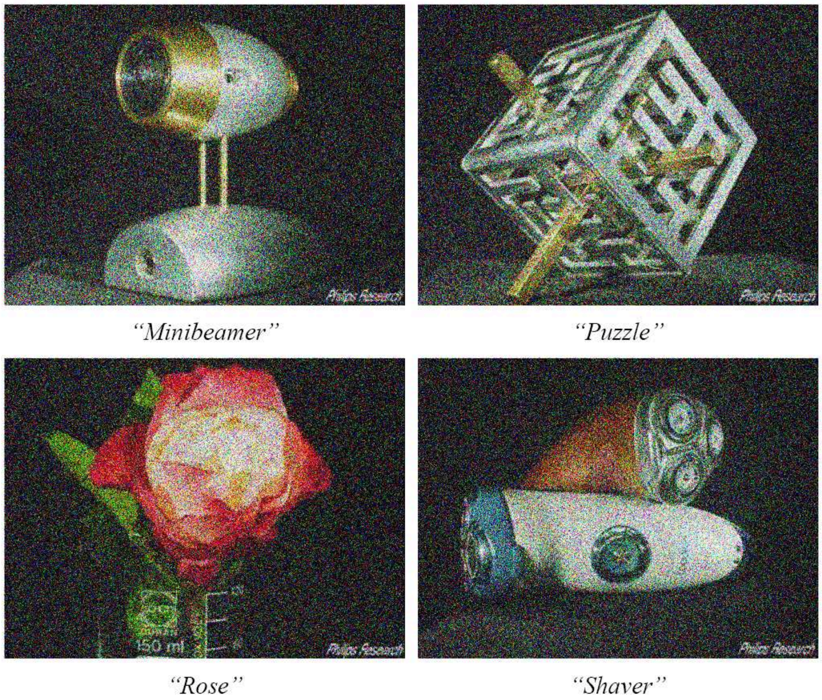
"Minibeamer"

"Puzzle"

"Rose"

"Shaver"

Figure 2.3: The four panels show the original scenes *Minibeamer*, *Puzzle*, *Rose* and *Shaver*.



"Minibeamer" *"Puzzle"*

"Rose" *"Shaver"*

Figure 2.4: Noise-impaired scenes (17 dB) *Minibeamer*, *Puzzle*, *Rose* and *Shaver*.

filter with five levels of noise ($\bar{x}$ = 0, σ^2 = 0.00125, 0.0025, 0.005, 0.01, 0.02). An increasing σ^2-parameter produced more noise in the images. Figure 2.4 shows the four scenes with additive noise ($\bar{x}$ = 0 and σ^2 = 0.02).

Procedure

The experiment consisted of two sessions: one for measuring viewing experience and one for measuring naturalness. In both sessions exactly the same set-up was used. The observers were given a brief instruction about the experiment on paper. Any remaining questions were answered and subsequently a short training session was conducted. The training session allowed the observers to get used to the setting as well as the tasks. In the training, six still images were presented with different noise levels, including the extremes used in the actual experiment. The rating scale for viewing experience and naturalness was labeled with the adjective terms [bad]-[poor]-[fair]-[good]-[excellent] according to the ITU (2000a) recommendation on subjective quality assessment. Observers were free to mark their assessment anywhere on the continuous rating scale. The order in which the images appeared was randomized throughout the experiment and each image was evaluated twice. The images were displayed for 10 seconds followed by a grey field for 3 seconds. In total, 20 observers were asked to indicate viewing experience for 4 (images) x 6 (noise impairment levels) x 2 (2D and 3D) x 2 (repetition) = 96 images. Exactly the same set-up was used for the naturalness ratings, only this session was done by 10 other observers. The lighting conditions of the room were constant for all observers and the level of light in the room was 25 lux, measured perpendicular to the display in the direction of the viewer.

2.3.2 Results

Viewing experience

Figure 2.5 shows the mean ratings for viewing experience averaged over the four images. On the x-axis the different noise levels are presented (increasing noise along the x-axis). The y-axis represents the averaged values for viewing experience from bad to excellent. The two lines in the figure represent the dimensions 2D and 3D. Error bars reflect the standard error of the mean.

A three-way repeated measures ANOVA (with Noise, Image and Dimension as factors) was carried out on the raw subjective ratings to test the main effects and interactions for statistical significance. The results revealed significant main effects of Image (F(3,17) = 6.413, $p<.01$), Dimension (F(1,19) = 5.251, $p<.05$) and Noise (F(5,15) = 46.521, $p<.001$) on the viewing experience ratings. No significant interactions between Image, Dimension and Noise were found. Figure 2.5 clearly shows the main effect of a decreasing viewing experience with increased noise level for both 2D and 3D images. The viewing experience of 3D images is rated systematically higher than the viewing experience of 2D images for all noise levels, explaining the main effect of Dimension. The main effect of Image was mainly caused by different parallel shifts in the four images, but the main effects of Noise and Dimension were clearly visible in all images.

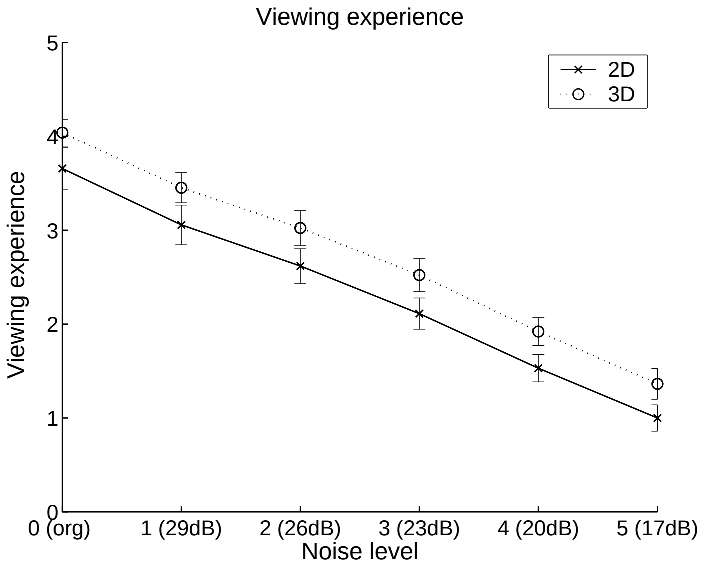


Figure 2.5: Viewing experience ratings averaged over all scenes. The x-axis represents the original image (org) and 5 noise-impaired images (PSNR) and the y-axis represents the subjective ratings for viewing experience. The lines in the figure represent Dimension (2D and 3D).

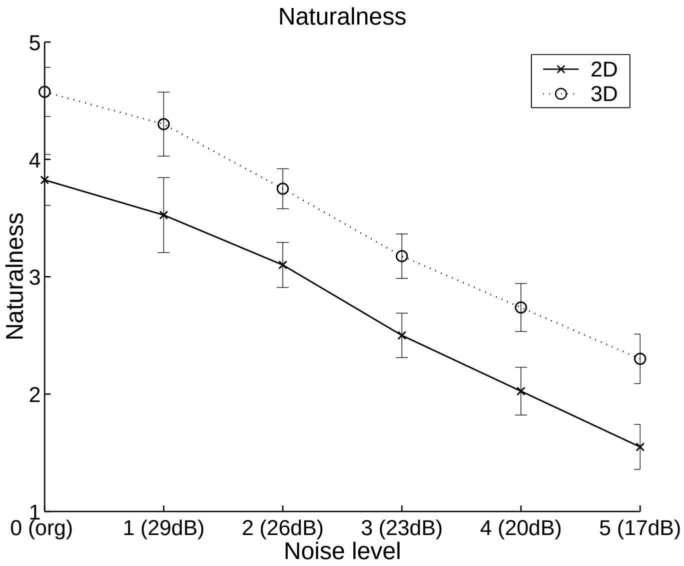


Figure 2.6: Naturalness ratings averaged over all scenes. The x-axis represents the original image (org) and 5 noise-impaired images (PSNR) and the y-axis represents the subjective ratings for naturalness. The lines in the figure represent Dimension (2D and 3D).

The difference in viewing experience between 2D and 3D is equivalent to a change in noise level of around 2 dB. Thus, 3D images with 2 dB more noise than their 2D counterparts result in the same viewing experience. So, the evaluation term viewing experience takes into account the addition of the stereoscopic cue to the total depth percept, as this is the primary difference between the 2D and 3D images.

Naturalness

Figure 2.6 shows the mean ratings for naturalness averaged over the four images. On the x-axis the different noise levels are presented (increasing noise along the x-axis). The y-axis represents the averaged values for naturalness from bad to excellent. The two lines in the figure represent the dimensions 2D and 3D. Error bars reflect the standard error of the mean.

A three-way repeated measures ANOVA (with Noise, Image and Dimension as factors) was carried out on the raw subjective ratings to test the main effects and interactions for statistical significance. The results revealed only significant main effects of Dimension (F(1,19) = 9.448, p<.013) and Noise (F(5,15) = 16.285, p<.004) on the naturalness ratings. No significant interactions between Image, Dimension and Noise were found. Figure 2.6 clearly shows the main effect of a decrease in naturalness with increasing noise level for both 2D and 3D images. The naturalness of 3D images is rated higher than for 2D images for all noise levels explaining the main effect of Dimension. The difference in naturalness between 2D and 3D is around 4 dB, when expressed in an equivalent difference in noise level.

Figure 2.5 and Figure 2.6 both show that noise considerably decreases the viewing experience and naturalness ratings both for 2D and 3D. Furthermore, both figures show a higher score for the 3D-mode than the 2D-mode, which implies that both viewing experience and naturalness take into account the added value of depth. The difference between 2D and 3D is larger for naturalness than for viewing experience, which implies that naturalness appears to be more sensitive to the addition of depth than viewing experience. The fact that the difference between 2D and 3D ratings remains constant over all the noise levels implies that the perceived depth is independent of the noise level.

2.3.3 Discussion

Our results show that both viewing experience and naturalness are sensitive image evaluation concepts when it comes to measuring the added value of stereoscopic depth using impaired 2D and 3D images. Earlier studies demonstrated that when observers are asked to rate image quality in impaired stereoscopic images, the added value of depth is hardly taken into account, if at all. However, when asking observers to assess viewing experience or naturalness, they do not only assess the level of impairment (in our case, the induced noise level), but also other aspects in the image, such as depth, which is illustrated by the fact that there are two distinctive lines for the assessment of 2D and 3D images. So, the added value of depth is taken into account when observers are assessing viewing experience, and even more so when they are assessing naturalness (see Figure 2.5 and 2.6). The results of the three-way repeated measures

ANOVA tests show that both Noise and Dimension significantly affect viewing experience and naturalness. For viewing experience also Image had a significant influence (vertical shift of the 2D and 3D line), but the added value of depth as measured by viewing experience was clearly recognized in all four images.

The method applied to quantify the added value of depth expressed in noise level yields an appropriate and useful measure. The potential stereoscopic advantage can thus be quantified in terms of dB noise-level. The difference in viewing experience and naturalness between 2D and 3D images expressed in noise level is, respectively, 2dB and 4dB. In other words, more noise is allowed in 3D images (respectively 2 dB and 4 dB) for an equal viewing experience and naturalness of 2D and 3D images.

The results in Figure 2.5 and 2.6 show a remarkably linear and thus predictable behavior, while being quite stable (low error) within the chosen stimulus set. Apparently observers are well capable of assessing the image impairment and added value of depth in the range used in this experiment.

Thus, quantifying naturalness or viewing experience by means of introducing a controlled impairment, such as noise, and expressing the results in units of this impairment yields a sensitive and reliable metric.

2.4 Conclusion

In sum, the experiments in Chapter 2 were performed to explore and determine which evaluation criterion is most appropriate to assess the performance of 3D-display systems. The explorative study in experiment 1 is based on a realistic set of 3D stimuli generated by new 2D-3D conversion techniques (not yet optimal). Results show that viewing experience and naturalness have the most discriminating power between the various 2D-3D conversion algorithms. However, the disadvantage of this experiment is that there was no experimental control over the stimuli at all.

Therefore, Experiment 2 was performed including both 'perfect' 2D and 3D material with no conversion or depth artefacts. Manually, several noise degradations were added making the added value of depth quantifiable in terms of units of noise impairment. Results show that the difference in viewing experience and naturalness between 2D and 3D images expressed in noise level is, respectively, 2dB and 4dB. From this experiment, it is concluded that naturalness is the most sensitive metric measuring the added value of 3D and will be used in Chapter 3 and Chapter 4.

In Chapter 3 and 4 more insight is gained in the behavior of the concept naturalness for 2D (JPEG) and 3D (crosstalk) artefacts. Also new knowledge and understanding will be build up about the behavior of 2D and 3D artefacts in 3D image material. The focus point of both chapters is the direct comparison between the traditional image quality concept and the naturalness concept.

Chapter 3

Perceived quality of JPEG coded 3D images

Abstract

This chapter describes two experiments to investigate the effects of symmetric and asymmetric JPEG coding and camera base distance on several evaluation criteria. The first experiment presents results on the effects of camera-base distance and JPEG-coding on overall image quality, perceived depth, perceived sharpness and perceived eye-strain. Results show that an increase in JPEG coding artefacts has a negative effect on image quality, sharpness and eye-strain but has no effect on perceived depth. An increase in camera-base distance increases perceived depth and reported eye-strain but has no effect on image quality and perceived sharpness. It is concluded that the added value of depth is not taken into account when using the image quality concept. Results on symmetric and asymmetric JPEG coding of the left- and right-eye view shows that the relationship between perceived image quality and average bit-rate is not straightforward. In some cases, image quality ratings of a symmetric coded pair can be higher than for an asymmetric coded pair, even if the averaged bit rate for the symmetric pair is lower than for the asymmetric pair. In experiment 2 the image quality concept is compared with the naturalness concept for different JPEG coding levels and camera base distances. Results show that naturalness weighs more equally the visibility of the distortion as well as the added value of depth in contrast to image quality.

[0] This chapter is based on Seuntiëns et al. (2005b)

3.1 Introduction

The transmission and storage of 3D image material involves a large amount of data due to the multiple views needed for 3D viewing. Therefore, recent research focuses on realizing new 3D coding and transmission standards to obtain savings in bandwidth and storage capacity. Developing new flexible formats is of particular relevance in the case of 3D HDTV, where a single uncompressed HDTV channel may cost up to one Gbit/s transmission bandwidth, or in the case of 3D video transmission over low-bandwidth transmission channels, such as the Internet (Johanson, 2001). The same compression techniques used in two-dimensional image material can also be applied independently on the left and right view of a stereoscopic image pair. Image compression may compromise perceived image quality however, through loss of detail and the introduction of compression artefacts such as blockiness, blur, or ringing. In order to ensure that the applied compression algorithms and levels still yield perceptually acceptable results, subjective testing using human viewers has been the only accurate method to date for assessing compressed stereoscopic video systems.

In this chapter, previous research on perceived image quality, perceived depth, perceived sharpness and perceived eye-strain in relation to 3D television is discussed first. Subsequently, in Experiment 3 the camera-base distance (B) and JPEG coding level are manipulated and the effects on perceived quality, sharpness, depth and eye-strain are measured to obtain a better understanding of the behavior of a 2D impairment in combination with different camera-base distance settings. The JPEG coding was applied in a symmetric (both views have the same compression ratio) and asymmetric (both views have a different compression ratio) way to investigate the concept of mixed JPEG coding. Experiment 4 investigates whether the added value of depth in JPEG impaired images (2D distortion) can be measured using the naturalness criterion.

3.1.1 Asymmetric coding and perceived image quality

Based on theories of binocular suppression, it is assumed that the binocular percept of a stereo image pair is dominated by the high quality component (Levelt, 1965). Thus, theoretically, when one image of the stereo pair is compressed such that a high quality is maintained, the other view can be coded more heavily without introducing visible artefacts in the binocular percept. The mixed resolution concept was introduced by Perkins (1992). Mixed resolution coding assumes that the binocular percept is not affected when one view is of high quality and the other view of lower quality. Perkins (1992) applied low-pass filtering (introducing blur) as compression algorithm resulting in a high-resolution and a low-resolution image for each view of a stereo image pair. The author concludes that mixed-resolution coding is easy to implement, and the reduction of the bit rate is significant with respect to a system that employs no coding.

Stelmach and Tam (1998) and Tam et al. (1998) applied a different compression ratio on the left- and right-eye views of a stereoscopic sequence using MPEG-2 (introducing blockiness) and low-pass filtering (introducing blur). The results showed that the subjective image quality of a stereo sequence was approximately the average of the monoscopic quality of the left- and right-eye images when MPEG-2 coding was used. Subjective image quality of an asymmetric low-pass filtered stereo sequence was dominated by the high quality component.

Meegan et al. (2001) studied the binocular combination of asymmetric blur and blockiness impairment images. In the case of asymmetric blur-impaired images the binocular percept was dominated by the high quality component. The binocular percept of the asymmetric MPEG-2 impaired images was approximately the average of the two monoscopic components. From this research it can be concluded that the success of asymmetric coding depends on the type of coding artefacts.

3.1.2 Perceived depth

The use of disparity information produces a compelling sense of depth, which defines the added value of stereoscopic TV. IJsselsteijn et al. (1998b) investigated the perception of depth and the naturalness of depth when viewing stereoscopic image material. As soon as binocular disparity was introduced, the ratings of perceived depth and naturalness of depth increased. Research of Westheimer and McKee (1980) documented a larger decrease in stereo acuity with asymmetric blur that with symmetric blur. Stelmach et al. (2000) investigated the effect of spatial and temporal low-pass filtering on perceived depth. The results indicate that spatial low-pass filtering has no effect on perceived depth. Temporal low-pass filtering produced poor image quality but the sensation of depth was relatively unaffected. An explanation is that low-pass filtering leaves the low spatial frequencies, that are sufficient to carry the disparity signal, unaffected. In their studies, depth shows a weak correlation with image quality and sharpness. These results suggest that depth is a dimension of perceptual experience that is largely independent of sharpness and overall image quality. This appears to be at variance with the results of IJsselsteijn et al. (2000b) where perceived image quality could be expressed as a function of perceived depth and experienced eye-strain. These results were obtained using uncompressed images that varied in terms of camera-base distance, convergence distance, and focal length. A number of stimuli contained excessive disparities, thus making it likely for observers to base their quality judgements on different image attributes than with the Stelmach et al. (2000) study.

3.1.3 Perceived sharpness

Perceived sharpness in stereoscopic images can be affected by several parameters, e.g., camera defocus, coding, or binocular disparity. Berthold (1997) reported that stereo images with different degrees of Gaussian blur were perceived sharper than non-stereo images. Tam et al. (1998) on the other hand found that the observers rated the MPEG-2 coded stereo and non-stereo images equally sharp or the stereo images even slightly less sharp. A high correlation was found between sharpness and image quality in both studies. Stelmach et al. (2000) investigated the effect of mixed-resolution on perceived sharpness and concluded that spatial low-pass filtering gives an acceptable sharpness. Sharpness was biased towards the view with the greater spatial resolution. On the other hand, temporal low-pass filtering produced very poor images with blurred edges. Meegan et al. (2001) confirmed these findings in an experiment measuring the visibility of blur in asymmetric processed stereo images. When the lower quality view contained blur artefacts, the higher quality view was over-weighted by the visual system.

3.1.4 Perceived eye-strain

Many studies report a clear preference for stereoscopic images over non-stereoscopic ones. However, viewing stereo images can be more fatiguing than viewing conventional two-dimensional images. Because eye-strain can be extremely annoying in stereoscopic displays, it is important to have an understanding of its subjective magnitude and impact on the user. IJsselsteijn et al. (2000b) investigated the effect of stereoscopic filming parameters and display duration on the subjective assessment of eye-strain. The averaged results of the eye-strain ratings show a clear linear increase with increasing disparities. There was no significant effect of display duration on the eye-strain scores, but the display durations were relatively short (1-15 seconds). Mitsuhashi (1996) found that observers experienced more eye-strain for binocular vision than with the conventional television picture, using an objective measure known as the critical flicker frequency (CFF). The critical flicker frequency is the highest frequency at which a particular person still sees flicker. At any higher frequency, the subject sees a steady light source. Watching stereoscopic television caused a significant CFF decrease within 30 minutes. It was also found that the CFF decreases are related to a subjective feeling of eye-strain. Okuyama (1999) evaluated visual fatigue with visual function testing (objective measure) and interviews (subjective measure). Visual function testing showed a mismatch between convergence and accommodation. The interviews reported more eye pain, an 'alien feeling' in the eyes and eyes filled with tears. Both evaluations show an increase in visual fatigue. Kooi and Toet (2004) concluded that disparity, crosstalk and blur are the most important parameters that cause eye-strain.

3.2 Experiment 3

In sum, it seems that the mixed resolution concept is appropriate for stereoscopic transmission although the quality of the binocular percept will depend on the type of distortion. Previous experiments on asymmetric coding, however, did not control camera-base distance. Since it is possible that different camera separation settings influence the visibility of coding artefacts, an experiment was performed that was aimed at investigating the effects of asymmetric/symmetric coding while using different levels of camera separation. The reason for varying camera separation is that larger separations lead to larger differences between the left and right eye view in terms of view perspective and image content at the borders of the image (image information appears and disappears with varying camera separation). Thus, image coders (such as JPEG or MPEG) produce different artefacts in the left and right eye view in terms of intensity, position and shape when varying camera separation. So, matching the left and right eye view (based on corresponding points in the images) may result in a different 3D percept or different perception of the artefacts. Secondly, there is full control for a possible effect of eye dominance, where images presented to the dominant eye would potentially contribute more to the overall stereoscopic percept than images presented to the non-dominant eye. Further, the study of the effects of camera-base distance and compression (JPEG coding) on perceived depth, sharpness, image quality and eye-strain is extended.

3.2.1 Method

Design

A mixed design experiment was performed with Image (2 images), Camera base distance (0, 8, and 12 cm) and JPEG coding (16 symmetric/ asymmetric combinations) as within subject factors, and four different attributes (perceived sharpness, perceived depth, perceived image quality and perceived eye-strain) tested between subjects.

Observers

Forty non-expert observers were paid to participate in this experiment. The observers, mostly students, came from the same age group (18-27 years old). All observers had a visual acuity of ≥ 1 (as tested with the Landolt-C chart), good stereo vision <30 seconds of arc (as tested with the Randot stereo test) and good color vision (as tested with the Ishihara test). Eye dominance and inter-pupillary distance were also measured. The Finger-Point method was used to determine eye dominance. Observers pointed naturally at an object with both eyes open and the face square to the object. The eyes were closed alternately. When the domi-

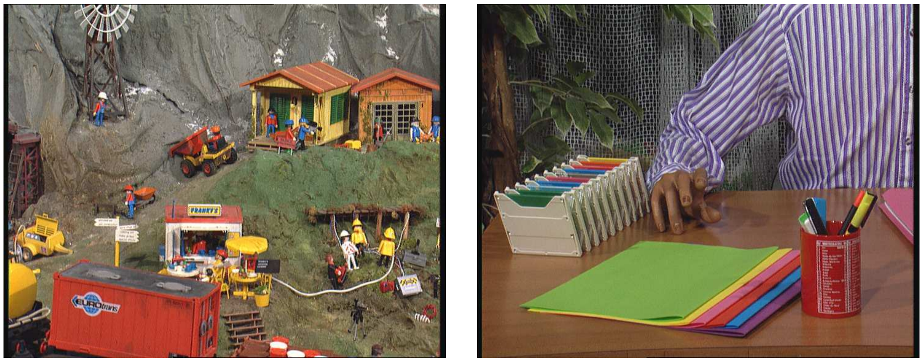


Figure 3.1: The left panel shows the original of the scene *Playmobiles* and the right panel shows the original of the scene *Bureau*.

nant eye is closed the finger appears to jump away from the original location. Thirty-one observers were right eye dominant and the average inter-pupillary distance was 6,2 cm, which is slightly below the population average of 6,3 cm (Dodgson, 2004).

Equipment

An AEA-Technology (AEAT) stereoscopic display was used in this experiment. For a detailed description see Chapter 1, section 1.3.3. The viewing distance was 80 cm.

Stimuli

The image material used in this experiment consisted of two still color scenes, *Playmobiles* and *Bureau*, that varied in camera-base distance (B) and compression ratio. The scene *Playmobiles* consists of a colorful toy landscape with mountains in the background and numerous *Playmobiles* in the foreground. The scene *Bureau* consists of a tailor's dummy sitting behind a desk with some office equipment. The original scenes are shown in Figure 3.1.

The scenes were recorded in a studio set-up, using a professional stereoscopic studio camera in a toed-in configuration. For each scene, lens focal length and convergence distance of the cameras to the scene were fixed to 20 mm and 1.30 m, respectively. Each scene was recorded at three different camera-base distances, namely 0 cm (i.e., monoscopic), 8 cm and 12 cm. The scenes were created in the European DISTIMA project and were kindly provided to us by CCETT in France.

An increase in camera-base distance results in an increase in disparity values and thus perceived depth, while the size of the objects and the field of view remains constant. A camera-base distance of 0 cm introduces no disparity between the left and the right image and thus no perceptible stereoscopic depth, while depth is highly noticeable with a camera-base distance of 8 cm and 12 cm.

The stimulus set contained the original, uncompressed version of each scene and three JPEG coded versions. The Baseline Sequential JPEG compression software package of the Independent JPEG Software Group[1] with default quantization table was used to generate for each scene different versions at various compression rates. The compression rate was determined by the 'Q-parameter'. Images with a high compression ratio were obtained by low 'Q-values', and therefore contained the most conspicuous distortions. The JPEG 'Q-parameters' used in this experiment were Q30, Q20 and Q10 for the scenes *Playmobiles* and *Bureau*. The coding levels were chosen carefully based on the visibility of artefacts. Stereo image pairs were formed by symmetric or asymmetric coding of the left and right eye images. In a symmetric stereoscopic image pair the same compression ratio is applied to the left- and right-eye view. In asymmetric coding the compression ratio of the two views is different. Table 3.1 gives the bytes per pixel (bpp) of the symmetric and asymmetric image pairs, which is the averaged bpp of the left and right eye views.

In order to test a potential effect of eye dominance, each combination of bit-rates was presented to both eyes. Each JPEG coding level was displayed an equal number of times to each eye (left and right) and all combinations of the coding levels were presented once. Thus in total, 2 scenes, 3 camera-base distances and 4x4 coding levels were used. This resulted in a stimulus set of 2x3x16 = 96 images.

[1] http://www.ijg.org

Table 3.1: Bytes per pixel (bpp) for the symmetric and asymmetric image pairs.

	bpp stereoscopic image			
	left eye			
right eye	Org	Q30	Q20	Q10
Org	3.00	1.55	1.54	1.53
Q30	1.55	0.11	0.10	0.08
Q20	1.54	0.10	0.08	0.07
Q10	1.53	0.08	0.07	0.05

Procedure

A set of 96 stereoscopic images was randomized and presented sequentially. Observers were asked to rate according to the single stimulus scaling method. Each attribute was rated by a different group of 10 observers and each observer rated only one attribute. The perceived overall image quality was rated on a categorical scale from 1 up to 5 corresponding to 1 for bad image quality and 5 for excellent image quality. The scale was labeled with the adjective terms [bad]-[poor]-[fair]-[good]-[excellent] according to the ITU (2000a) recommendation on subjective quality assessment. Perceived sharpness was rated on a numerical scale from 1 up to 5. The least sharp image corresponded to 1 and the sharpest image to 5. Experienced eye-strain was rated on an impairment scale from 1 up to 5 with the adjective terms [very annoying]-[annoying]-[slightly annoying]-[perceptible, but not annoying]-[imperceptible] according to the ITU (2000a) recommendation. Perceived depth was rated on a numerical scale from 1 up to 5. The image with no perceived depth was to be rated 1 and the image containing most perceived depth was to be rated 5. No adjectives were used on the depth and sharpness scale.

The stimulus set of 96 images was judged for each attribute in two subsessions, containing 48 stimuli each, with a small break in between. Each subsession of 48 images took approximately 20 minutes. Before the experiment started the observers were asked to read the instructions explaining the task and attribute they had to rate. After that the observers participated in a trial of 16 stimuli to get acquainted with the stimulus set and the range of variations in image parameters (camera-base distance, compression ratio).

To check whether there were any negative side effects as a result of watching 45 minutes of stereoscopic images, observers were asked to fill out a symptom checklist before and after the complete experiment. The symptom checklist consisted of 6 items which observers had to rate namely (1) General discomfort, (2) Fatigue, (3) Headache, (4) Eye-strain, (5) Difficulty focussing and (6) Blurred vision.

3.2.2 Results

A way to analyze data obtained by numerical category scaling experiments is to transform the data into an interval scale assuming a psychologically linear continuum. Thurstone's law of categorical judgement can be used for such a transformation (Thurstone, 1927). The Thurstone model assumes that the attribute strength is measured on an internal psychological scale, i.e., an interval scale with Gaussian noise distribution. For all observers, the raw category scaling data obtained in the experiment were transformed to a Thurstone scale using the software package ThurcatD (Boschman, 2000). As input, the program ThurcatD needs frequency distributions per category for each stimulus that was presented in the experiment. From the input frequency distributions of ratings over the categories, ThurcatD calculates the stimulus scale values in standard deviation units and, also, the interval borders that define the intervals on the psychometrical scale. Equal distances on the scale correspond with equal differences in the percept judged because the Thurstone scale is a true interval scale. The Thurstone values are scaled back to the original scale using a linear transform function.

Eye dominance

The effect of eye dominance was tested for the asymmetric combinations between ratings of left-eye dominant (averaged over 9) and right-eye dominant (averaged over 31) observers for each attribute. As tested with a Wilcoxon Signed Rank Test, which is a conservative test, there were no significant differences between the ratings of left-eye dominant and right-eye dominant observers for all asymmetrical combinations and attributes. Therefore, the data of the left- and right-eye dominant observers was pooled. Next, the effect between the ratings of the asymmetric combinations and their reverse version per attribute was tested and also found no significant difference and therefore pooled the data.

Perceived Image Quality

Figure 3.2 shows the Thurstone values for image quality and the standard errors for the scenes *Playmobiles* and *Bureau*. On the x-axis the symmetrical and asymmetrical coding combinations of the stereoscopic images are presented with increasing bit-rate to the right. The y-axis represents the estimated Thurstone scale values for each data point scaled back to the original scale. The ThurcatD analysis results show a good model fit (χ^2=93.89, p=0.2879 for *Playmobiles* and χ^2=89.43, p=0.4076 for *Bureau*). This result means that equal distances on the scale correspond with equal differences in the percept judged. The three lines in the figure represent the 3 camera-base distances 0, 8 and 12 cm.

The quality scores show an increasing trend with increasing bit-rate for both scenes. Two quality dips are clearly visible in the *Bureau* scene. At these points one of the two views of the stereoscopic image is coded with a JPEG compression factor of Q10. This image contains a lot of annoying artefacts (mostly blockiness) which can hardly be reduced by the high quality component of the stereoscopic image pair. The *Bureau* scene contains some homogeneous areas were the blockiness artefact is more visible. The quality dip in the *Playmobiles* scene also occurs at Q10, but is smaller because this image contains less homogeneous areas. These quality dips indicate that the relationship between perceived image quality and average bit-rate is not straightforward, at least for stereoscopic images. For example, image quality ratings of a symmetric coded pair (Q30_Q30) can be higher than for an asymmetric coded pair (Q10_org), even if the averaged bit rate for the symmetric pair (Q30_Q30) is lower, than for the asymmetric pair (Q10_org). An increase in camera-base distance (B) has almost no effect on perceived image quality. The *Bureau* scene shows almost no differences between the three camera-base distances and the quality judgements of the *Playmobiles* scene differ only slightly.

The weighting of bi-ocular inputs (left and right view) in binocular combination was also investigated. When the left and right view of an asym-

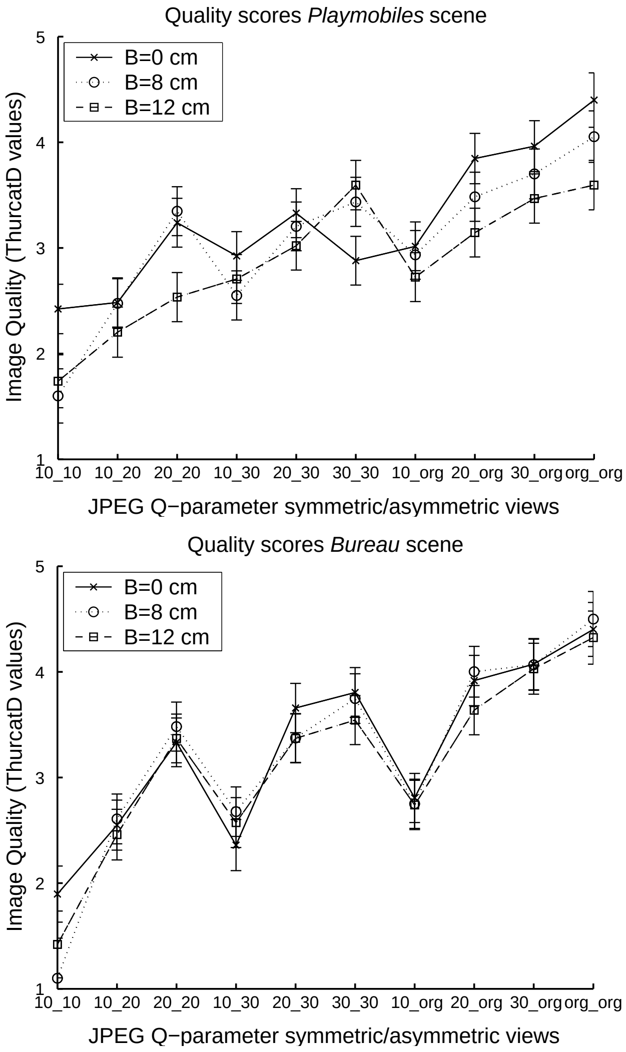


Figure 3.2: Thurstone values and error bars for image quality for the scenes *Playmobiles* and *Bureau*. The x-axis represents the JPEG Q-parameter (increasing bitrate to the right) for the symmetrical and asymmetrical image pairs and the three lines in the figure represent the 3 camera-base distances (B).

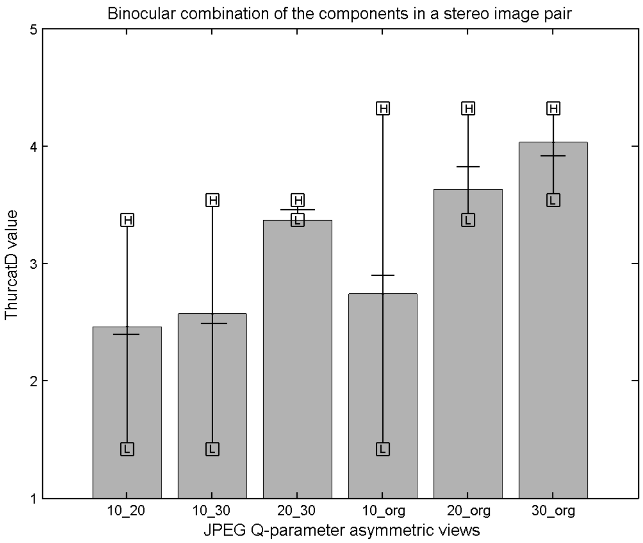


Figure 3.3: Binocular weighting of image quality for the *Bureau* scene with a camera-base distance of 12 cm. In the first case, labeled as 10_20 on the x-axis, the judged image quality of the symmetric image pairs are labeled as 'L' and 'H' corresponding to JPEG coding Q10_Q10 and Q20_Q20. The judged image quality of the asymmetric image pair is the height of the bar (Q10_Q20).

metric coding pair are viewed separately, the less compressed view would have a higher subjective image quality. On the other hand, the highly compressed view would have a lower subjective image quality, and JPEG coding artefacts are visible. In Figure 3.3, the image quality of the symmetric inputs (high 'H' quality e.g., Q20_Q20, and low 'L' quality e.g., Q10_Q10 of the views) are presented by the whiskers (endpoints) and the asymmetric combination (stereo view of the high and low quality inputs, e.g., Q10_Q20) is presented by the height of the bar. The image quality of the symmetric inputs is the bi-ocular combination of two images with the same compression ratio. Figure 3.3 presents the symmetric and asymmetric stereoscopic image pairs of the *Bureau* scene with a camera-base distance of 12 cm.

The results of Figure 3.3 show that the perceived image quality of the binocular combination was approximately the average of the bi-ocular high quality and low quality input. There are some differences between the scenes but these are small. The tendency towards the low quality input is a bit stronger in the *Bureau* scene than in the *Playmobiles* scene when Q10 is in the asymmetric image pair. This can be explained by the fact that the blockiness artefact is more visible in the *Bureau* scene at high compression rates because there are more homogeneous areas.

Perceived Depth

The Thurstone values for perceived depth and the error bars for the *Playmobiles* and *Bureau* scene are presented in Figure 3.4. The x-axis represents the symmetric and asymmetric coding combinations in increasing bit-rate. The y-axis represents the estimated Thurstone scale values for each data point scaled back to the original scale. The ThurcatD analysis results show a good model fit (χ^2=61.65, p=0.9820 for *Playmobiles* and χ^2=76.07, p=0.7924 for *Bureau*). This result means that equal distances on the scale correspond with equal differences in the percept judged. The three lines in the figure represent the 3 camera-base distances 0, 8 and 12 cm.

As expected, the perceived depth scores increased when the camera-base distance increased. The perceived depth between camera-base distance 8 and 12 cm increased less than between 0 and 8 cm. Furthermore, JPEG coding had no clear effect on perceived depth. For all JPEG compression levels and combinations, the perceived depth remains nearly the same for each camera-base distance.

Perceived Sharpness

In Figure 3.5 the Thurstone values for sharpness of the *Playmobiles* scene and the *Bureau* scene are given. The x-axis represents the symmetric and asymmetric coding combinations in increasing bit-rate. The y-axis represents the estimated Thurstone scale values for each data point scaled back to the original scale. The ThurcatD analysis results show a good model fit (χ^2=71.73, p=0.8814 for *Playmobiles* and χ^2=91.93, p=0.3381 for *Bureau*). This result means that equal distances on the scale correspond with equal differences in the percept judged. The three lines in the figure represent the 3 camera-base distances 0, 8 and 12 cm.

The results for perceived sharpness show great similarity with the per-

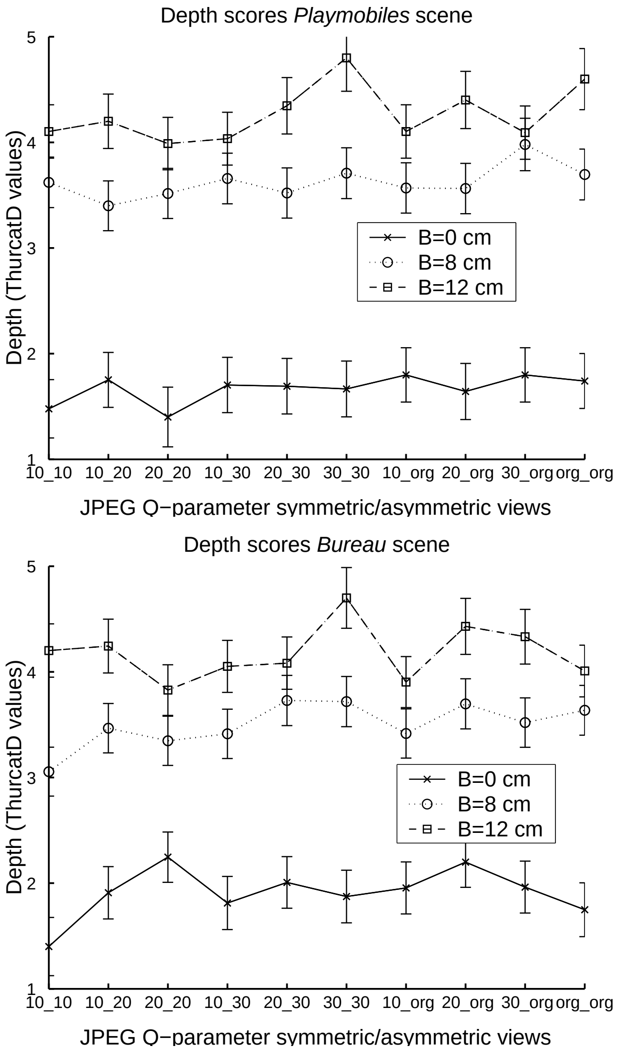


Figure 3.4: Thurstone values and error bars for perceived depth for the *Playmobiles* and *Bureau* scenes. The x-axis represents the JPEG Q-parameter (increasing bit-rate to the right) for the symmetrical and asymmetrical image pairs and the three lines in the figure represent the 3 camera-base distances (B).

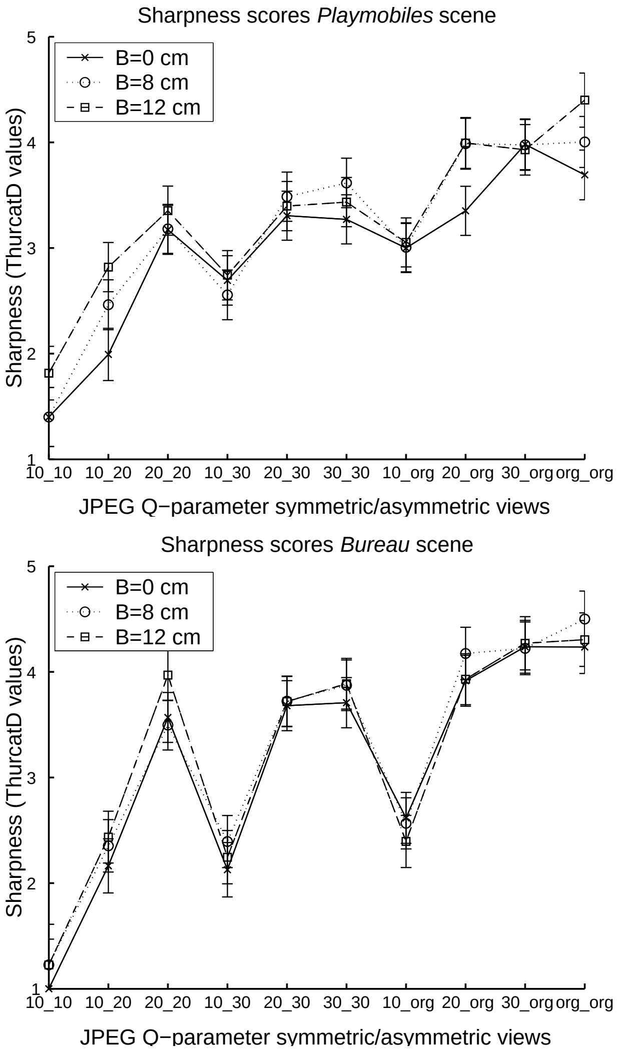


Figure 3.5: Thurstone values and error bars for sharpness for the *Playmobiles* and *Bureau* scenes. The x-axis represents the JPEG Q-parameter (increasing bit-rate to the right) for the symmetrical and asymmetrical image pairs and the three lines in the figure represent the 3 camera-base distances (B).

ceived image quality results. Perceived sharpness increased when the bit rate increased. Also in these figures the perceived sharpness scores dropped dramatically as soon as JPEG compression level Q10 was presented in one of the two views of the stereoscopic pair. The sharpness scores in the *Bureau* scene were approximately the same for the three camera-base distances. There were little differences visible in the *Playmobiles* scene between the three camera-base distances. So, introducing image disparity appears to have no effect on perceived sharpness in our stimulus set.

Perceived Eye-strain

Figure 3.6 represents the Thurstone scale values for the eye-strain scores of the observers. The x-axis represents the symmetric and asymmetric coding combinations in increasing bit-rate. The y-axis represents the estimated Thurstone scale values for each data point scaled back to the original scale. The ThurcatD analysis results show a good model fit (χ^2=66.91, p=0.9460 for *Playmobiles* and χ^2=75.38, p=0.8084 for *Bureau*). This result means that equal distances on the scale correspond with equal differences in the percept judged. The three lines in the figure represent the 3 camera-base distances 0, 8 and 12 cm.

The results show less annoyance with increasing bit-rate (less compression) and more annoyance with increasing camera-base distance. As in the quality and sharpness figures, there is an increase in reported eye-strain as soon as JPEG coding level Q10 is presented in one of the two views of a stereoscopic image pair.

Correlation between attributes

In this experiment high correlation coefficients were found between image quality and sharpness (R=0.93) and between image quality and eye-strain (R=0.76). No significant correlation was found between image quality and depth as is also obvious when comparing Figures 3.2, 3.4, and 3.5. A moderate negative correlation was found between depth and eye-strain because introducing more depth in the image leads to an increase in eye-strain. Table 3.2 shows the correlation coefficients (R) for all attributes.

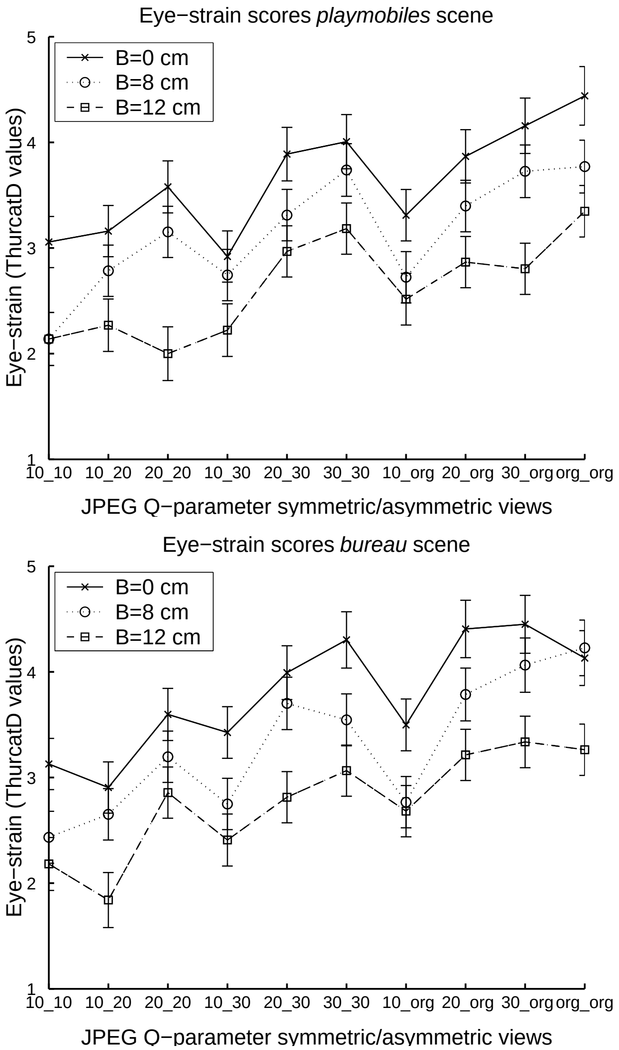


Figure 3.6: Thurstone values and error bars for eye-strain for the scenes *Playmobiles* and *Bureau*. The x-axis represents the JPEG Q-parameter (increasing bit-rate to the right) for the symmetrical and asymmetrical image pairs and the three lines in the figure represent the 3 camera-base distances (B).

Table 3.2: Correlation coefficients (R) for all attributes

	Quality	Sharpness	Depth	Eye-strain
Quality	1	0.93	-0.04	0.76
Sharpness	0.93	1	0.18	0.62
Depth	-0.04	0.18	1	-0.52
Eye-strain	0.76	0.62	-0.52	1

Symptom Checklist

Figure 3.7 shows the averaged results of the symptom checklist. The categories correspond with the labels on the x-axis. The y-axis shows the averaged scores (0 = none, 1 = slight, 2= moderate, 3 = severe) over 10 observers for each attribute. The figures are shown separately for each of the subjective ratings. This is done so as to visualise potential priming or sensitisation processes in relation to negative side effects as a consequence of having provided a certain attribute rating (e.g., eye-strain). In all figures a slight increase in symptoms can be observed. A Wilcoxon Matched-Pairs Signed Ranks test was used to reveal statistical significant differences (criterion p< 0.05) between the averaged checklist results before and after the experiment. The symptom checklist of the attribute quality shows a significant difference for the items headache (p=0.034) and eye-strain (p=0.034). No significant differences were found for the attributes depth and sharpness. The attribute eye-strain reveals significant differences for the items general discomfort (p=0.014), fatigue (p=0.034) and eye-strain (p=0.014). The items headache and difficulty focussing almost reached significance. The averaged ratings of the attribute eye-strain differed remarkably from the averaged ratings of the other attributes, although all observers saw the same stimuli set. A possible explanation maybe the direct association of the attribute eye-strain with the items on the symptom checklist, sensitizing observers to potential negative effects associated with viewing a stereo display.

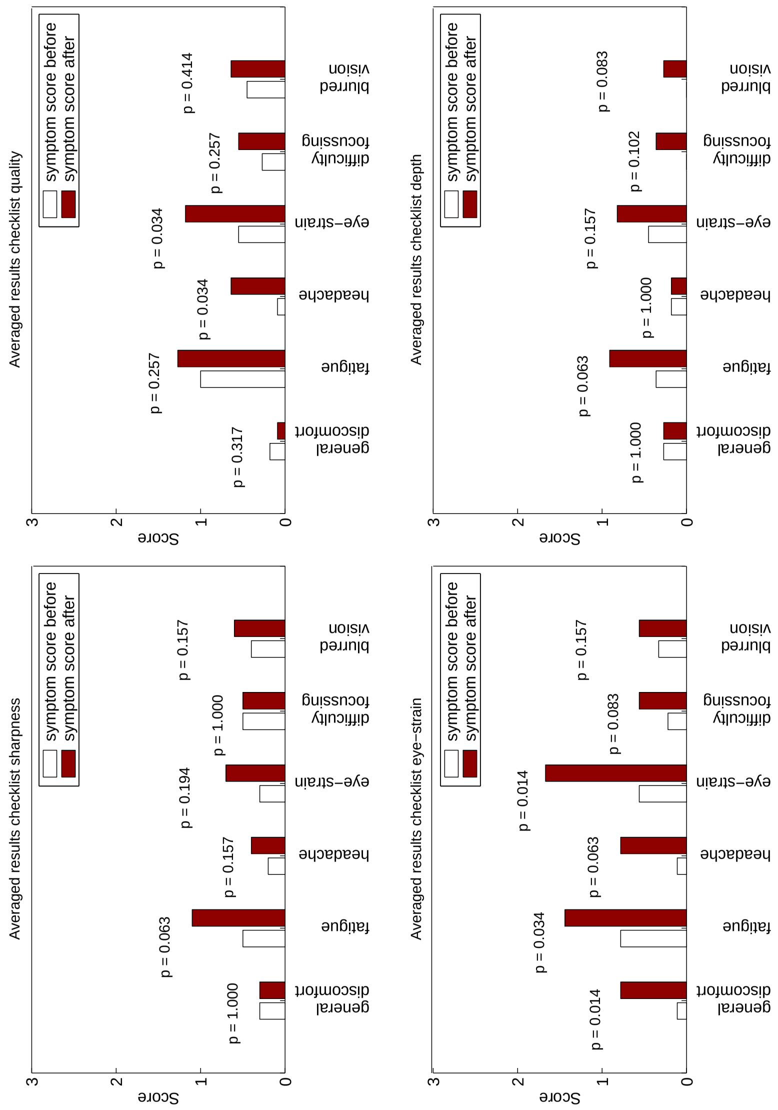


Figure 3.7: Averaged symptom checklists for all attributes. The x-axis shows the items (1) general discomfort, (2) fatigue, (3) headache, (4) eye-strain, (5) difficulty focussing and (6) blurred vision. The y-axis shows the averaged scores (0 = none, 1 = slight, 2= moderate and 3 = severe).

3.2.3 Discussion

The perceived image quality of monoscopic images (B=0 cm) was rated about the same as that of stereoscopic images (B=8 and 12 cm). The reason for this finding may be that the added value of perceived depth was not taken into account when judging image quality. This may be due to the used stimulus set and experimental paradigm. Stelmach et al. (2000) also found a low correlation between image quality and perceived depth using low-pass filtering. On the other hand, IJsselsteijn et al. (2000b) described an empirical relation between perceived depth, eye-strain and image quality for uncompressed stereoscopic images. The authors showed that an increase in image quality ratings could be attributed to an increase in perceived depth (when kept within natural bounds). However, quality judgements were attenuated by the eye-strain ratings, thus arriving at a simple stereoscopic image quality model for uncompressed images, describing quality as the difference between the added value of depth diminished by experienced eye-strain. In the current experiment, image artefacts were introduced which may have dominated the image quality assessment, making observers focus less on the depth dimension.

The results of this experiment show that an increase in JPEG coding decreases perceived image quality and sharpness and slightly increases perceived eye-strain. No effect of JPEG coding was found on perceived depth. Results of Stelmach et al. (2000) also showed that spatial low-pass filtering had no effect on perceived depth. In an earlier study, Stelmach and Tam (1998) showed that subjective image quality of a stereo sequence was approximately the average of the monoscopic quality of the left- and right-eye images. In this experiment, the stereoscopic image quality of the binocular combination was approximately the average for all disparities (B= 0, 8 and 12 cm). From the image quality figures, it can be concluded that there is almost no decrease in image quality when coding at org_org, org_Q30 and org_Q20. The quality drops heavily when one of the views is coded with a JPEG compression of Q10. This can be explained by the fact that the blockiness artefact is highly visible in Q10. These results are in line with Meegan et al. (2001) who also found an under-weighting of the high quality component when blockiness artefacts are present in the image. So, asymmetric coding is a valuable way to save bandwidth but one view must be of high quality (preferable the original) and the compression level of the coded view must be within an acceptable range (in this experiment Q20 is allowed). Also an interesting conclusion of this experiment is, that the relation between perceived image quality and average bit-rate is not always straightforward, at least for stereoscopic images. Thus, an increase in average bit-rate does not always result in an increase in perceived image quality. The image quality of Q30_Q30 is higher than the image quality of Q10_org, whereas the average bit-rate of Q30_Q30 is lower than that of Q10_org.

Camera-base distance had no significant effect on perceived image quality. In the *Playmobiles* scene there were slight differences and in the *Bureau* scene there were no differences. As expected, perceived depth increased with increasing camera-base distance. An increase in camera-base distance from 0 cm to 8 cm significantly increased perceived depth. The increase from 8 cm to 12 cm resulted in a smaller increase in perceived depth. Increasing camera-base distance appears to have no effect on perceived sharpness in our stimulus set. The last attribute, eye-strain, also showed an increase when camera-base distance increased. This was also found by IJsselsteijn et al. (2000b) where averaged results of the eye-strain ratings showed a clear linear increase with increasing disparities. Also the results of Kooi and Toet (2004) showed that increasing disparity is one of the most important parameters that cause eye fatigue.

The results of the symptom checklist show that the averaged ratings of the attribute eye-strain differed remarkably from the averaged ratings of the other attributes, although all observers saw the same stimulus set. This finding can be explained by a possible priming effect. Observers directly associated the attribute eye-strain with the items on the symptom list, sensitizing them to potential negative effects associated with viewing a 3D display.

This research showed that the added value of depth is not always or consistently taken into account when judging image quality, since no increase in image quality was found when depth was increased. Whether or not depth is taken into account when assessing image quality appears to depend rather heavily on the type of image impairments present in the stereoscopic pair. So, for the evaluation of 3D TV other concepts than image quality may be needed. The next experiment addresses the effect of JPEG coding on 2D and 3D image content comparing the traditional image quality criterion and the naturalness criterion. The results presented in Chapter 2 suggest that naturalness may be a more sensitive evaluation criterion to assess the added value of depth.

3.3 Experiment 4

The results of Experiment 3 showed that the image quality of 3D images is rated similarly to the quality of their 2D counterparts when 2D artefacts (JPEG-coding) are visible. This suggests that for the assessment of 3D content, image quality is mainly determined by the visibility of artefacts in the images and not so much by the added value of depth. It is expected that naturalness will take the added value of 3D into account, even when artefacts are visible in the images. Therefore, Experiment 4 is performed, determining image quality and naturalness of JPEG impaired images.

3.3.1 Method

Design

The experiment had a mixed design with Image (2 images), Camera base distance (0, 4, and 8 cm) and JPEG coding (5 levels) as within subject factors, and two different evaluation concepts (naturalness and image quality) tested between subjects.

Observers

Twenty observers participated in this experiment. Ten observers judged the image quality whereas the other ten observers judged the naturalness of the 3D scenes. The observers, mostly students and Philips employees, had a visual acuity >1 per eye (as tested with the Landolt C chart), a stereo vision <30 seconds of arc (as tested with the Randot Stereo Test) and no color deficiencies (as tested with the Ishihara test).

Equipment

The Screenscope$^{\mathrm{TM}}$(mirror stereoscope) was used to direct the left- and right-eye image of a side-by-side displayed stereo pair to the appropriate eye. For detailed information see Chapter 1, section 1.3.3.

Stimuli

The image material used in this experiment consisted of two still scenes, *Playmobiles* and *Bureau*, which are the same as in Experiment 3. For these scenes various levels of depth (0, 4, and 8 cm camera base distance) and JPEG coding (org, Q30, Q20, Q15, Q10) were simulated. The scenes were recorded in a studio set-up, using a professional stereoscopic studio camera in a toed-in configuration. For each scene, lens focal length and convergence distance of the cameras to the scene were fixed to 20 mm and 1.30 m, respectively. Each scene was recorded at three different camera-base distances, namely 0 cm (i.e., monoscopic), 4 cm and 8 cm. The scenes were originated in the European DISTIMA project and were kindly provided to us by CCETT in France. The spatial resolution of both scenes was 720x576 pixels per eye.

In total 2 (scenes) x 3 (depth levels)x 5 (JPEG levels) = 30 images were used in the stimulus set. The experiment started off with a training session of 6 stereoscopic images and also included one repetition measurement per stimulus, so in total 66 stereoscopic images were shown.

Procedure

The set of 60 stereoscopic images was presented sequentially to the observers in a random order. After rating an image a grey adaptation field was shown (3 seconds) before the next image appears. The images were rated according to the single stimulus scaling method. Each attribute (image quality and naturalness) was rated by 10 observers. A quality categorical scale ranging from 1 to 5 was used. On this scale 1 corresponds to a bad image quality/naturalness and 5 to an excellent image quality/naturalness. The scale was labeled with the adjective terms [bad]-[poor]-[fair]-[good]-[excellent] according to the ITU BT500-10 (2000). During the training session (6 stereoscopic images) the observers got acquainted with the stimulus set and its variation in image quality/naturalness.

3.3.2 Results

Image Quality

In Figure 3.8, the image quality ratings are shown for the *Bureau* and *Playmobiles* scenes. On the x-axis the JPEG coding levels are indicated and the y-axis represents the image quality in terms of ThurCatD scale values. The three lines in the figure represent the three camera base distances.

Figure 3.8 shows for both scenes that the image quality for 2D (i.e., 0 cm camera base distance) and 3D (i.e., 4 and 8 cm camera base distance) scenes is rated about the same. Observers anchored their image quality judgements on the most salient features (e.g., JPEG coding artefacts) while the added value of depth is not recognized. So, the evaluation term image quality does not take into account the added value of depth.

Naturalness

Figure 3.9 shows the naturalness ratings for the scenes *Bureau* and *Playmobiles*. The x-axis indicates the JPEG coding level and the y-axis represents the rescaled ThurcatD scale values for naturalness. The different camera base distances are represented by the three lines in the graph.

Figure 3.9 shows a substantial difference in naturalness between 2D (i.e., 0 cm camera base distance) and 3D (i.e., 4 and 8 cm camera base distance) scenes. The naturalness of 3D scenes is rated higher than 2D scenes for all compression levels. Naturalness ratings decrease with increasing JPEG compression for both 2D and 3D. As JPEG compression increases, the difference between 2D and 3D ratings becomes less, indicating that naturalness is mainly determined by the introduced artefacts at higher JPEG compression levels. Nevertheless, the added value of depth is taken into account in the naturalness concept for all compression levels. Apparently, when observers are asked to rate naturalness instead of image quality, they weigh the added value of 3D more and are also more tolerant for some JPEG artefacts in the images. This suggests that people are more focussed on image distortions when asked to rate image quality, whereas they pay less attention to these distortions when asked to rate naturalness. Hence, naturalness seems to be a nice example of a concept weighting more equally the visibility of the distortion as well as the 3D percept.

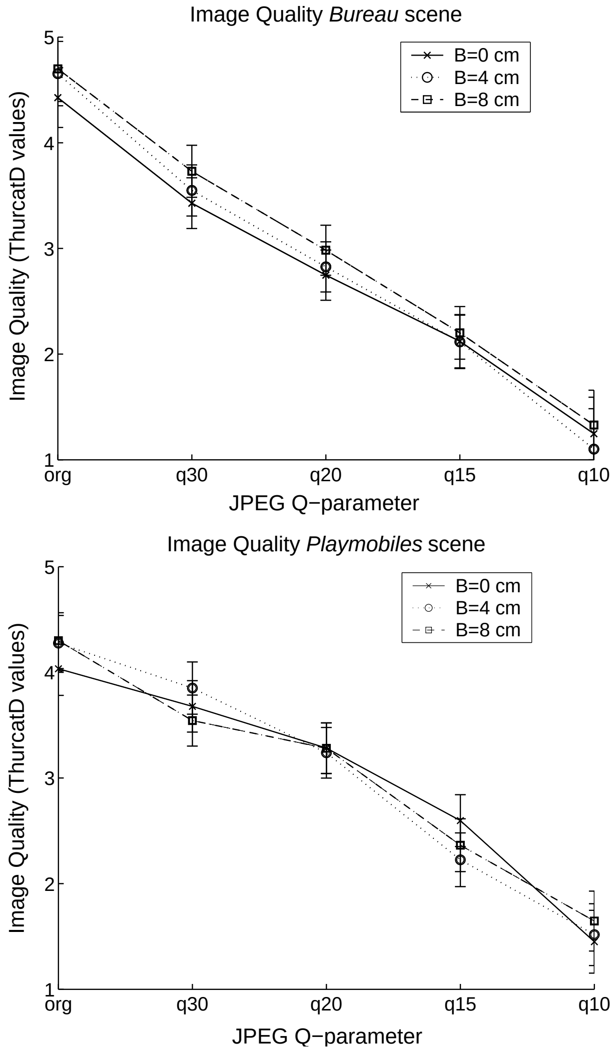


Figure 3.8: Image quality ratings of the *Bureau* and *Playmobiles* scene on a Thurstone true interval scale as a function of the JPEG coding level (lower bit-rate to the right). The three lines represent the different camera base distances. A scale value of 1 corresponds to a bad image quality and a scale value of 5 to an excellent image quality.

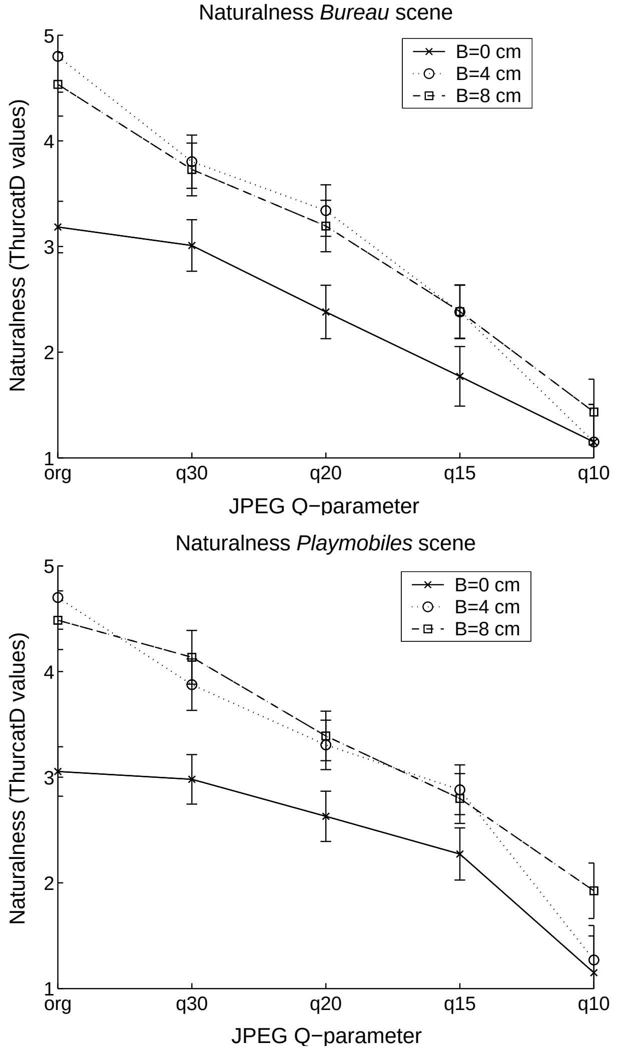


Figure 3.9: The naturalness ratings of the *Bureau* and *Playmobiles* scene expressed in Thurstone scale values as a function of the JPEG coding level (lower bit-rate to the right). The three lines represent the different camera base distances. A scale value of 1 corresponds to a bad naturalness and a scale value of 5 to an excellent naturalness.

3.4 Conclusion

The main conclusion of Experiment 3 is that the evaluation criterion image quality is not sensitive to depth when displaying JPEG impaired image material. In other words, people are more focussed on the JPEG distortions in the 2D and 3D images and do not take into account the added value of depth when judging image quality.

Experiment 2 concludes that naturalness seems to be sensitive to depth. In Experiment 4, the naturalness and image quality concept are directly compared for JPEG coded image material. Again, observers anchored their image quality judgements on the most salient features (e.g., JPEG coding artefacts) while the added value of depth is not recognized, whereas, the added value of depth is clearly taken into account in the naturalness concept for all compression levels. Hence, naturalness seems to be a nice example of a concept taking into account the 2D JPEG distortion as well as the 3D percept, and therefore, seems well suited for the evaluation of 3D content in contrast to image quality.

This chapter focussed on a 2D distortion (JPEG coding) available in both 2D and 3D images. The focus of the next chapter will be on crosstalk, a typical 3D distortion only visible in 3D image material. The focus point of the next chapter will be building up general knowledge about the behavior of crosstalk in 3D images and try to measure the added value of depth with the evaluation criterion naturalness.

Chapter 4

Perceptual attributes of crosstalk in 3D images

Abstract

Nowadays, crosstalk is probably one of the most annoying distortions in 3D displays. So far, display designers still have a relative lack of knowledge about the relevant subjective attributes affected by crosstalk and how they are combined in an overall 3D visual experience model. Perceptual 'benefits' of perceived depth can be nullified by the perceptual 'costs' of crosstalk, as increasing camera base distance is the manipulation that both increases depth and crosstalk. The aim of the first experiment described in this chapter is to investigate three perceptually important attributes influencing the overall visual experience: perceived image distortion, perceived depth, and visual strain. The stimulus material used consisted of two natural scenes varying in depth (0, 4, and 12 cm camera base distance) and crosstalk level (0, 5, 10, and 15%). Observers rated the attributes according to the ITU BT.500-10 in a controlled experiment. Results show that image distortion ratings show a clear increase with increasing crosstalk and increasing camera base distance. Especially higher crosstalk levels are more visible at larger camera base distances. Ratings of visual strain and perceived depth only increase with increasing camera base distance and remain constant with increasing crosstalk (at least until 15% crosstalk). In the second experiment, the naturalness concept is compared with the image quality concept for different crosstalk levels and camera base distances. Results show that naturalness weighs more equally the visibility of the distortion as well as the added value of depth in contrast to image quality.

[0] This chapter is based on Seuntiëns et al. (2005c)

4.1 Introduction

Stereoscopic display techniques are based on the principle of displaying two views, with a slightly different perspective, in such a way that the left eye view is only seen by the left eye and the right eye view only by the right eye. In both stereoscopic and auto-stereoscopic displays, separation of the left and right eye view is one of the major challenges for display designers. Imperfect separation makes a small proportion of one eye's image perceptible to the other eye, a phenomenon known as crosstalk or image ghosting. Crosstalk is generally believed to be undesirable, and display developers are working to minimize crosstalk as much as possible. In view of the fact that a perfect left-right image separation may not always be feasible, in particular in auto-stereoscopic displays, a deeper understanding is needed of the subjective acceptability (rather than mere detectability) of crosstalk, as well as of the relation between the perceptual attributes that contribute to the overall appreciation of an (auto)stereoscopic display. For example, crosstalk becomes generally more noticeable with an increase in left-right image separation. As this is the image manipulation that also introduces stereoscopic depth, an optimal balance between the added value of depth and the negative effect of crosstalk needs to be found.

The research done on crosstalk to date mainly focuses on characterizing the factors contributing to crosstalk in several (auto)stereoscopic display devices. Significant attention has been directed to quantifying crosstalk in time-sequential stereoscopic displays and design of anti-crosstalk models for these displays. The most important factors contributing to crosstalk in Liquid Crystal Shutter (LCS) systems are slow shuttering, shutter leakage, and phosphor afterglow of the (CRT-type) monitor. Woods and Tan (2002) measured and quantified sources of image ghosting such as imperfect shuttering and phosphor afterglow. They found that shutters in the opaque state still had a measurable amount of transmission (shutter leakage) and the decay time of the red phosphor was much longer than the decay time of the blue and green phosphor. There was also a considerable amount of variation in crosstalk for several LCS glasses in combination with different monitors.

Pastoor (1995) performed an experiment investigating visibility thresholds of crosstalk in grey-scale patches. He found that the visibility of crosstalk increases with increasing contrast and increasing binocular parallax of the image. On a high-contrast display (contrast-ratio 100:1) and a reasonable depth range (40 min of arc) crosstalk below 0.3% is invisible in grey-scale patches. Hanazato et al. (1999) measured the visibility thresholds of crosstalk in relation to the amount of binocular parallax and contrast, using geometrical test patterns and found that crosstalk levels below 0.2% were invisible. Importantly, they also found that for natural images crosstalk levels of up to 2% were unnoticed by observers. This is in line with Lipton (1987) who found that natural images may produce ghosting which is less visible than in computer-generated wire-frames, because wire-frame images have hard edges and high contrast. Images with more texture or more detail tend to conceal crosstalk. Kooi and Toet (2004) found that vertical disparity, crosstalk and blur are the most important factors that determine stereoscopic viewing comfort. In summary, crosstalk becomes more visible with increasing parallax, which appears to be less visible or disturbing when images do not contain sharp edges.

Lipscomb and Wooten (1994) designed a crosstalk reduction algorithm. The algorithm brings the background luminance up to 0.3 (0 = black, 1 = white) so that any crosstalk up to a value of 0.3 can be eliminated by darkening the grey background. This anti-crosstalk algorithm is effective but limited to artificial representations, because backgrounds in natural images are not homogeneous or black. The anti-crosstalk model of Konrad et al. (2000) is based on psychovisual experiments quantifying the crosstalk in the system. After quantifying the crosstalk, a look-up table can be created to generate anti-crosstalk. An advantage is that it can be implemented in real time, but a disadvantage is that for every system accurate measurements have to be carried out to quantify the crosstalk.

4.2 Experiment 5

So far, display designers still have a relative lack of knowledge about the relevant subjective attributes affected by crosstalk and how they are combined in an overall 3D visual experience model. Therefore, it is very useful to quantify perceptually relevant attributes such as image distortion, visual eye-strain and perceived depth in relation to crosstalk. This chapter reports on an experiment that was carried out to obtain a better understanding of the perceptual image attributes that can be affected by crosstalk using natural images. It is assumed that observers' preference (visual experience) is a trade-off between perceptual attributes such as image distortions (ghosting, double lines), visual strain (eye-strain) and image enhancements (added value of perceived depth). The aim of the current experiment is to investigate these perceptually relevant attributes when varying the crosstalk level and the amount of depth (image separation). Session 1 focuses on monocular and binocular image distortions and tries to find a combination rule for the left and right eye. Session 2 focuses on visual strain and perceived depth.

4.2.1 Method

Design

The experiment had a mixed design with Image (2 images), Camera base distance (0, 4, and 12 cm) and Crosstalk (4 levels) as within subject factors, and the evaluation concepts tested between subjects.

Observers

Twenty non-expert observers participated in the crosstalk experiment with natural images. Ten observers participated in session 1 and 10 observers in session 2. The observers, mostly students, came from the same age group (20-30 years old). All observers had a visual acuity of >1 (as tested with the Landolt C test), good stereo vision <30 seconds of arc (as tested with the Randot stereo test) and good color vision (as tested with the Ishihara color vision test).

Equipment

For this experiment, a Screenscope$^{\text{TM}}$(mirror stereoscope) was used to direct the left- and right-eye image of a side-by-side displayed stereo pair to the appropriate eye. The main advantage of using this display is that it has perfect image separation, thus allowing for complete experimental control over the percentage of crosstalk. A detailed description of the Screenscope$^{\text{TM}}$can be found in Chapter 1, section 1.3.3.

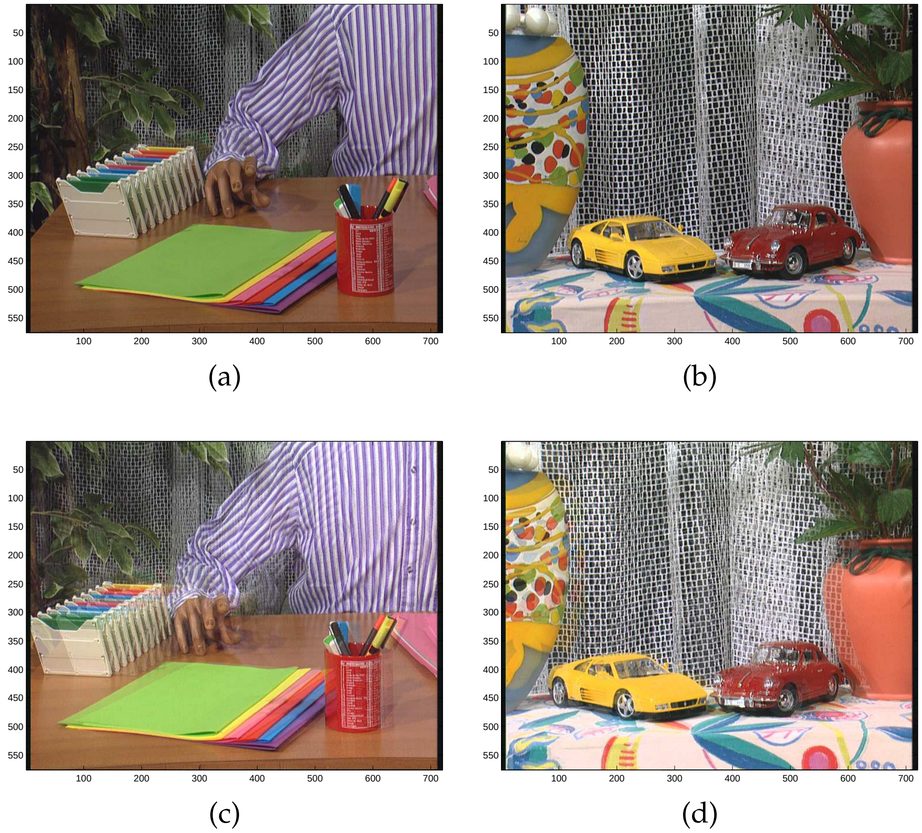
Stimuli

The stimulus material used in this experiment consisted of two natural scenes varying in depth (0, 4, and 12 cm camera base distance) and crosstalk level (0, 5, 10, and 15%). The scenes were recorded in a studio set-up, using a professional stereoscopic studio camera in a toed-in configuration. For each scene, lens focal length and convergence distance of the cameras to the scene were fixed to 20 mm and 1.30 m, respectively. The three recorded camera base distances represented no stereoscopic depth (0 cm), pleasant stereoscopic depth (4 cm) and substantial stereoscopic depth (12 cm). The *Bureau* scene consisted of a tailor's dummy sitting behind a desk with some office tools. The *Voitures* scene showed a table with two miniature cars and two vases located on the desk. The spatial resolution of both scenes was 720x576 pixels per eye. The scenes are shown in Figure 4.1 panel (a) and (b).

In order to simulate crosstalk in the left and right images, a certain percentage of the RGB values of the right image was superimposed in the left image and vica versa. The new images with induced crosstalk were created according to the following equations. In this example, the crosstalk is induced in the left image for all pixels (x,y). The calculations for the right image are done in a similar way.

$$\begin{aligned} R_l^{'}(x,y) &= min(R_l + \frac{R_r \times p}{100}, 255) \\ G_l^{'}(x,y) &= min(G_l + \frac{G_r \times p}{100}, 255) \\ B_l^{'}(x,y) &= min(B_l + \frac{B_r \times p}{100}, 255) \end{aligned}$$

$R_l^{'}$, $G_l^{'}$ and $B_l^{'}$ represent the new values of the left image with induced crosstalk. R_l, G_l and B_l are the original values of the left image. R_r, G_r and B_r represent the original values of the right image. A certain percentage, p, of the right image is added to the left image. In our experiment this was either 0, 5, 10 or 15%. An example of the two scenes with induced crosstalk is shown in Figure 4.1 panel (c) and (d).



(a)

(b)

(c)

(d)

Figure 4.1: Original scenes *Bureau* and *Voitures* are shown in panel (a) and (b). Scenes with induced crosstalk are shown in panel (c) and (d).

Procedure

In the first session, ten observers assessed the degree of perceived image distortion (ghosting, double lines) for static stereoscopic images with induced crosstalk. In order to be able to separate out the effects of (stereoscopic) visual strain from (monoscopic) image distortions (ghosting, double lines), observers were asked to first rate the monocular image distortion for the left eye and right eye separately (one eye covered) and next, the binocular image distortion (both eyes open).

In the second session, another ten observers assessed the degree of visual strain and the degree of perceived depth in two subsessions. The two subsessions were randomized and contained the same stereoscopic static images as in the first session. All images were presented sequentially to the observers and rated according to the single stimulus scaling method of the ITU (2000a) using a 5-point numerical categorical scale. The value 0 on the scale corresponds with the absence of an attribute in the image. Observers were free to decide when to go to the next image by pressing the space bar.

Each session took approximately 30 minutes. Before the experiment started the observers were asked to read the instructions explaining the task and attribute they had to rate. Subsequently, the observers participated in a trial of twelve stimuli to get acquainted with the stimulus set and the range of variations in scene, camera base distance and induced crosstalk.

4.2.2 Results

Monocular versus binocular image distortion

In the first experiment the perceptual combination of the left and right monocular image distortion with different crosstalk levels and camera base distances was investigated. Therefore, the monocular and binocular image distortions were measured separately and an attempted to find a combination rule was made.

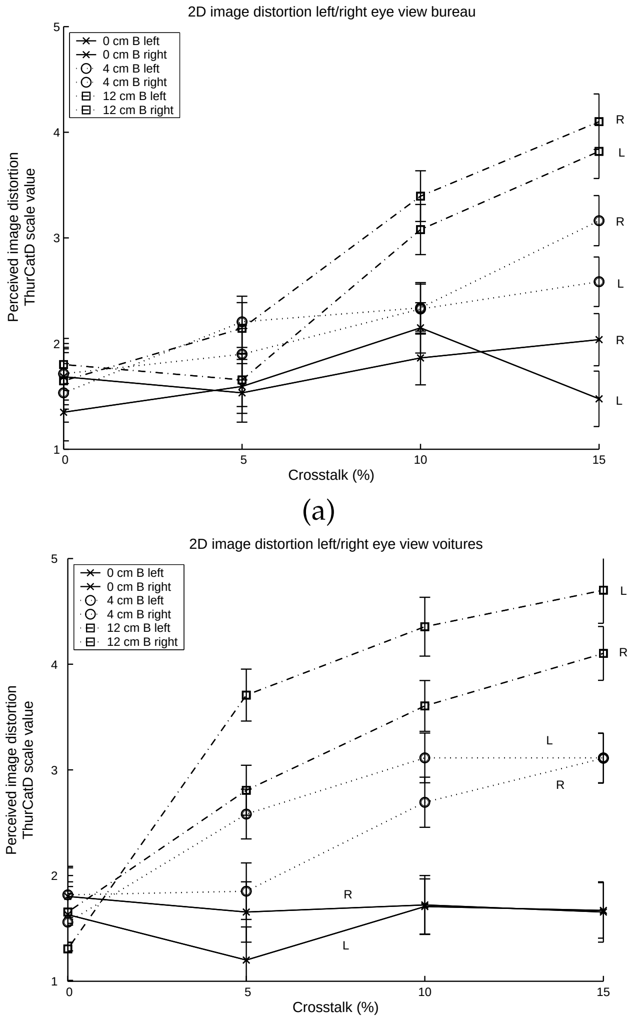
Panel (a) and (b) in Figure 4.3 show the results of the monocular image distortion ratings of the left and right eye view as a function of disparity and crosstalk for both scenes. The amount of implemented crosstalk is shown on the x-axis (0, 5, 10, and 15%). The y-axis represents the estimated Thurstone scale values for each data point scaled back to the original scale. The ThurcatD analysis results show a good model fit (χ^2=22.17, p=0.8048 for *Bureau* and χ^2=25.87, p=0.8312 for *Voitures*). This result means that equal distances on the scale correspond with equal differences in the percept judged. The lines in the graph represent both left and right eye views for three camera base distances (0, 4, and 12 cm).

A repeated measures ANOVA showed no significant main effect of left or right eye view on monocular image distortion for both *Bureau* scene (F(1,9)=18,21, p=0.87) and *Voitures* scene (F(1,9)=16.63, p=0.93). So, the perceived monocular image distortion is nearly the same for the left and right eye in both scenes.

The *Bureau* scene shows a constant and low degree of image distortion for 0, 4, and 12 cm camera base distance when crosstalk increases from 0 to 5%. Crosstalk percentages higher than 5% show an increase in perceived monocular image distortion for a camera base distance of 4 cm and even more for a camera base distance of 12 cm. In contrast to the *Bureau* scene, perceived monocular image distortions already increase for low crosstalk percentages at 4 and 12 cm camera base distance in the *Voitures* scene. Obviously, no crosstalk appears in both scenes for camera base distance of 0 cm (left and right view are identical) and this can be regarded as a control condition. Both figures in panel (a) and (b) show indeed that monocular image distortion is rated at a constant low level for this condition, which bodes well for the reliability of the ratings.

Panel (c) and (d) in Figure 4.3 show the degree of binocular image distortion for the scenes *Bureau* and *Voitures*. The crosstalk percentages, camera base distances and the used scale are the same as in the monocular image distortion figures. The ThurcatD analysis results show a good model fit (χ^2=25.57, p=0.7048 for *Bureau* and χ^2=35.33, p=0.7545 for *Voitures*). This result means that equal distances on the scale correspond with equal differences in the percept judged.

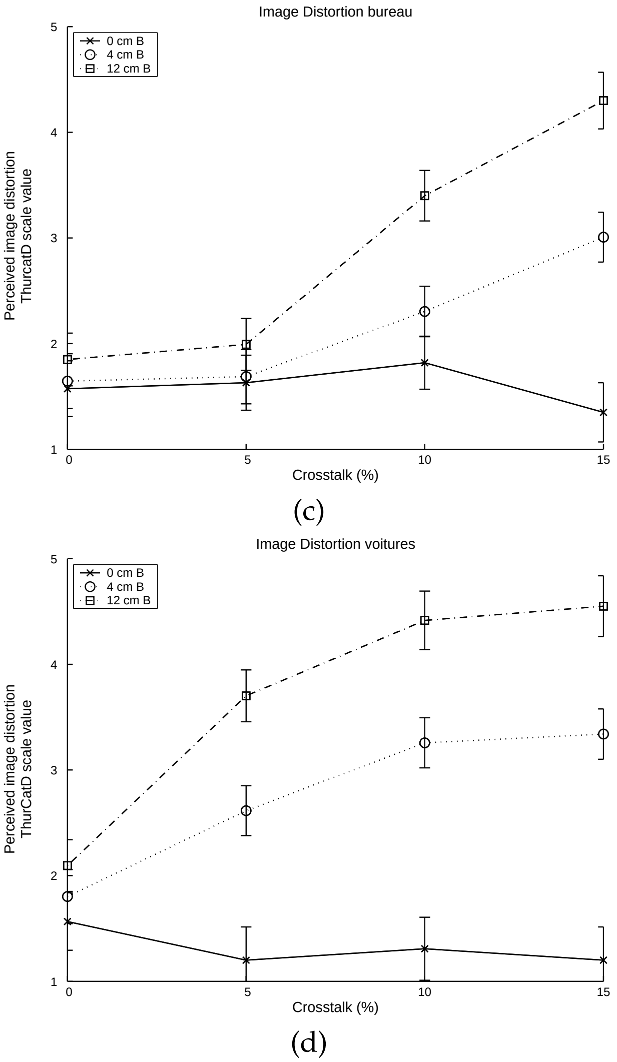
The perceived binocular image distortion in the *Bureau* and *Voitures* scene in Figure 4.3 panel (c) and (d) follows the same trend as the monocular image distortion of the left and right eye view in Figure 4.3 panel (a) and (b).



(a)

(b)

Figure 4.2: Results of the monocular image distortion ratings for the left and right eye view of the scenes *Bureau* and *Voitures* are shown in panel (a) and (b). The x-axis represents the crosstalk percentage and the y-axis represents the Thurstone values for three camera base distances.



(c)

(d)

Figure 4.3: Results of the binocular image distortion ratings are shown in panel (c) and (d). The x-axis represents the crosstalk percentage and the y-axis represents the Thurstone values for three camera base distances.

The binocular image distortion ratings of the *Bureau* and *Voitures* scene in Figure 4.3 panel (c) and (d) show an increasing trend with increasing camera base distance and crosstalk. A repeated measure ANOVA shows a significant main effect of camera base distance on binocular image distortion for *Bureau* (F(2,18)=43.09, p<0.01) and *Voitures* (F(2,18)=209.55, p<0.01) and a main effect of crosstalk for *Bureau* (F(3,27)=46.95, p<0.01) and *Voitures* F(3,27)=99.28, p<0.01). In addition, a significant interaction was found between camera base distance and crosstalk on binocular image distortion for *Bureau* (F(6,54)=16.63, p<0.01) and *Voitures* (F(6,54)=49.04, p<0.01). This interaction has implications for the interpretation of the main effects. The main effect of crosstalk on binocular image distortion is only valid for a camera base distance of 4 and 12 cm, because the ratings on camera base distance 0 cm remain constant with increasing crosstalk.

The data suggests the following model for the binocular image distortion as a function of the monocular image distortions of the left and right eye.

$$D_{3D} = 1/2(D_l + D_r)$$

In this model, D_{3D} represents the calculated binocular image distortion and D_l and D_r represent the left and right eye monocular image distortion respectively. This model seems suited because the correlations between the calculated binocular image distortion and the measured binocular image distortion are very high for both scenes (*Bureau* r = 0.98 and *Voitures* r = 0.97). So, the perceived binocular image distortion can be represented by the average of the perceived monocular image distortions for the left and right eye.

Visual strain

Figure 4.4 shows the degree of visual strain for the scenes *Bureau* and *Voitures*. The amount of implemented crosstalk is shown on the x-axis (0, 5, 10, 15%). The y-axis represents the estimated Thurstone scale values for each data point scaled back to the original scale. The ThurcatD analysis results show a good model fit (χ^2=21.52, p=0.9377 for *Bureau* and χ^2=19.24, p=0.9729 for *Voitures*). This result means that equal distances on the scale correspond with equal differences in the percept judged. The lines in the graph represent the three camera base distances (0, 4 and 12 cm).

In Figure 4.4, visual strain remains constant with increasing crosstalk percentages for both scenes *Bureau* and *Voitures*. Only an increase in camera base distance increases the visual strain. For both the *Bureau* and *Voitures* scene, a repeated measures ANOVA only showed a main effect of camera base distance on visual strain (F(2,18)=35.62, $p<0.01$) and F(2,18)=47.93, $p<0.01$) respectively. No interaction effects of camera base distance and crosstalk were found for visual strain.

Perceived depth

Figure 4.5 shows the perceived depth for the scenes *Bureau* and *Voitures*. The amount of implemented crosstalk is shown on the x-axis (0, 5, 10, and 15%). The y-axis represents the estimated Thurstone scale values for each data point scaled back to the original scale. The ThurcatD analysis results show a good model fit (χ^2=36.70, p=0.3010 for *Bureau* and χ^2=18.35, p=0.9825 for *Voitures*). This result means that equal distances on the scale correspond with equal differences in the percept judged. The lines in the graph represent the three camera base distances (0, 4, and 12 cm).

In Figure 4.5, perceived depth only increases with increasing camera base distance. For increasing crosstalk, the perceived depth remains constant for both scenes. A repeated measures ANOVA only showed a main effect of camera base distance on perceived depth (F(2,18)=175.17, $p<0.01$) for *Bureau* and (F(2,18)=77.95, $p<0.01$) for *Voitures*. No interaction effects of camera base distance and crosstalk were found for perceived depth.

4.2.3 Discussion

The results of the binocular image distortion ratings in Figure 4.3 panel (c) and (d) show a clear increase with increasing crosstalk and increasing camera base distance. Especially higher crosstalk levels are more visible at larger camera base distances. Ratings of visual strain and perceived depth (Figure 4.4 and 4.5) only increase with increasing camera base distance and remain constant with increasing crosstalk (at least until 15% crosstalk), which is noteworthy as crosstalk has, in the past, often been

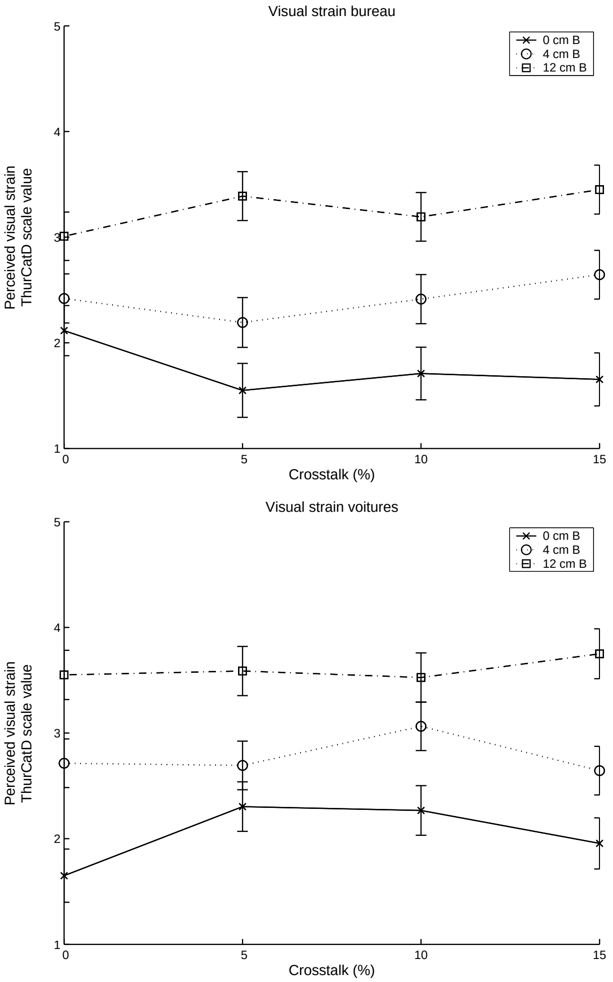


Figure 4.4: Visual strain ratings for the 3D scenes *Bureau* and *Voitures*. The x-axis represents the crosstalk percentage and the y-axis represents the Thurstone values for 3 camera base distances.

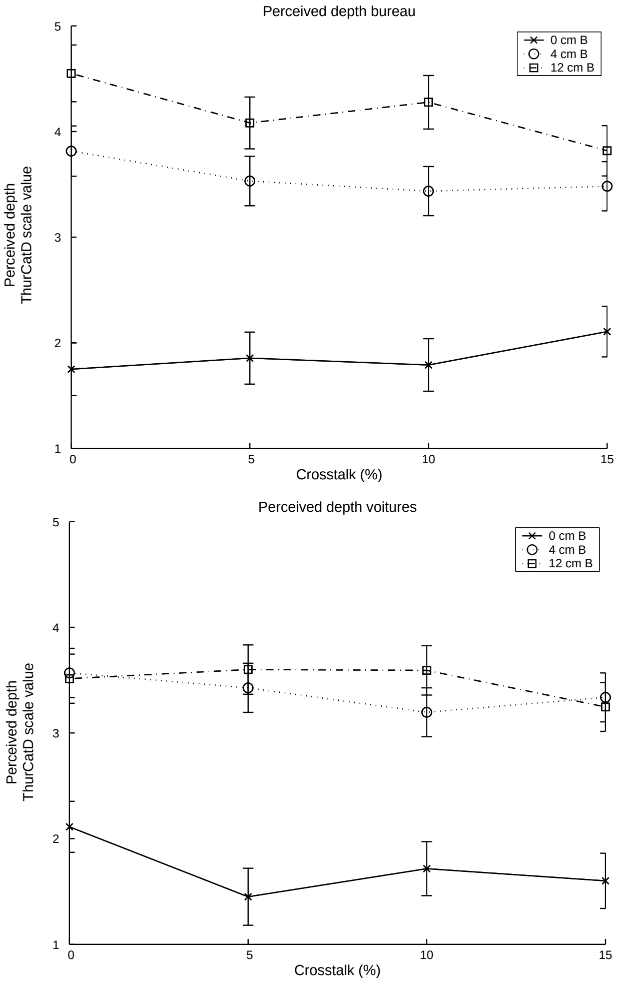


Figure 4.5: Perceived depth ratings for the 3D scenes *Bureau* and *Voitures*. The x-axis represents the crosstalk percentage and the y-axis represents the Thurstone values for 3 camera base distances.

cited as one of the potential causes of eye strain (Pommeray et al., 2003). However, in the current experiment relatively short display times were employed (between 5-10 seconds), thus it is possible that watching stereoscopic content for a longer time will increase visual strain with higher crosstalk levels. In natural image material, some crosstalk is allowed but it depends strongly on the scene content as shown in this experiment. Image distortions caused by crosstalk percentages up to 5% are hardly visible in the *Bureau* scene for all camera base distances, as in contrast to the *Voitures* scene were crosstalk is visible as soon as it is introduced. Difficult to say is where crosstalk becomes visible in the *Voitures* scene because the only measurements were at 0% and 5% crosstalk but our results are in line with Hanazato et al. (1999) who suggested a threshold for natural images around 2%. Based on their information, display designers should keep the system crosstalk below 2% in order to show natural stereoscopic content without any problems. Also, small depth settings reduce the visibility of image distortions due to crosstalk as shown in this experiment. Image distortions at 4 cm camera base distance are less visible than image distortions with the same intensity at 12 cm camera base distance. Perceived depth of both scenes increases significantly as soon as camera base distance is increased from 0 cm to 4 cm. The added value of perceived depth is very small, when increasing the camera base distance further to 12 cm, so, a nice depth percept is already perceived at 4 cm camera base distance with the added advantage that perceived image distortion and eyestrain are significantly lower than at larger camera base distances. This experiment investigated only crosstalk percentages up to 15% crosstalk but it is well possible that visual strain will increase with crosstalk percentages above 15%. Also, perceived depth may be affected negatively for crosstalk percentages higher than 15% because of fusion problems. In the next experiment, image quality and naturalness concepts are directly compared for crosstalk impaired images.

4.3 Experiment 6

This experiment focuses on both aspects depth reproduction and crosstalk. More precisely, both image quality and naturalness are evaluated on crosstalk- impaired images. Previous experiments show that the image quality of 3D images is rated approximately equal to their 2D counterparts when 2D or 3D artefacts are visible. Thus it appears that image quality is mainly determined by the visibility of artefacts in the images and not so much by the added value of depth. It is expected that naturalness will take the added value of 3D into account, even when artefacts are visible in the images.

4.3.1 Method

Design

The experiment had a mixed design with Image (2 images), Camera base distance (0, 4, and 8cm) and Crosstalk (5 levels) as within subject factors, and the two different evaluation concepts (image quality and naturalness) tested between subjects.

Observers

Twenty observers participated in this experiment. Ten of them judged image quality whereas another ten observers judged naturalness of the 3D scenes. The observers, mostly students and employees of a research department, had a visual acuity >1 per eye (as tested with the Landolt C chart), a stereo vision <30 seconds of arc (as tested with the Randot Stereo Test) and no color deficiencies (as tested with the Ishihara test).

Equipment

For running the experiment the PORT (Perceptie Onderzoek Research Tool) system was used, which was developed at Philips to automatically create test scripts and collect data. It consists of 3 hardware components connected via a network:

- Port console: This is an ordinary PC showing the test leader the user-interface and controls of the experiment.
- Port participant interface: This PC (in our case a notebook) displays a user interface giving the participant the opportunity to enter his or her score.

- Video streamer: A high-end PC with dedicated hardware (Radeon 9000 graphics card, hard disk) and inhouse developed software (VideoSim) that allows to show sequences in real-time. The video streamer is connected to a Philips Brilliance 107P2 CRT monitor (1600x1200).

Twenty test scripts were generated using the port system and each test script contained a randomized sequence of the stimulus set. The port console used these scripts as input and controlled the video streamer showing the images on the CRT monitor. Observers assessed the image using the port participant interface and their score was saved in the port console. The final results were saved in a text file containing subject number, stimulus set, and accompanying scores.

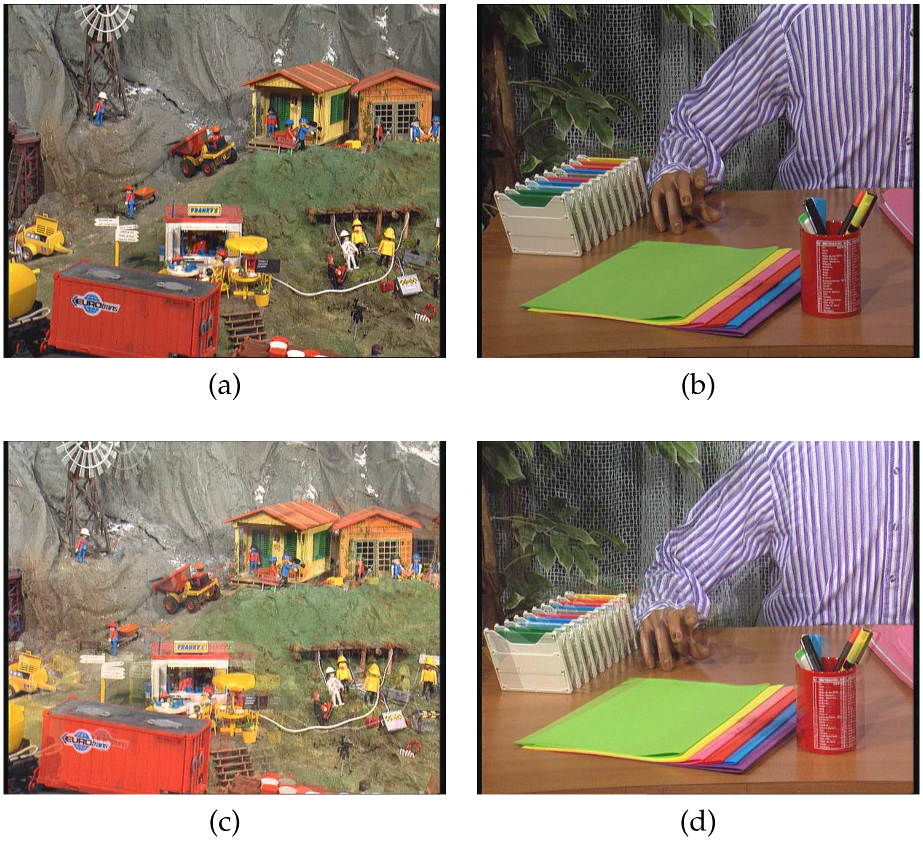
Also for this experiment, a Screenscope^{TM}(mirror stereoscope) was used to direct the left- and right-eye image of a side-by-side displayed stereo pair to the appropriate eye. A detailed description of the Screenscope can be found in Chapter 1, section 1.3.3.

Stimuli

The image material used in this experiment consisted of two still scenes, *Playmobiles* and *Bureau*, recorded with the same professional stereoscopic studio camera as in Experiment 5. The scene *Playmobiles* consists of a colorful toy landscape with mountains in the background and numerous *Playmobiles* in the foreground. The scene *Bureau* consists of a tailor's dummy sitting behind a desk with some office equipment. For these scenes various levels of depth (0, 4, and 8 cm camera base distance) and crosstalk (0, 5, 10, 15, and 20%) were simulated. The spatial resolution of both scenes was 720x576 pixels per eye.

The simulation of crosstalk is described in section 4.2.1. An example of the two original scenes (panel (a) and (b)) and crosstalk induced scenes (panel (c) and (d)) is shown in Figure 4.6

In total 2 (scenes) x 3 (depth levels)x 5 (crosstalk levels) = 30 images were used in the stimulus set. The experiment consisted of first a training session of 6 stereoscopic images and also included one repetition measurement per stimulus, so in total 66 stereoscopic images were shown.



(a)

(b)

(c)

(d)

Figure 4.6: Original scenes *Playmobiles* and *Bureau* are shown in panel (a) and (b). Scenes with induced crosstalk are shown in panel (c) and (d).

Procedure

The set of 66 stereoscopic images was presented sequentially to the observers in a random order. A grey adaptation field was shown between two consecutive stimuli. The images were rated according to the single stimulus scaling method. Each attribute (image quality and naturalness) was rated by 10 observers. A categorical scale ranging from 1 to 5 was used. On this scale 1 corresponds to bad image quality/naturalness and 5 to excellent image quality/naturalness. The scale was labeled with the adjective terms [bad]-[poor]-[fair]-[good]-[excellent] according to the ITU BT500-10 (2000). During the training the observers got acquainted with the stimulus set and its variation in image quality/naturalness.

4.3.2 Results

Image Quality

In Figure 4.7, the image quality ratings are shown for the *Bureau* and *Playmobiles* scenes. On the x-axis the crosstalk levels are indicated. The y-axis represents the estimated Thurstone scale values for each data point scaled back to the original scale. The ThurcatD analysis results show a good model fit (χ^2=25.67, p=0.9778 for *Bureau* and χ^2=37.68, p=0.6609 for *Playmobiles*). This result means that equal distances on the scale correspond with equal differences in the percept judged. The three lines in the figure represent the three camera base distances.

Figure 4.7 shows for both scenes that the 2D image quality (i.e., B = 0 cm) is independent of crosstalk level and is rated as good over the whole range of crosstalk levels. This was expected because crosstalk is a typical 3D distortion linked to the disparity between the left and right eye. Hence, it does not occur in 2D images. 3D image quality only exceeds 2D image quality when the 3D image is undistorted (0% crosstalk). As soon as some crosstalk is introduced in the natural images, image quality is rated the same (for 5% crosstalk) or lower (for higher levels of crosstalk) in 3D images compared to 2D images. A higher disparity (e.g., B = 8 cm) results in a lower image quality rating compared to lower disparity values (e.g., B = 4 cm) when crosstalk is introduced in the images. Since this is not the case at zero crosstalk distortion, there clearly is some interac-

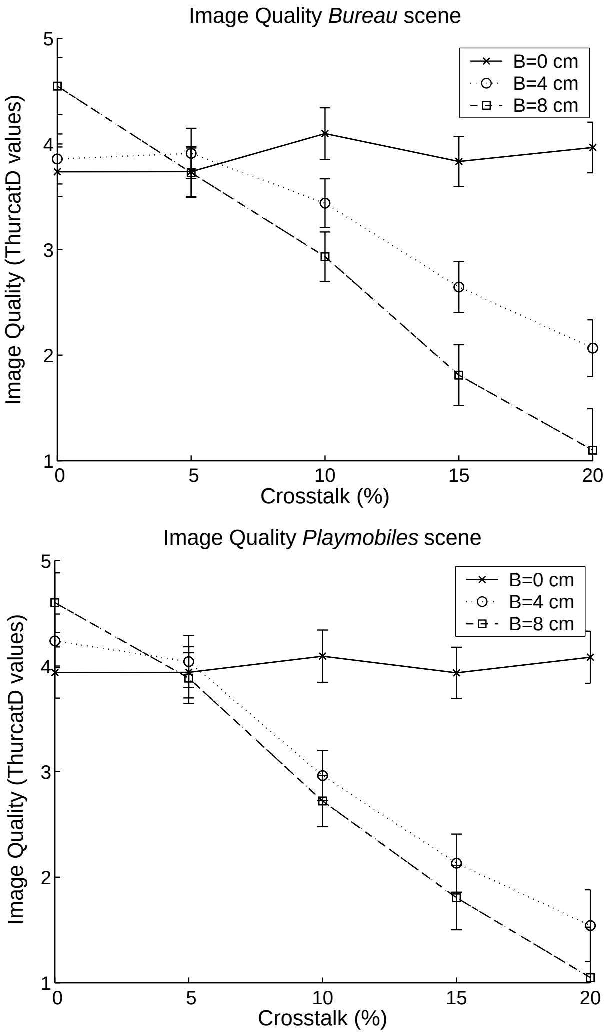


Figure 4.7: Image quality ratings of the *Bureau* and *Playmobiles* scene on a Thurstone true interval scale as a function of the crosstalk level. The three lines represent the different camera base distances. A scale value of 1 corresponds to a bad image quality and a scale value of 5 to an excellent image quality.

tion between camera base distance and crosstalk level affecting the image quality. This is due to the higher visibility of double edges in images with a higher disparity.

Naturalness

Figure 4.8 shows the naturalness ratings for the scenes *Bureau* and *Playmobiles*. The x-axis indicates the crosstalk level. The y-axis represents the estimated Thurstone scale values for each data point scaled back to the original scale. The ThurcatD analysis results show a good model fit (χ^2=40.03, p=0.5579 for *Bureau* and χ^2=45.70, p=0.3210 for *Playmobiles*). This result means that equal distances on the scale correspond with equal differences in the percept judged.The different camera base distances are represented by the three lines in the graph.

Similarly to the image quality scores, also the naturalness scores are independent of crosstalk level when the camera base distance is 0 cm. As mentioned above, this is not surprising, since crosstalk is a 3D distortion not affecting the quality of a 2D image. The image quality of the 2D images is on average considered to be good, whereas their naturalness is judged to be just fair. The naturalness of 3D images is rated higher than that of 2D images up to crosstalk levels of 10%. Only when the crosstalk level exceeds 10%, naturalness is rated lower in 3D images than in 2D images. Observers indicated that at this level the crosstalk distortion gets very annoying and influences the naturalness percept negatively. Apparently, when observers are asked to rate naturalness instead of image quality, they weight the added value of 3D more and are also more tolerant for some crosstalk in the images. This suggests that people are more focussed on image distortions when asked to rate image quality, whereas they pay less attention to these distortions when asked to rate naturalness. Hence, naturalness seems to be a nice example of a concept weighting more equally the visibility of the distortion as well as the 3D percept, as was also shown in previous chapters in relation to other image distortions.

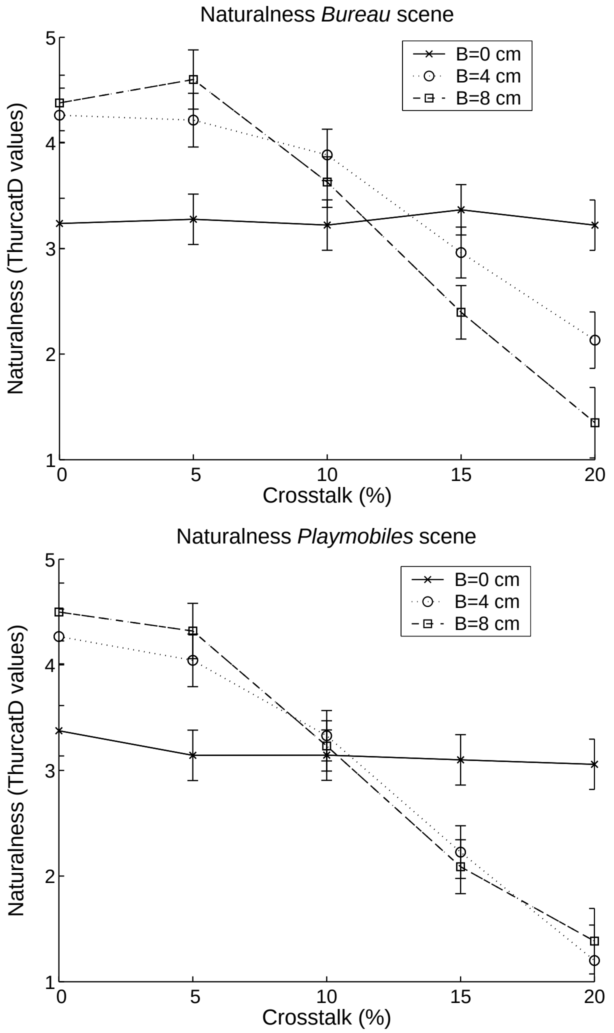


Figure 4.8: The naturalness ratings of the *Bureau* and *Playmobiles* scene expressed in Thurstone scale values as a function of the crosstalk level. The three lines represent the different camera base distances. A scale value of 1 corresponds to a bad naturalness and a scale value of 5 to an excellent naturalness.

4.4 Conclusion

In sum, the experiments in Chapter 4 were performed to gain better insights in the behavior of a typical 3D distortion (crosstalk) in relation to perceptually relevant attributes such as perceived image distortion, visual strain, and perceived depth. The 'benefits' of the added value of depth can be nullified by the 'costs' of crosstalk, as increasing camera base distance is the parameter that both increases crosstalk and depth.

The main conclusion of Experiment 5 is that perceived image distortion increases with increasing crosstalk levels and increasing camera base distance. Visual strain and perceived depth are not affected at all by increasing crosstalk ($< 15\%$). Increasing camera base distance is the only parameter that increases visual strain and perceived depth.

Experiment 6 concludes that naturalness of 3D images is rated higher than the naturalness of 2D images up to crosstalk levels of 10%. Image quality ratings are always higher for the 2D images as soon as crosstalk is visible in the 3D images. This suggests that people judging image quality are more focussed on image distortions, whereas they take into account the added value of depth when judging naturalness.

Now that we have more insights in the behavior of 2D (JPEG) and 3D (crosstalk) image distortions and how to measure the added value of depth, we make a next step and try to model two evaluation concepts in terms of image quality and depth in Chapter 5. An additional experiment is performed in Chapter 5 to have a complete and valid data set in which naturalness, viewing experience, image quality, and depth are evaluated. In the second part of Chapter 5, the modeling part is described.

Chapter 5

Modeling the added value of 3D

Abstract

The goal of this chapter is to model the concepts viewing experience and naturalness in terms of image quality and depth. In order to achieve this, reliable weighs are needed and therefore all data should come from experiments containing the same stimulus set and set-up. An additional experiment is described in this chapter to complete the data set. The main conclusion of this additional experiment is that image quality exhibited a higher score for the '2D' depth level than for higher depth levels, whereas the opposite is true for naturalness. In the case of viewing experience all the depth levels got more or less similar scores. Thus, naturalness takes into account the added value of depth, whereas image quality does not.

The variables in the overall data set used for the actual modeling are the type of artefact (2D or 3D artefacts) and the content generation method (conversion or multi-view recording). Naturalness and viewing experience are modeled in terms of image quality and depth according to a linear combination of image quality and depth. Results show that naturalness incorporates more image quality than viewing experience in those cases where only the perceived image quality is varied, and not the amount of depth. In cases where depth is varied, naturalness takes the added value of depth more into account than viewing experience. Hence, the criterion naturalness seems to balance the perceived image quality and the perceived depth.

[0] This chapter is based on Lambooij et al. (2005)

5.1 Introduction

The main objective of this chapter is to model the concepts viewing experience and naturalness in terms of image quality and depth. A complete data set must be obtained before both concepts can be modeled appropriately. The variables in the final data set are different artefacts (2D and 3D) and different content generation methods (2D-3D conversion or multi-view recording). A comparison is made between 2D-3D conversion using a single image as input and multi-view recording with 9 cameras. The advantage of multi-view recording is that full information from 9 angles (9 images) is available and can be displayed on the multi-view display. The disadvantage of the 2D-3D conversion algorithm is the occlusion problem which gets constantly worse when increasing the virtual camera base distance in the algorithm. The advantage is that the format is flexible and the amount of depth can be adjusted. In Experiment 7, the data set for modeling the concepts viewing experience and naturalness will be completed. Next, the actual modeling will be done.

5.2 Experiment 7

This experiment focuses on image quality, perceived depth, viewing experience and naturalness of 2D and 3D image material. 2D-3D conversion software (introducing depth artefacts) and multi-view camera recordings (accurate depth representation) are directly compared in one experiment to complete the data set for the actual modeling.

5.2.1 Method

Design

The experiment had a within subjects design with Content (four scenes), Depth (four levels: 2D, 2D-3D conversion with two settings of the camera base distance, and multi-view recording), and Noise (three levels) as dependent variables and image quality, depth, viewing experience, and naturalness as independent variables.

Observers

One female and thirteen male naive observers participated in the experiment. Seven observers worked in a research environment, five observers were internal graduation students and two observers were external students all with a technical background. Their ages ranged from 26 to 34. All observers had good stereo vision (<40 seconds of arc as tested with the Randot stereo test).

Equipment

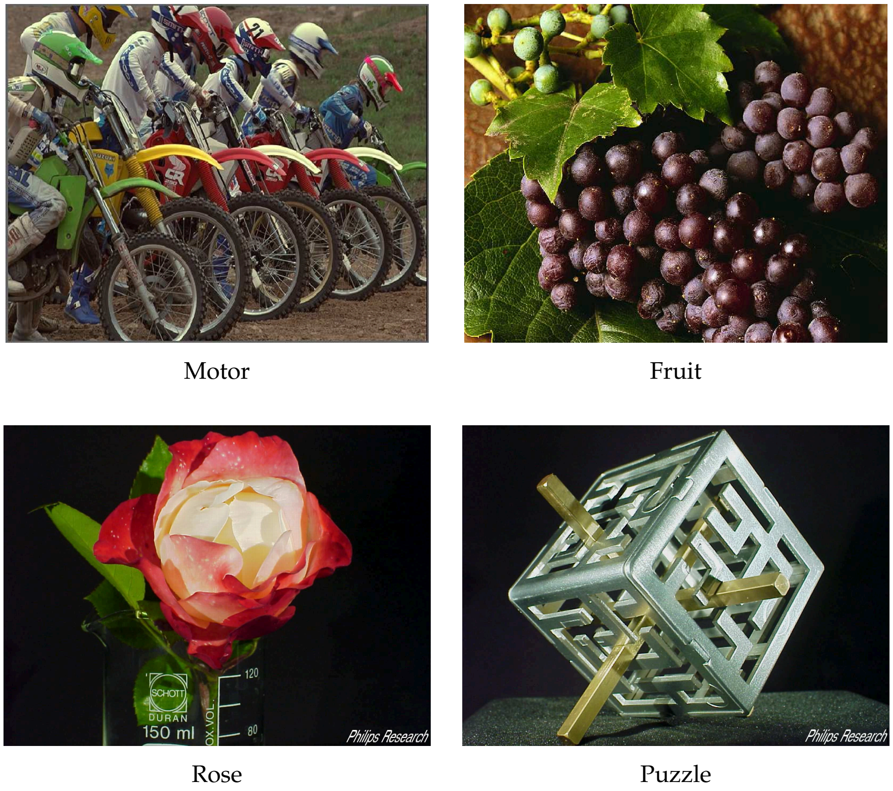
A 20″ Philips multi-view auto-stereoscopic display as described in Chapter 1, section 1.3.3 was used in this experiment . Nine different views were generated either using 2D-3D conversion software or using a nine-camera recording setup. Custom built software (PORT, Perceptie Onderzoek Research Tool) was used to conduct this psychophysical experiment.

Stimuli

The image material used in this experiment consisted of four original static scenes (see Figure 5.1). Two images from Experiment 1 (*Motor* and *Fruit*) and two from Experiment 2 (*Rose* and *Puzzle*). The two originals *Rose* and *Puzzle* were shot using the nine-camera recording setup (accurate depth information). The middle view (camera five) was used as 2D input for the conversion algorithm. The two other originals (*Motor* and *Fruit*) were originally 2D, and hence, for these stimuli nine-view camera recordings were not available.

For all four originals the 2D content was converted into 3D with the 'realtime' algorithm (best performing algorithm) as described in Experiment 1. Different depth levels (2D, 0.01 cbd, 0.02 cbd) were applied by varying the camera base distance parameter in the 'realtime' algorithm. The algorithm generates 3D images containing nine views. A camera base distance of 0.01 in the algorithm corresponds to a distance of 1 cm between each of the nine virtual views. The view-offset, i.e., the part that is displayed in front of the screen was 1/3 (2/3 was displayed behind the screen).

All output images had a resolution of 1600x1200 pixels and were directly



Motor

Fruit

Rose

Puzzle

Figure 5.1: Original scenes *Motor*, *Fruit*, *Rose* and *Puzzle* as used in the experiment.

displayed on the 20″ Philips multi-view auto-stereoscopic monitor. Three levels of white Gaussian noise ($\bar{x} = 0$, $\sigma^2 = 0, 0.005, 0.02$) were added to all output images after 2D-3D conversion. This resulted in stimuli with no distortion (original image) to stimuli with a PSNR (Peak Signal to Noise Ratio) of 19 dB for both 2D and 3D images.

Procedure

The stimuli were evaluated by four different evaluation criteria, namely image quality, depth, naturalness and viewing experience. The rating scale for all evaluation criteria was labeled with the adjective terms [bad]-[poor]-[fair]-[good]-[excellent] according to the ITU (2000a) recommendation regarding subjective quality assessment. In every session observers had to assess the same set of stimuli, but for different evaluation criteria. The evaluation criteria were assigned in random order to observers to cancel out order effects. Two criteria were combined into one session, so each subject had to come back once with at least five days in between two rating sessions. Prior to the experiment, observers were given a brief introduction on paper about the experiment. Any remaining questions were answered and subsequently a short training session was conducted. The training consisted of six stills that were also present in the experiment. The training session was implemented to make the observers familiar with the assessment method and with the extremes with respect to depth and noise levels. The actual experiment consisted of 42 stimuli. Four images, three depth levels (conversion), and three noise levels (4x3x3) plus a nine-view depth level for two images with three noise levels each (1x2x3). The lighting conditions of the room were constant for all observers and the level of light in the room was 3 lux, measured perpendicular to the display in the direction of the viewer.

5.2.2 Results

The statistical analyses applied on the subjective data were Thurstone scaling and MANOVA. The MANOVA's independent variables were Depth, Content, and Noise and the dependent variables were the evaluation criteria image quality, depth, viewing experience, and naturalness.

A Thurstone analysis (as described in Chapter 3, paragraph 3.2.2) on the subjective data revealed a good model fit which implies that the distance between scale values was perceived as equal. Further analysis of this experiment was divided into two separate parts, namely with and without the nine-view depth level. These sets must be analyzed separately, because the nine-view depth level was an additional depth level that was only available for two of the four original images. So, when the nine-view depth level was incorporated, the analysis was performed on two original images and when the nine-view depth level was not incorporated, the analysis was performed on all four original images.

Effects of content, depth and noise without the nine-view depth level

The averaged assessment scores as a function of noise level and depth level with their error bars are plotted in Figure 5.2 for each of the assessment criteria.

From Figure 5.2, it is clear that naturalness, viewing experience and image quality all seem similarly influenced by the introduced noise, that is, the assessment criteria revealed similar slopes as a function of the noise level. Perceived depth however, seems less influenced by noise. It is also clear that the ′0.02′ depth level received the lowest scores for naturalness, viewing experience and image quality. This low score is mainly due to annoying depth artefacts and some color artefacts caused by the ′realtime′ algorithm at higher depth levels. The perceived depth for the ′0.02′ level was not scored higher than for the ′0.01′ level, although the introduced depth was higher for the ′0.02′ depth level. As expected, depth shows the lowest score for the ′2D′ depth level. Finally, it should be noted, that image quality exhibited a higher score for the ′2D′ depth level than for the ′0.01′ depth level, whereas the opposite was true for naturalness and viewing experience. Thus, naturalness and viewing experience take into account the added value of depth whereas image quality does not.

The main effect of Depth level was significant for all evaluation criteria. This result is clearly visible in Figure 5.2 (in all panels), where the depth lines are significantly different from each other. The ′0.02′ depth level showed low assessments for naturalness, viewing experience and image quality. The main effect of Noise was also significant for all evaluation criteria, although the perceived depth seems less influenced by noise. Three out of four evaluation criteria take into account the impairment

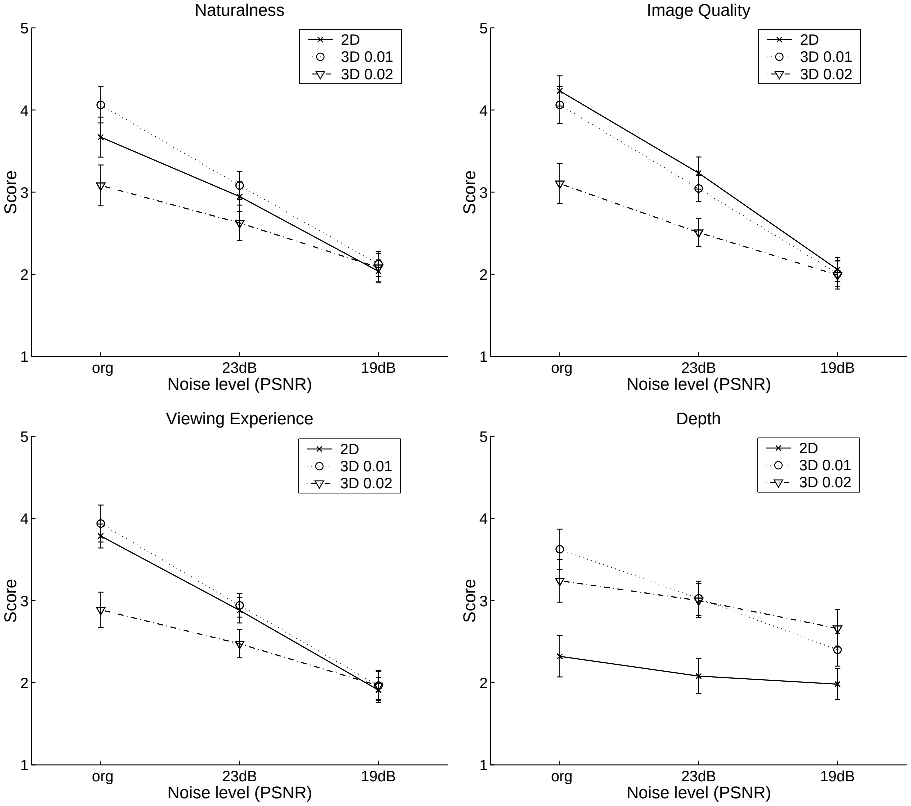


Figure 5.2: Ratings for Naturalness, Image quality, Viewing experience and Depth of the *Motor*, *Fruit*, *Rose* and *Puzzle* scenes as a function of noise level. The three lines represent the different camera base distances (2D, 0.01 and 0.02 cbd). A scale value of 1 corresponds to bad and a scale value of 5 to excellent.

introduced by noise. The last main effect, namely Content, was also significant for all evaluation criteria and this was mainly due to some clear conversion artefacts in the images *Fruit* and *Puzzle*. Although there was a significant effect of Content, the scores of the four scenes were averaged for each assessment criterion because similar main effects and trends were visible and in the same direction in the data of all images.

The three possible two-way interactions between the three fixed factors differed in significance. The interaction Content x Noise was not significant for all evaluation criteria. The Content x Depth interaction was significant for viewing experience (F(6,479)=2.278, p=0.035) and image quality (F(6,479)=3.715, p=0.001) and was not significant for naturalness (F(6,479)=1.905, p=0.078) and depth (F(6,479)=1.002, p=0.423). The behaviour of the content was fairly similar for the ′0.01′ depth level and for the 2D depth level, but considerably different for the ′0.02′ depth level. The depth artefacts introduced at the ′0.01′ depth were subtle and difficult to see. However, the ′0.02′ depth level exaggerated these artefacts. A stimulus containing many depth artefacts (objects in wrong depth planes) received a lower score than a stimulus with fewer depth artefacts. The interaction Depth x Noise was also significant for all evaluation criteria. Further analysis revealed that increasing noise resulted in smaller differences between the different depth levels. This seems plausible, since when there is too much noise, the differences between the depth levels become unclear and overwhelmed by the noise. The 2D stimuli were scored higher than the 3D stimuli in terms of image quality ($p < .05$), but lower in terms of naturalness. The assessment in terms of viewing experience did not differ for the 2D and ′0.01′ depth levels. As expected, the 2D depth level gave the lowest score for the evaluation criterion depth. For image quality, the 2D depth level gave the best results, whereas the ′0.01′ level resulted in the highest scores for naturalness and viewing experience.

Effects of content, depth and noise including the nine-view depth level

The averaged (over all observers) assessment scores as a function of noise and depth level and their error bars are plotted in Figure 5.3 for the different assessment criteria to get a better understanding of the results. These results are based on the images *Rose* and *Puzzle*.

Similar trends can be observed as in Figure 5.2 (i.e., without the nine-

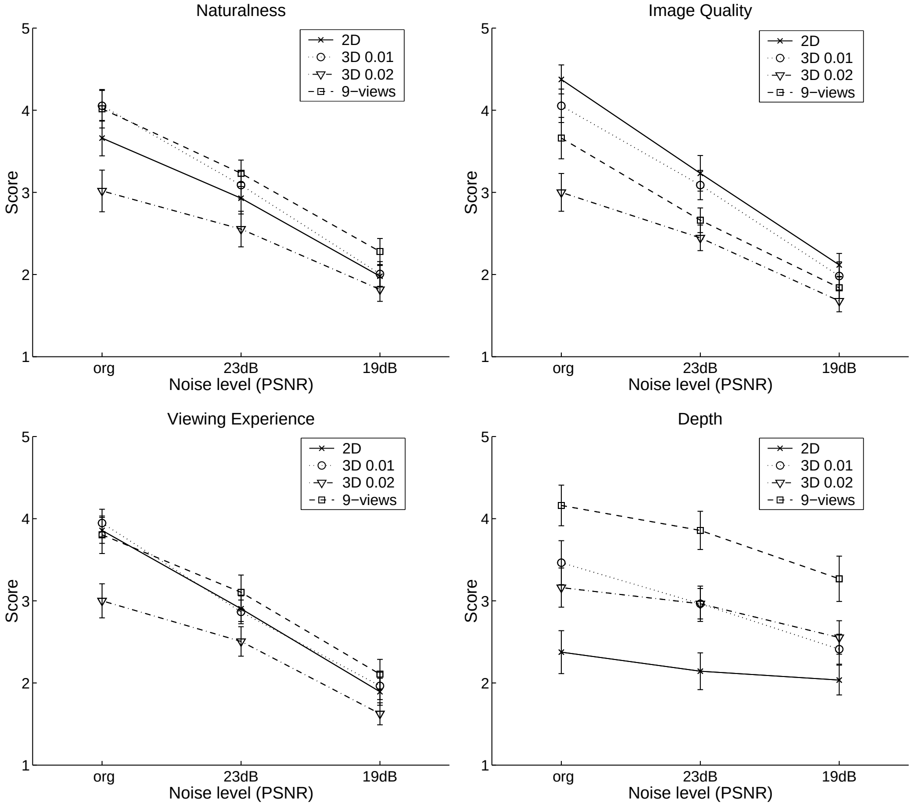


Figure 5.3: Ratings for Naturalness, Image quality, Viewing experience and Depth of the *Rose* and *Puzzle* scenes as a function of noise level. The four lines represent the different camera base distances (2D, 0.01 cbd, 0.02 cbd and 9-view). A scale value of 1 corresponds to bad and a scale value of 5 to excellent.

view depth level). Increasing noise levels seem to have a similar effect on viewing experience, naturalness and image quality, but depth ratings seem to be less affected by increasing noise levels. The ′0.02′ depth level received the lowest scores for naturalness, viewing experience and image quality and is valued almost the same as the ′0.01′ depth level in terms of depth although it introduced more depth. Depth ratings for the nine-view recordings received the highest scores. And finally, image quality was rated higher for the ′2D′ depth level than for the nine-view and the ′0.01′ depth level, whereas the opposite is true for naturalness. In the case of viewing experience all the depth levels, except for the ′0.02′ level, got similar scores.

The main effect of Depth was significant for all evaluation criteria. As expected, also Noise had a strong influence on the assessment of the stimuli and was also significant for all criteria, but to a lesser degree for the depth criterion. Finally, also a significant main effect of Content was found for all criteria. The scene *Puzzle* was scored considerably higher than the scene *Rose*. Nevertheless, they both exhibited the same trends regarding Noise and Depth.

The three possible two-way interactions behave very similar to the interactions found without the nine-view depth level. The interaction Content x Noise was not significant for all evaluation criteria. The Content x Depth interaction was significant for all criteria. The image *Puzzle* scored higher except in the ′nine-view′ case. This could be the result of an artefact in the nine-view depth level of the image *Puzzle*. The corner of the cube that points out of the screen towards the viewer was a little blurred in the nine-view depth level. The interaction Depth x Noise was also significant for all evaluation criteria. Again, increasing noise resulted in smaller differences between the different depth levels. The assessment in terms of image quality was lower for the nine-view depth level than for the 2D depth level, but higher in terms of naturalness. No differences were found in terms of viewing experience.

5.3 Towards a 3D Visual Experience model

As mentioned before, one of the goals of this thesis is to determine which evaluation criterion takes into account the added value of depth. The term Visual Experience is used to denote an overarching category characterizing the final outcome of an assessment proces where perceptual costs and benefits are weighted against each other in relation to relevant characteristics of visual display systems. In the case of a stereoscopic 3D system, Visual Experience is a mix of perceived image quality, perceived depth and visual comfort as stated in Chapter 1. The contribution of visual comfort to the 3D Visual experience is not incorporated in the model described below. Future research is needed to give more insights into visual comfort related to 3D Visual experience. Thus, the model is limited to the contributions of image quality and perceived depth.

In this chapter different evaluation concepts that incorporate to some degree the added value of depth, more particularly naturalness and viewing experience, were applied. Since also image quality and perceived depth were assessed, it is possible to find a relation for both evaluation criteria (naturalness and viewing experience) in terms of image quality and perceived depth. In general terms, this relation is given in Equation 5.1 with EC the evaluation criterion.

$$EC = \alpha \cdot IQ + \beta \cdot D + \gamma \tag{5.1}$$

Both evaluation criteria naturalness and viewing experience (EC) take into account image quality (α · IQ) as well as perceived depth (β · D) and a residual term (γ). Given that observers scored all evaluation criteria in different sessions and used the same scale and range, it might be assumed that for most observers the given scores are not absolute scores, but rather relative scores. As a consequence, the coefficients α and β are only relative contributions and no absolute coefficients.

Table 5.1 describes five stimulus sets and their variations in 2D artefacts (noise), 3D artefacts (depth artefacts due to conversion) and depth variation (2D, 0.01 cbd, 0.02 cbd, 9 views). These five sets are based on the stimuli from Experiment 1 (only the stills) and Experiment 7. Stimulus set I and II result from Experiment 1 and stimulus set III, IV, and V result from Experiment 7.

Table 5.2 depicts the results of a regression analysis on Equation 5.1 based on the data from Experiment 1 (only the still images) and Experiment 7. In order to get a better understanding of the impact of 2D artefacts, 3D

Table 5.1: Description of the different stimulus variations including 2D artefacts, 3D artefacts and depth variation ($\surd$ = present).

Stimulus set	2D artefact noise	3D artefact depth	Depth variation			
			2D	0.01	0.02	9v
I	-	$\surd$	-	$\surd$	-	-
II	-	$\surd$	-	$\surd$	$\surd$	-
III	$\surd$	-	$\surd$	-	-	$\surd$
IV	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	-
V	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$

artefacts, and depth variation on the model, a distinction is made between five stimulus sets.

The results for stimulus set I in Table 5.2 show that for naturalness the coefficient for image quality (α) is relatively high and for depth (β) is very low. Experiment 1 confirms this, i.e., the behavior of naturalness with respect to the different algorithms is very similar to the behavior of image quality and is completely different from the behavior of depth. Comparable results hold for the coefficients of viewing experience, although for image quality the coefficient is somewhat lower, and for depth somewhat higher compared to naturalness.

Results for stimulus set II reveal that both naturalness and viewing experience mostly incorporate image quality, but they also incorporate depth. Naturalness incorporates more depth than viewing experience as shown by the depth coefficient (β) and viewing experience incorporates more image quality as shown by the image quality coefficient (α). These results are not in line with the regression results of stimulus set I, where naturalness incorporated more image quality than viewing experience and the influence of the perceived depth on both naturalness and viewing experience was almost negligible.

Regression results for stimulus set III reveal that the coefficient of depth (β) is higher for naturalness than for viewing experience. This is in line with results found in Experiment 2 where the difference between 2D and

Table 5.2: R^2 and regression coefficients of the linear regression model based on the results of Experiment 1 and Experiment 7.

Stimulus	Naturalness			Viewing Experience		
set	R^2	(α)	(β)	R^2	(α)	(β)
I	0.951	0.951	0.063	0.919	0.874	0.191
II	0.942	0.741	0.325	0.857	0.854	0.113
III	0.969	0.798	0.395	0.975	0.878	0.269
IV	0.980	0.884	0.210	0.981	0.923	0.121
V	0.980	0.864	0.246	0.981	0.921	0.144

3D is larger for naturalness than for viewing experience. So, naturalness takes better account of depth than viewing experience.

The results of stimulus set IV show that both naturalness and viewing experience have high coefficients for image quality (α) and low coefficients for depth (β). So, both naturalness and viewing experience are mainly determined by image quality, but they both do incorporate depth and for this experiment naturalness to a higher degree than viewing experience. It is also clear that viewing experience incorporates image quality more than naturalness.

Stimulus set V shows similar results. Naturalness takes into account depth more than viewing experience, and viewing experience incorporates image quality more than naturalness.

5.4 Conclusion

One of the main conclusions of this chapter is that it makes a clear difference which evaluation criterion is applied to measure the added value of 3D stimuli compared to 2D stimuli. The evaluation criterion naturalness seems to have the highest discriminating power between the different depth levels. The results show that the data sets with and without nine-view depth level are assessed relatively similarly. The fact that only the perceived depth is strongly affected by the nine-view depth level seems plausible. The nine-view depth level contains the largest and most accurate depth information (hardly any depth artefacts are present).

Earlier chapters revealed that naturalness in contrast to image quality does incorporate depth. In this chapter (comparing 2D, 3D converted and 3D recorded images in one experiment), 2D images receive also higher scores for image quality than the 3D stimuli. For viewing experience, there are only small, but no significant, differences between 2D and 3D, with a small preference for 3D. Naturalness, on the other hand, is scored higher for 3D. This is also in line with Chapter 2, concluding that the shift between 2D and 3D is larger for naturalness than for viewing experience.

The relationship modeled in Equation 5.1 is applied to five different data sets. In four cases naturalness includes depth to a greater extent than viewing experience. This is true when 2D artefacts (noise) and 3D artefacts (depth artefacts due to 2D-3D conversion) in combination with depth variation are introduced. Only when no depth variation is introduced (stimulus set I), both naturalness and viewing experience do not take into account the added value of depth, and are mainly related to image quality. Apparently, naturalness incorporates image quality more than viewing experience in those cases where only perceived image quality is varied, and not the amount of depth. In cases where also depth is varied, naturalness takes the added value of depth more into account than viewing experience. Hence, the criterion naturalness seems to balance the perceived image quality and the perceived depth more than the criterion viewing experience. Therefore, it is concluded that naturalness is the best concept to evaluate the quality of 3D image content.

The last chapter looks back on the insights gained from the work on 3D quality. A redefined 3D Visual Experience model is presented and the applicability of the model is discussed.

Chapter 6

General Discussion

6.1 Main Conclusions

Stereoscopic television has a long history and the last few years a consensus has been reached that a successful introduction of 3D television broadcast services can only be a lasting success if the perceived image quality and the viewing comfort are at least comparable to conventional 2D television. In addition, 3D television technology should be compatible with conventional 2D television to ensure a gradual transition from one system to the other. This is becoming increasingly feasible because of recent progress in capturing, coding, and display technology. Central to these developments is the viewer's experience which will signify success or failure of the proposed technological innovations. Potential positive effects of introducing the depth dimension in television have often been assumed, but paradoxically, the majority of image quality studies of stereoscopic displays failed to show the added value of 3D. Thus, one central goal of this thesis was to identify and test subjective evaluation paradigms that would be able to characterize this experience. A second goal of this thesis was to describe a 3D visual experience model incorporating image quality, depth and visual comfort. In this way we will be able to systematically analyse perceptual issues regarding 3D TV and produce requirements for design improvements that will eventually lead to a 3D TV system that will be accepted by consumers in their homes.

Chapter 1 provided some basic principles of the human visual system and discussed how monocular and binocular cues are used to construct a 3D representation of the world surrounding us. Furthermore, we described an end-to-end 3D broadcast chain that includes content generation, coding, transmission, and display. The acceptance and commercial succes of such a system aimed at the consumer market depend to a large extent on the users' experience with and responses towards the system. Therefore, it is important to have a clear understanding of the visual experience of 3D TV, both looking at the potential added value of 3D, as well as the potential drawbacks for users. In Chapter 1, a number of existing subjective assessment methods for the evaluation of 2D and 3D imaging systems are reviewed. Engeldrum (2004) describes a model of subjective image quality for display systems, which helps manufacturers to implement and integrate image quality into their products. This model was developed for 2D image quality and was taken as a point of departure for the current 3D work. We believe that a 3D visual experience model should incorporate image quality, depth and visual comfort. A first concept is described in Chapter 1 along with a description of possible new evaluation criteria for 3D TV like presence, naturalness and viewing experience. It was hypothesized that these concepts would take into account the added value of depth and this was investigated in the following chapters.

The main objective of **Chapter 2** was to explore which evaluation criterion is most appropriate to assess 3D quality. Therefore, an explorative experiment was performed using different (not yet optimal) 2D-3D conversion algorithms to generate the 3D image material and measured assessment of image quality, depth, viewing experience, naturalness, and presence. An important caveat for the results of Experiment 1 is that the results are based on image material containing 3D depth artefacts due to 2D-3D conversion algorithms, only 3D images (no 2D reference), and no 2D image artefacts (for instance noise). Results of Experiment 1 indicate that viewing experience and naturalness have the highest discriminating power between the different 2D-3D conversion algorithms. Experiment 2 investigated the concepts viewing experience and naturalness in more detail having complete experimental control over the stimulus set. Perfect 2D and 3D image material was used and several controlled noise degradations were added manually to the images making the added value of depth quantifiable in terms of units of noise impairments. Robust results show that naturalness and viewing experience both take into account the added value of depth as well as the introduced noise distortion. Naturalness seems to have the highest discriminating power between 2D and 3D images.

The focus of **Chapter 3** is the direct comparison between the traditional image quality concept and the naturalness concept for 2D JPEG artefacts. Results from Experiment 3 show that perceived depth is not affected by JPEG coding and perceived image quality is equal for 2D and 3D images. This means that image quality is sensitive for coding distortions (as expected) but does not account for the depth dimension. Experiment 4 shows that the naturalness concept clearly weighs the image distortion and the added value of depth in contrast to image quality and therefore is well suited for the evaluation of 3D content containing 2D artefacts such as JPEG coding.

Chapter 4 focuses on the direct comparison between the traditional image quality concept and the naturalness concept for 3D crosstalk artefacts. Experiment 5 shows that perceived depth is not affected at all by increasing crosstalk. An important observation is that perceived image distortion scales with increasing crosstalk and increasing depth, which is not the case with 2D distortions (such as JPEG) which are independent of the amount of depth (see chapter 3). The results from Experiment 6 show that as soon as crosstalk is visible in the 3D images the image quality ratings for 3D are lower than the image quality ratings for 2D. Crosstalk levels up to 10% exhibit a higher naturalness score for 3D than for 2D images. So, naturalness also seems an appropriate concept measuring the added value of 3D for image material containing 3D artefacts like crosstalk.

Modeling the added value of 3D is the main topic in **Chapter 5**. From image quality and depth measurements the viewing experience and naturalness were predicted with a linear regression analysis and convincing fits were found for the evaluation criteria viewing experience and naturalness. Our results show that naturalness is the best concept weighting both image quality and depth.

6.2 Adapted 3D Visual Experience model

The traditional Image Quality Circle (as proposed by Engeldrum (2004), described in Chapter 1) seems to be a useful framework for the optimization of 3D display systems. However, as the experiments in this thesis show, the image quality concept alone is not enough to measure the vi-

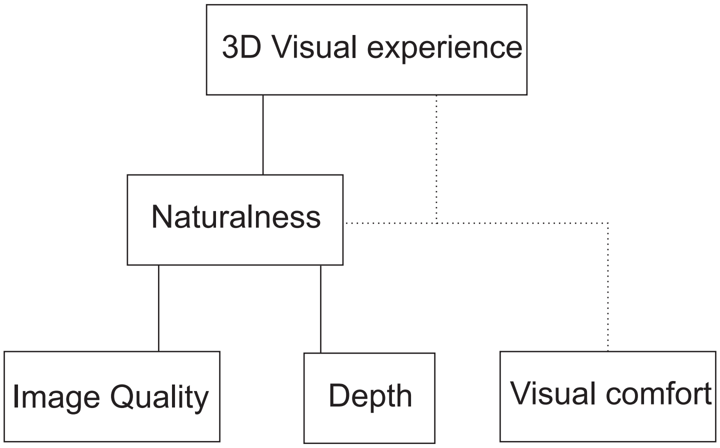


Figure 6.1: 3D Visual experience model

sual experience of a 3D display, since the depth dimension is not incorporated. Image quality mainly appears to be sensitive to 2D image distortions. Naturalness, on the other hand, appears to be more useful in characterizing the experience related to a 3D display. In Figure 6.1, an expansion of the Image Quality Circle for the optimization of 3D display systems is proposed.

The image quality box in this model is the same as the upper box in the model of Engeldrum (2004). The image quality ratings of a 3D TV system can be predicted from the technological variables by using the traditional Image Quality model. The concept naturalness is a higher order concept in the model, weighting both image quality and depth. Visual comfort (dashed line in the model) was not central to this thesis, but will be of great importance for the success of 3D TV. To date, no comprehensive 3D visual experience model is available combining image quality, depth, and visual comfort. This model is a first step forward in measuring overall 3D visual experience. Future research will investigate the role of visual comfort in our proposed model. In the future, this model will be extended with other applications providing us with an enhanced visual experience, e.g., AmbiLight TV (Philips, 2004). This feature introduced by Philips provides ambient light on the walls behind the TV to improve the visual experience. So, not the depth dimension but the light dimension is supposed to enhance the visual experience.

6.3 Applicability of the 3D Visual Experience model

The applicability of the 3D visual experience model depends on the type of application. 3D TV is a typical appreciation-oriented application and the goal is to enhance the 3D visual experience of observers. So, the emphasis of broadcasting and entertainment applications is to provide a whole new visual experience. Our model is well suited for appreciation-oriented applications and takes into account typical appreciation-oriented attributes such as the added value of depth and also, potentially, factors of visual comfort. The goal is to present the images as ′pleasing′ as possible.

On the other hand, the model may be less suited for performance-oriented applications such as medical imaging. The intention of images for medical imaging is not to provide the viewer with a whole new visual experience, but to enable the doctor to make the right diagnosis. 2D image quality for broadcast applications depends on important attributes such as brightness, color rendering, contrast, and sharpness. Some of these 2D image quality attributes are also important for medical imaging, but normally medical image quality is measured in terms of sensitivity, diagnostic accuracy, and specifity and not in terms of image quality. The diagnostic results also strongly depend on the observer′s experience in a particular medical area. Another example of a performance-oriented application is a surveillance camera system. Although some image quality aspects are important, the main goal of such a system is to provide image material which is optimized for face recognition.

6.4 Future Research

Future research will focus on visual comfort of 3D display systems which still is an important issue in current displays. The difficulty of this problem is that increasing the depth level decreases the visual comfort. So, a trade-off is necessary between the depth range of a 3D display and the visual comfort. It is important to have a complete understanding of the advantages and drawbacks of a 3D TV system. Therefore, our 3D visual experience model could contribute to a lower cost design cycle for 3D TV taking into account the perceptual costs and benefits of the complete system so that the technological parameters can be optimized to an optimal visual experience. Although 3D technology has made significant advances over the last years, still some progress must be made before 3D TV can be introduced to the consumer's market. In this respect, human factors research plays an essential role addressing perceptual and usability issues, and optimizing the design of 3D technology from a user-centered perspective.

So will 3D TV replace conventional television anytime soon? Probably not in the short term, because the TV market has recently been confronted with some technology pushes, e.g. HDTV and digital broadcasting. However, the new visual experience that 3D TV offers to the user is too attractive to be ignored. Within two years, 3D TV will start as a new application for conventional PCs, probably with later adaptation to the gaming industry, which are then already hooked up to the TV set situated in the living room. The progress speed will depend on the effort required to create attractive content. Nevertheless, from today's developments one can be optimistic that the scientific and technological challenges of 3D TV will be surmountable: A new visual experience lies ahead of us.

Bibliography

Ahumada, A. J. (1993). Computational image quality metrics: a review. *SID Digest*, 24:305–308.

Berthold, A. (1997). *The influence of blur on the perceived quality and sensation of depth of 2D and stereo images*. Technical report, ATR Human Information Processing Research Laboratories.

Boschman, M. (2000). ThurcatD: A tool for analyzing ratings on an ordinal category scale. *Behavior Research Methods, Instruments, and Computers*, 32:379–388.

Daly, S. (1993). The visible differences predictor: An algorithm for the assessment of image fidelity. In Watson, A. B., editor, *Digital Images and Human Vision*, pages 179–206. New York: MIT Press.

de Ridder, H. (1992). Minkowski-metrics as a combination rule for digital-image-coding impairments. *Proceedings of the SPIE*, 1666:16–26.

de Ridder, H. (1996). Naturalness and image quality: Saturation and lightness variation in color images of natural scenes. *Journal of Imaging Science and Technology*, 40:487–498.

de Ridder, H., Blommaert, F., and Fedorovskaya, E. (1995). Naturalness and image quality: Chroma and hue variations in color images of natural scenes. *Proceedings of the SPIE*, 2411:51–61.

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

Engeldrum, P. (2000). *Psychometric Scaling*. Imcotek Press, Winchester, Massachusetts, USA.

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

Eskicioglu, A. A. and Fisher, P. S. (1995). Image quality measures and their performance. *IEEE Transactions on Communications*, 43:2959–2965.

Fedorovskaya, E., de Ridder, H., and Blommaert, F. (1997). Chroma variations and perceived quality of color images of natural scenes. *Color Research and applications*, 22:96–110.

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

Gonzalez, R. C. and Woods, R. E. (1992). *Digital Image Processing*. Addison-Wesley publishing company, Inc.

Hanazato, A., Okui, M., Yamanoue, H., and Yuyama, I. (1999). Evaluation of crosstalk in stereoscopic display. *Proceedings 3D Image Conference*, pages 258–263.

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

IJsselsteijn, W., de Ridder, H., and Freeman, J. (2001). Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence. *Presence: Teleoperators and Virtual Environments*, 10:298–311.

IJsselsteijn, W., de Ridder, H., Freeman, J., and Avons, S. (2000a). Presence: Concept, determinants and measurement. *Proceedings of the SPIE*, 3959:520–529.

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

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

IJsselsteijn, W., de Ridder, H., and Vliegen, J. (2000b). Effects of stereoscopic filming parameters and display duration on the subjective assessment of eye strain. *Proceedings of the SPIE*, 3957:12–22.

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

IJsselsteijn, W., Seuntiëns, P., and Meesters, L. (2002). *ATTEST Deliverable 1: State of the art in human factors and quality issues of stereoscopic broadcast television*. Technical report, Eindhoven University of Technology.

ITU (2000a). Methodology for the subjective assessment of the quality of television pictures. *Recommendation BT.500-10*.

ITU (2000b). Subjective assessment of stereoscopic television pictures. *Recommendation BT.1438*.

Janssen, T. and Blommaert, F. (2000). Visual metrics: Discriminative power through flexibility. *Perception*, 29:965–980.

Johanson, M. (2001). Stereoscopic video transmission over the internet. *Proceedings of the IEEE Workshop on Internet Applications*, pages 12–19.

Karunasekera, S. A. and Kingsbury, N. G. (1995). A distortion measure for blocking artifacts in images based on human visual sensitivity. *IEEE Transactions on image processing*, 4:713–724.

Kayargadde, V. and Martens, J.-B. (1996a). Perceptual characterization of images degraded by blur and noise: Experiments. *Journal of the Optical Society of America A*, 13:1166–1177.

Kayargadde, V. and Martens, J.-B. (1996b). Perceptual characterization of images degraded by blur and noise: Model. *Journal of the Optical Society of America A*, 13:1178–1188.

Konrad, J., Lacotte, B., and Dubois, E. (2000). Cancellation of image crosstalk in time-sequential displays of stereoscopic video. *IEEE Transactions on Image Processing*, 9:897–908.

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

Laihanen, P., Lindholm, M., Rouhiainen, S., Saarelma, N., and Tuuteeeri, L. (1994). Automatic color correction. *Proceedings of the Second IS&T/SID Color Imaging Conference: Transportability of Color*, pages 97–102.

Lambooij, M., IJsselsteijn, W., Heynderickx, I., and Seuntiëns, P. (2005). *What is the appropriate attribute to measure 3D quality?* Technical report, Eindhoven University of Technology.

Levelt, W. (1965). *On Binocular Rivalry*. Assen, The Netherlands: Royal VanGorcum.

Libert, J. M. and Fenimore, C. P. (1999). Visibility thresholds for compression-induced image blocking: measurement and models. *Proceedings of the SPIE*, 3644:197–206.

Lipscomb, J. and Wooten, W. (1994). Reducing crosstalk between stereoscopic views. *Proceedings of the SPIE*, 2177:92–96.

Lipton, L. (1987). Factors affecting ghosting in time-multiplexed planostereoscopic CRT display systems. *Proceedings of the SPIE*, 761:75–78.

Lubin, J. (1993). The use of psychophysical data and models in the analysis of display system performance. In Watson, A. B., editor, *Digital Images and Human Vision*, pages 163–178. New York: The MIT Press.

Mansson, J. (1998). *Stereovision: A Model for Human Stereopsis*. Technical report, Lund University Cognitive Studies, Sweden.

Marr, D. (1982). *Vision*. W.H. Freeman and Co, San Francisco.

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

Meesters, L. (2002). *Predicted and perceived quality of bit-reduced gray-scale still images*. Ph.D. thesis, Eindhoven University of Technology, The Netherlands.

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

Mitsuhashi, T. (1996). Evaluation of stereoscopic picture quality with CFF. *Ergonomics*, 39:1344–1356.

Okuyama, F. (1999). Evaluation of stereoscopic display with visual function and interview. *Proceedings of the SPIE*, 3639:28–35.

Palmer, S. (1999). *Vision Science*. MIT Press, Cambridge, Massachusetts, USA.

Pastoor, S. (1995). Wahrnehmungsgrenzen für stationäres Übersprechen. *Internal Report HHI*, BMFT-Vorhaben: 01 BK 101.

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

Perkins, M. G. (1992). Data compression of stereopairs. *IEEE Transactions on Communications*, 40:684–696.

Philips (2004). *Philips Ambilight technology illuminates the future of home entertainment*. Technical report.

Pommeray, M., Kastelik, J., and Gazalet, M. (2003). Image crosstalk reduction in stereoscopic laser-based display systems. *Journal of Electronic Imaging*, 12:689–696.

Roufs, J. A. J. (1992). Perceptual image quality: Concept and measurement. *Philips Journal of Research*, 47:35–62.

Schreer, O., Kauff, P., and Sikora, T. (2005). *3D Videocommunication: Algorithms, concepts and real-time systems in human centred communication*. Wiley.

Seuntiëns, P., Heynderickx, I., and IJsselsteijn, W. (2005a). Viewing experience and naturalness of 3D images. *Proceedings of the SPIE*, 6016:43–49.

Seuntiëns, P., Meesters, L., and IJsselsteijn, W. (2005b). Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation. *ACM Transactions on Applied Perception*, accepted 2005 and to be published 2006.

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

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

Slater, M., Steed, A., and Chrysanthou, Y. (2002). *Computer Graphics and Virtual Environments*. Addison-Wesley Publishing Company, Inc.

Smith, C. and Dumbreck, A. (1988). 3-D TV: The practical requirements. *Television: Journal of the Royal Television Society*, pages 9–15.

Stelmach, L., Tam, W., Speranza, F., Renaud, R., and Martin, T. (2003). Improving the visual comfort of stereoscopic images. *Proceedings of the SPIE*, 5006:269–282.

Stelmach, L., Tam, W. J., Meegan, D., and Vincent, A. (2000). Stereo image quality: Effects of mixed spatio-temporal resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 10:188–193.

Stelmach, L. B. and Tam, W. J. (1998). Stereoscopic image coding: effect of disparate image-quality in left- and right-eye views. *Signal Processing: Image Communications*, 14:111–117.

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

Thurstone, L. (1927). A law of comparative judgment. *Psychological Review*, 34:273–286.

van Berkel, C. and Clarke, J. (1997). Characterization and optimization of 3d-lcd module design. *Proceedings of the SPIE*, 3012:179–186.

van den Branden Lambrecht, C. J. (1996). *Perceptual Models and Architectures for Video Coding Applications*. PhD thesis, Ecole Polytechnique Federale de Lausanne.

Watson, A. B. (1987). Efficiency of a model human image code. *Journal of the Optical Society of America A*, 4:2401–2417.

Westheimer, G. and McKee, S. (1980). Stereoscopic acuity with defocused and spatially filtered retinal images. *Journal of the Optical Society of America*, 7:772–778.

Wheatstone, C. (1838). Contributions to the physiology of vision: Ii. on some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society*, 128:371–394.

Winkler, S. (1999). Issues in vision modelling for perceptual video quality assessment. *Signal Processing*, 78:231–252.

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

Woods, A. and Tan, S. (2002). Characterising sources of ghosting in time-sequential stereoscopic video displays. *Proceedings of the SPIE*, 4660:66–77.

Yamanoue, H., M.Nagayama, M.Bitou, and and, J. (1998). Orthostereoscopic conditions for 3D HDTV. *Proceedings of the SPIE*, 3295:111–120.

Yano, S. (1991). Experimental stereoscopic high-definition television. *Displays*, 12:58–64.

Samenvatting

Sinds de komst van de televisie is er veel onderzoek verricht naar het verbeteren van o.a. kleurweergave, beeldkwaliteit, en geluidskwaliteit. Een nieuwe stap voorwaarts is de introductie van drie-dimensionale televisie waarbij het totale beeld een diepte impressie geeft aan de kijker. Objecten komen uit het scherm of verdwijnen achter het scherm.

In de echte wereld nemen we diepte waar omdat onze ogen gemiddeld 6.5 cm uit elkaar staan. Ieder oog ziet daardoor een ander perspectief van de wereld. Onze hersenen mappen deze twee beelden en extraheren diepte-informatie uit de verschuiving van de twee beelden. Dit principe wordt ook toegepast in de drie-dimensionale televisie waarbij een lenzen systeem ieder oog van een verschillend beeld voorziet.

Bij het optimaliseren van televisiebeelden wordt vaak gekeken naar de algehele beeldkwaliteit. Subjectieve testen met gebruikers worden gebruikt om de optimale instelling van de technische parameters te vinden. Voor de conventionele televisie (twee- dimensionaal) zijn er al modellen beschikbaar die de algehele beeldkwaliteit voorspellen, maar zijn deze modellen toereikend om de drie-dimensionale ervaring te meten? Wij denken dat voor een drie-dimensional model alleen beeldkwaliteit de volledige lading niet dekt maar dat ook factoren zoals de kwaliteit van de gereproduceerde diepte en het visueel comfort een belangrijke rol spelen.

Het doel van dit proefschrift is het begrijpen, meten en het modelleren van de drie-dimensionale visuele ervaring. Dit proefschrift onderzoekt verschillende nieuwe concepten voor de evaluatie van drie-dimensionale beelden en vergelijkt deze met een bestaand beeldkwaliteitsmodel.

In hoofdstuk 2 bekijken we welk evaluatiecriterium het meest geschikt is om drie-dimensionale visuele ervaring the meten. De criteria beeldkwaliteit, diepte, natuurlijkheid, presence, en kijkervaring zijn onderzocht in experiment 1 waarbij verschillende 3D conversie algoritmen zijn getest. Deze algoritmen zijn nog niet perfect en bevatten nog enkele beeldfouten. De resultaten laten zien dat kijkervaring en natuurlijkheid het grootste onderscheidend vermogen hebben tussen 2D en 3D geconverteerde beelden. In experiment 2 leggen we de focus op de criteria kijkervaring en natuurlijkheid en gebruiken we 2D en 3D beeldmateriaal zonder diepte fouten. Als gecontroleerde verstoring hebben we witte ruis gebruikt en resultaten laten zien dat de criteria kijkervaring en natuurlijkheid de ruisverstoring in gelijke maten meewegen. Tevens blijkt dat natuurlijkheid gevoeliger is voor diepte dan kijkervaring.

Hoofdstuk 3 beschrijft experimenten waarbij gekeken wordt naar het effect van symmetrische en asymmetrische JPEG codering (typische 2D verstoring) en camera basis afstand op verschillende evaluatiecriteria. In experiment 3 onderzoeken we het effect van JPEG codering en diepte op de algehele beeldkwaliteit, diepte, waargenomen scherpte, en waargenomen eye-strain (oogpijn). Resultaten laten zien dat een toename van JPEG codering een negatief effect heeft op waargenomen beeldkwaliteit, scherpte en eye-strain maar geen effect heeft op waargenomen diepte. Een toename in camera basis afstand (meer diepte) heeft een effect op waargenomen diepte en eye-strain maar heeft geen effect op algehele beeldkwaliteit en waargenomen scherpte. Het traditionele concept algehele beeldkwaliteit neemt de toegevoegde waarde van diepte dus niet mee. In experiment 4 wordt het criterium beeldkwaliteit vergeleken met het criterium natuurlijkheid. Resultaten laten zien dat natuurlijkheid de zichtbaarheid van de JPEG verstoring en ook de toegevoegde waarde van diepte meeneemt in tegenstelling tot beeldkwaliteit. Beeldkwaliteit neemt alleen de JPEG verstoring mee in het oordeel.

In hoofdstuk 4 onderzoeken we het effect van crosstalk (overspraak, typische 3D verstoring) en camera basis afstand op de criteria beeldkwaliteit, waargenomen diepte, en waargenomen eye-strain. Resultaten van experiment 5 laten zien dat de waargenomen beeldverstoring toeneemt met een stijging in crosstalk en camera basis afstand. Waargenomen diepte en waargenomen eye-strain stijgen alleen met een toenemende camera basis afstand en blijven gelijk met toenemende crosstalk. Resultaten van experiment 6 laten zien dat het criterium natuurlijkheid de crosstalk verstoring en de toegevoegde waarde van diepte meeneemt in tegenstelling tot beeldkwaliteit.

Hoofdstuk 5 gaat over het modelleren van de concepten kijkervaring en natuurlijkheid in termen van beeldkwaliteit en diepte. Visueel comfort speelt ook een belangrijke rol in het nieuwe model, maar additioneel onderzoek is vereist om hierover uitspraken te doen. Het model zal zich dus richten op de bijdrage van beeldkwaliteit en diepte. Het modelleren is gedaan op 5 verschillende data-sets uit dit proefschrift. In vier van de vijf gevallen neemt natuurlijkheid diepte meer mee in de beoordeling dan kijkervaring. Het modelleren laat zien dat het criterium natuurlijkheid de beste balans heeft tussen beeldkwaliteit enerzijds en diepte anderzijds. Hieruit concluderen we dat natuurlijkheid het beste concept is voor de evaluatie van drie-dimensionale beelden. Hoe visueel comfort ingepast moet worden in het model heeft zeker nog meer onderzoek nodig. Ook het effect van bewegende beelden kan zeker van invloed zijn op de beoordeling en moet daarom ook nog nader onderzocht worden.

In hoofdstuk 6 kijken we terug en beschrijven we welke inzichten we hebben verkregen. Ook bespreken we het nieuwe 3D model en de toepasbaarheid ervan.

Summary

Since the introduction of television, much has been done to improve color, picture quality, and sound quality of conventional television. A new step forward is the introduction of three-dimensional television enabling people to watch their content in three dimensions.

In the real world we perceive depth due to the fact that our eyes are separated by 6.3 cm on average. Each eye receives a slightly different perspective of the world. Our brain fuses these two images, and because each image is slightly displaced with respect to the other - a phenomenon known as retinal disparity - the relative and absolute depths of objects in space are perceived. This principle is also applied in the 3D-TV were the left- and right-eye image are separated by optical lenses in the 3D-TV.

The image quality criterion is often used to optimize television sets. Subjective tests (user-tests) are performed to find the optimal settings of the technical parameters. For conventional television (2D), there already exist models predicting the overal image quality, but are these models sufficient to measure the 3-D visual experience? We believe that a 3D visual experience model is required that is multidimensional, incorporating perceptual factors related to reproduced depth, 3D image impairments, and visual comfort.

The research aim of this thesis is how to understand, measure, and model the 3D visual experience. This thesis investigates several new concepts for the evaluation of 3D image material and compares it with an existing image quality model.

The goal of chapter 2 is to explore and determine which evaluation criterion is most appropriate to assess the performance of 3D-display systems. Experiment 1 explores the assessment criteria image quality, depth, nat-

uralness, presence and viewing experience using new 2D-3D conversion techniques (not yet optimal). Results show that viewing experience and naturalness have the most discriminating power between the various 2D-3D conversion algorithms. Experiment 2 focuses on the criteria viewing experience and naturalness and uses 'perfect' 3D content with no conversion or depth artefacts. Several noise levels were added to the 2D and 3D images to degrade image quality. Results show that the noise distortion is weighted equally both with viewing experience and naturalness. Naturalness is more sensitive to depth than viewing experience.

Chapter 3 describes two experiments to investigate the effects of symmetric and asymmetric JPEG coding (typical 2D distortion) and camera base distance on several evaluation criteria. Experiment 3 investigates the effects of camera-base distance and JPEG-coding on overall image quality, perceived depth, perceived sharpness and perceived eye-strain. Results show that an increase in JPEG coding artefacts has a negative effect on image quality, sharpness and eye-strain but has no effect on perceived depth. An increase in camera-base distance (more depth) increases perceived depth and reported eye-strain but has no effect on image quality and perceived sharpness. So, the traditional image quality concept does not take into account the added value of depth Experiment 4 compares the image quality concept and the naturalness concept. Results show that naturalness weighs more equally the visibility of the distortion as well as the added value of depth in contrast to image quality.

Chapter 4 investigates the effect of crosstalk (ghosting, typical 3D distortion) and camera-base distance on the criteria perceived image distortion, perceived depth, and perceived eye-strain. Results of Experiment 5 show that image distortion ratings show a clear increase with increasing crosstalk and increasing camera base distance. Ratings of visual strain and perceived depth only increase with increasing camera base distance and remain constant with increasing crosstalk. Results of Experiment 6 show that naturalness weighs more equally the visibility of the distortion as well as the added value of depth in contrast to image quality.

The goal of chapter 5 is to model the concepts viewing experience and naturalness in terms of image quality and depth. Also visual comfort plays an important role in this model, but additional research is required in this area. The model will only focus on the contribution of image quality and depth. The modeling is based on 5 different data-sets from this thesis.

Results show that in four out of five cases naturalness weighs more depth than viewing experience. The modeling shows that the criterion naturalness seems to have the best balance between image quality and the depth. Therefore, it is concluded that naturalness is the best concept to evaluate the quality of 3D image content.

In Chapter 6, we will briefly look back on the previous chapters and discuss the most important findings. The new 3D visual experience model is presented and we discuss the applicability of the model.

Acknowledgements

I would like to thank all my colleagues at the Human Technology Interaction group and Philips Research for the excellent working atmosphere and the wonderful time together. I acknowledge with gratitude Don Bouwhuis, Ingrid Heynderickx, Wijnand IJsselsteijn and Lydia Meesters for the valuable advice and feedback during this project.

Biography

Pieter J. H. Seuntiëns was born in Eindhoven, The Netherlands, on February 6, 1976. He received the B.Sc. degree in mechanical engineering from Fontys University Eindhoven and the M.Sc. degree in Human-Technology Interaction from the Eindhoven University of Technology, in 1999 and 2002, respectively. In 2002 he started his PhD research at the J.F. Schouten School for User-System Interaction. His research was funded by the IST-2001-34396 European project 'ATTEST' and Philips Research Laboratories, Eindhoven. In close collaboration with the Video Processing and Visual Perception Group of Philips Research his work focuses on perceptual issues and quality metrics regarding two- and three-dimensional imaging systems.