



Deliverable 1

State-of-the-art in human factors and quality issues of stereoscopic broadcast television

Originator: W.A. IJsselsteijn, P.J.H. Seuntjens and L.M.J. Meesters, TU/e



State-of-the-art in human factors and quality issues of stereoscopic broadcast television

August 2002

W.A. IJsselsteijn,
P.J.H. Seuntjens, and
L.M.J. Meesters

© 2002, W.A. IJsselsteijn, P.J.H. Seuntjens, L.M.J. Meesters.

All rights reserved.

Published in August 2002.

State-of-the-art in human factors and quality issues of stereoscopic broadcast television

Deliverable ATTEST/WP5/01

Advanced Three-dimensional Television System Technologies, ATTEST

Project number: IST-2001-34396

Eindhoven University of Technology
Department Technology Management
Den Dolech 2
P.O. Box 513
5600MB Eindhoven
The Netherlands

Executive summary

Stereoscopic television has a long history and over the years a consensus has been reached that a successful introduction of 3-D television broadcast services can only be a lasting success if the perceived image quality and the viewing comfort is at least comparable to conventional 2-D television. In addition, 3-D television technology should be compatible with conventional 2-D television to ensure a gradual transition from one system to the other. This is becoming increasingly feasible because of recent advanced in capture, coding, and display technology. Central to these developments is the viewers' experience which will signify success or failure of the proposed technological innovations. This deliverable describes the state-of-the-art in human factors and quality issues of 3-D broadcast television.

In Chapter 2 we address fundamentals of binocular vision such as stereopsis, and binocular disparity. Additionally, the integration of monocular and binocular depth cues, and individual differences in stereoscopic ability within the population are discussed to provide relevant background information that is needed to understand human factor issues in stereoscopic content generation, compression, display and viewing conditions.

Chapters 3-5 give an overview of current stereoscopic technological developments in content generation, compression and displays. The primary perspective taken in these chapters is technology-centred though various perceptual issues will be addressed as well. In Chapter 3 three approaches to producing stereoscopic content are described: (i) stereoscopic dual-camera approach, (ii) depth-range camera approach, and (iii) 2D-to-3D video conversion approach. We describe the geometry and viewing factors of traditional stereoscopic parallel and converging dual-cameras. The novel content generation concepts of depth-range camera, producing a regular 2D sequence with synchronized depth information, and the 2D-to-3D video conversion are discussed in relation to the approaches taken in the IST ATTEST project.

In Chapter 4 we discuss stereoscopic compression schemes and resulting coding distortions. Most stereoscopic compression schemes utilize traditional 2-D coding techniques in combination with disparity or depth based coding to exploit the image redundancy. Better compression can be achieved by incorporating properties of the human visual system (e.g. spatio-temporal contrast sensitivity, masking and binocular rivalry). In compressed stereoscopic pictures two classes of coding distortions can be distinguished: (i) conventional coding artifacts (e.g. blockiness and blurring), and (ii) artifacts specific to stereoscopic coding (e.g. depth distortions such as the cardboard effect). The physical characteristics of the coding artifacts and experimental results related to these perceived impairments in stereoscopic pictures are reviewed. The goal within ATTEST is to develop a high quality stereoscopic video data-format that is compatible with traditional coding standards MPEG-2/4/7), conventional 2-D and novel 3-D displays.

In Chapter 5 we review time-parallel and time-multiplexed stereoscopic displays using viewing aids (e.g. glasses). In addition, also autostereoscopic display techniques are described such as direction multiplexed, holographic and volumetric techniques. The advantages and disadvantages of the different display techniques are discussed and we conclude with preliminary requirements of a 3-D TV system.

In Chapter 6 we describe subjective testing paradigms to evaluate stereoscopic image systems and present results of subjective assessment studies where human observers judged the perceived sharpness, depth, image quality, naturalness, presence and/or eye-strain of stereoscopic pictures.

Subjective studies indicate that image-size, MPEG-2 coding and low-pass filtering do not affect the perceived depth. However, an increase of perceived depth can be gained by increasing the viewing distance or incorporating motion parallax. Negative effects on the depth percept were reported for inappropriate shooting parameters causing for instance diplopia or a cardboard effect.

Furthermore, several studies showed that human observers prefer stereoscopic pictures over monoscopic pictures. This was demonstrated for unimpaired images and video sequences. Except for MPEG-2 impaired sequences, the same trend was found for impaired pictures. Subjective studies of binocular rivalry, in particular relevant for image compression, showed that the image quality of an asymmetric blurred stereo image pair is dominated by the high image quality component. On the other hand, the perceived image quality of asymmetric processed stereo pairs containing blockiness is an average of the quality of the monoscopic views.

A small deviation between the experienced image quality and naturalness (truthful reproduction) of stereoscopic pictures was found. It seems that observers prefer slightly exaggerated depth even though the appearance of the image is perceived as slightly unnatural. Naturalness and depth are related but can vary independently depending on the scene content and the shooting parameters. Other factors that can affect the perceived naturalness of a stereoscopic picture are: (i) size distortions whereby a 3-D object looks unnaturally small (puppet theatre effect), (ii) lack of motion parallax where the shape of 3-D object can be distorted if the viewers' position deviates from the centre of the screen. Size distortion can be avoided by using an orthoscopic camera configuration and the visual distortion due to a lack of motion parallax can be reduced by multiview stereoscopic displays.

Presence was applied as an evaluation concept to measure the overall psychological impact of a stereoscopic display on the viewer. It was demonstrated that subjects experience a sense of "being there" more for stereoscopic images than for monoscopic images.

Stereoscopic pictures can enhance the viewers experience of perceptual constructs such as image quality, naturalness and presence. However an attenuating factor can be the experienced eye-strain caused by the accommodation-vergence conflict. Additional factors that may increase eye-strain are exaggerated horizontal disparities, the introduction of vertical disparities in converging camera configurations, crosstalk and blur.

Finally, an approach to model the perceived quality of a stereoscopic image system is proposed. This model can be used to describe the relationship between technical parameters, induced by camera configuration, compression and displays, and the underlying attributes of perceived image quality such as sharpness, depth and eye-strain. This model will be used to guide future efforts in the area of stereoscopic quality and human factors.

Contents

1	Introduction	1
1.1	Stereoscopic cinema and television	1
1.2	The ATTEST contribution	3
1.3	Structure of the current deliverable	4
2	Principles of stereoscopic depth perception	7
2.1	Binocular disparity and stereopsis	7
2.2	Cue combination in depth perception	10
2.3	Individual differences	12
3	Content generation	15
3.1	Introduction	15
3.2	Stereoscopic dual-camera video production	15
3.2.1	Stereoscopic video geometry	16
3.2.2	Parallel cameras	16
3.2.3	Converging cameras	18
3.3	Depth range camera	22
3.4	2D-3D video conversion	23
4	Coding of 3-D imagery	25
4.1	Image data redundancy	25
4.2	3-D compression schemes	27
4.2.1	Disparity and depth based coding	27
4.2.2	Mixed resolution coding	29
4.2.3	Multiview coding	30
4.3	Evaluation of compression schemes	31
4.3.1	Conventional coding artifacts	31
4.3.2	Artifacts specific to stereoscopic coding	34

5 Stereoscopic display techniques	37
5.1 Introduction	37
5.2 Stereoscopic displays	37
5.3 Autostereoscopic displays	39
5.4 Preliminary requirements of a 3-D TV system	41
6 Human factors	43
6.1 Subjective assessment methods	43
6.1.1 Explorative study: focus groups	43
6.1.2 Direct ratings of subjective image quality	44
6.1.3 Context effects	46
6.2 Subjective attributes of stereoscopic viewing	47
6.2.1 Perceived sharpness	47
6.2.2 Depth	48
6.2.3 Image quality	49
6.2.4 Naturalness	51
6.2.5 Presence and enjoyment	52
6.2.6 Eye strain	53
6.3 Stereoscopic image quality model	55
6.3.1 Approaches towards image quality modelling	55
6.3.2 ATTEST's quality model for stereoscopic image systems	56
Bibliography	59
Index	67

Chapter 1

Introduction

"I come," cried he, "to proclaim that there is a land of Three Dimensions"

Edwin A. Abbott, Flatland, 1884.

1.1 Stereoscopic cinema and television

The history of stereoscopy (literally "solid sight") can be traced back to 1833 when Sir Charles Wheatstone created a mirror device (see figure 1.1) that enabled the viewer to fuse two slightly different views of the same painting or drawing into one stereoscopic image, resulting in a compelling three-dimensional perception of the original picture¹. In 1838 he presented his classic paper 'On some remarkable, and hitherto unobserved, phenomena of binocular vision' to the Royal Society of London, which discussed his theories on stereopsis. With the realization of still photography in 1839, it was only years before the paintings and drawings were replaced by photographs in his stereoscopic viewing device. In 1844, Sir David Brewster further developed the stereoscope by utilizing prismatic lenses to magnify and fuse the stereo images. Before the American Civil War (mid-19th century) many homes had stereoscopic viewing devices of various sorts (Hayes, 1989). Viewing stereoscopic still images became a popular pastime in both Europe and the US, and from 1860 to the 1930s stereography flourished. Brewster's lenticular stereoscope became a commercial success, selling 250.000 in a short time (Zone, 1996).

The development of motion pictures affected the popularity of the stereoscope, but stereoscopic cinema (early 1900s) and stereoscopic television (1920s) were present at the dawn of their monoscopic counterparts. Although stereoscopic films date back to 1903 when the Lumière brothers made the first 3-D motion picture available to the public, it was not until the 1950s that Hollywood turned to 3-D as the 'next big thing', trying to counteract the dropping box office receipts that occurred as a consequence of the increasing popularity of a competing technology: *television*.

¹Of course, the history of the study of *binocular vision* can be traced back much further, at least to Aristotle (ca. 330 B.C.) who considered that both eyes were moved from a single source and also notes the occurrence of double images, known as *diplopia* (Wade, 1998)

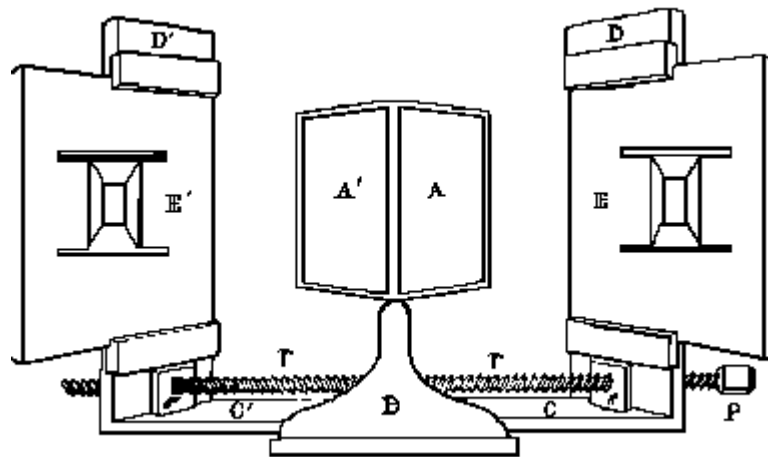


Figure 1.1: The Wheatstone stereoscope. It used mirrors (A' and A) placed at an angle to reflect the left and right eye drawings (E' and E) towards the viewer's eyes.

From 1952 to 1954, with 3-D movies at the height of their popularity, Hollywood produced over 65 stereoscopic films. One of the first 3-D feature films, *Bwana Devil*, released in 1952, became one of the top-grossing films to that time, earning nearly \$100.000 in its first week. Producers seemed well aware of the enhanced psychological impact stereoscopic films had on the viewing public, judging from the catch-phrases used on publicity ads, such as: "You are there!"², "It happens to YOU in three dimensions"³, or the slightly ambiguous "Puts the action right in your lap"⁴, which was used for an X-rated movie (Hayes, 1989). Although such statements were clearly sales pitches of the films' marketing people, they do illustrate the fact that the aim of the cinema experience was to enhance the film's impact and entertainment value by making the viewer feel part of the movie - turning it into a *first-person* experience. Stereoscopic cinema has the ability to generate a compelling sense of physical space, and allows images to emerge from the screen and enter further into the spectator's space than is possible with conventional 2-D or 'flat' cinema. This effect was often exaggerated by throwing or poking objects from the screen at the viewer.

Although many good stereoscopic movies were produced in the 1950s, stereoscopic cinema got a bad reputation with the public because of the discomfort experienced when viewing misaligned and overdone stereoscopic movies. Today, stereoscopic cinema is commercially relatively successful, with 3D-IMAX theatres being perhaps the most well-known exponents.

Already in the 1920s did television pioneers envision stereoscopic television. On August 10 1928, British engineer John Logie Baird demonstrated the principle of stereo-TV before an audience of scientists and representatives of the press at the Baird Laboratories in Long Acre. The first known experimental television 3-D broadcast in the U.S. was on April 29, 1953. While members of the 31st Convention of the National Association of Radio and Television Broadcasters saw a true polarised 3-D program, reportedly *A Time for Beanie*, viewers at home were left to their own non-existent

² *Inferno*, 1953

³ *Man in the Dark*, 1953

⁴ *The Starlets*, 1976

devices and saw a blurred mess (Hayes, 1989). The first "non experimental" 3-D television broadcast occurred some 30 years later with the Super Bowl halftime show and Coca-Cola commercial on NBC, and the Rose Bowl Parade on Fox.

Today, stereoscopic displays are being utilized in many application areas including simulation systems (e.g. flight simulators), medical systems (e.g. endoscopy), telerobotics (e.g. remote inspection of hazardous environments), computer-aided design (e.g. car interior design), telecommunication (e.g. videoconferencing) and entertainment (e.g. 3D-IMAX, VR games). Stereoscopic systems reinforce the perception of physical space in both camera-recorded and computer-generated environments. The capacity to provide an accurate representation of structured layout, distance and shape can be utilized for precise perception and manipulation of objects, and has added value for communication and entertainment purposes.

Although stereoscopic television has received considerable attention in the past, it has not yet delivered on its promise of providing a high-quality broadcast 3-D TV service that guarantees an optimal and strainfree viewing experience. However, a number of converging trends are likely to change this situation, possibly within this decade, making a full 3-D TV application available to the mass consumer market. The following developments are of particular significance:

- New camera technologies and coding algorithms are being developed that allow for efficient content generation and coding of 3-D images, as well as 2D-3D conversion, and computer animation/motion picture hybrids.
- The advent and increased acceptance of digital TV transmission in Europe (Meyer and Fontaine, 2000) enables broadcasters to transmit two synchronized digital channels (one for the left and one for the right eye) in bandwidths smaller than those utilized by one analogue TV channel.
- Stereoscopic display technology has evolved greatly over the past decades (Okoshi, 1980; Pastoor and Wöpking, 1997; Sexton and Surman, 1999), with particularly promising recent developments in the area of multiview autostereoscopic displays.
- The growing interest in stereoscopic broadcast television services has stimulated a number of laboratories, most notably in Japan, Canada, Germany, France, UK, and The Netherlands, to investigate the human factors requirements necessary for high quality stereoscopic television systems. A significant number of these studies will be reviewed in the current deliverable, particularly in Chapter 6.

At present, 3-D TV is considered to be the logical next step complementing HDTV, and will generate a potentially huge replacement market for current 2-D TV sets.

1.2 The ATTEST contribution

Several projects funded by the European Commission (e.g. COST 230, RACE-II DISTIMA, ACTS PANORAMA, ACTS MIRAGE, ACTS TAPESTRIES) have been aimed at developing and evaluating the standards, technology, signal processing and content production facilities needed for stereoscopic broadcast services. The *IST ATTEST* project aims to build on this considerable knowledge

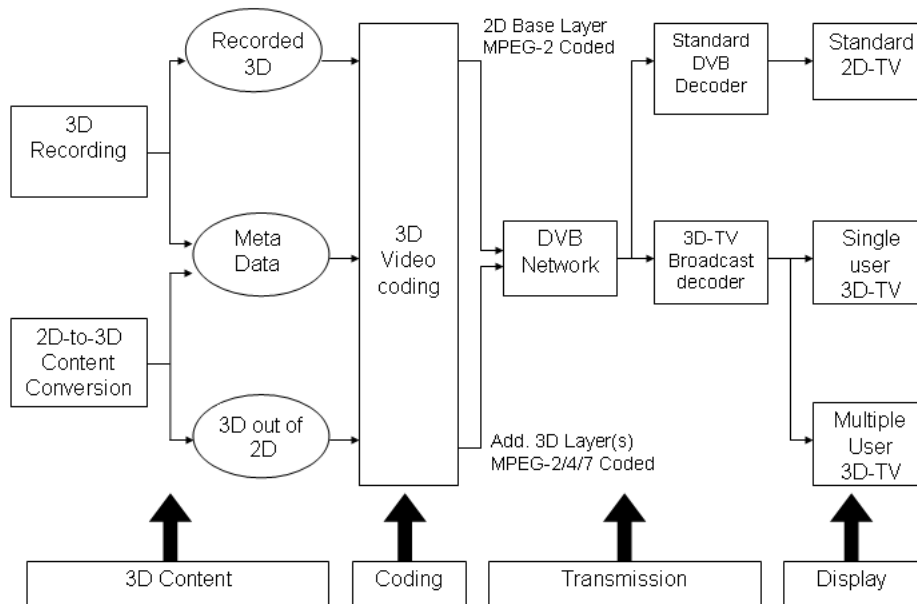


Figure 1.2: The ATTEST 3-D video processing chain

base, and make significant advances in developing a flexible, 2D-compatible, and commercially feasible end-to-end 3-D broadcast chain that includes 3-D (image plus depth) content generation, coding, transmission, and display. By incorporating the complete broadcast chain, the ATTEST project will be at the forefront of the various trends mentioned earlier. Research into the human perceptual and usability requirements will play a central role throughout the project. Particular contributions of the ATTEST project include a novel 3D-video camera system, fast 2D-3D conversion algorithms, a flexible and scalable syntax for image-based 3-D data representation, and autostereoscopic single and multiviewer displays (see figure 1.2).

The acceptance, uptake and commercial success of any advanced technology aimed at the consumer market depends to a large extent on the users' experiences with and responses towards the system. Therefore, it is vital to have a clear understanding of the in-the-home viewing experience of 3-D TV, both looking at the potential added value of the 3-D TV systems, as well as the potential drawbacks for users. The aim of Workpackage 5 of the ATTEST project is to arrive at a set of requirements and recommendations for an optimal 3-D TV system, and to contribute to each individual step in the ATTEST 3-D video chain through perceptual and usability evaluations of the proposed technological innovations.

1.3 Structure of the current deliverable

The current deliverable describes the state-of-the-art in human factors and perception studies relevant to the optimisation of stereoscopic broadcast applications. First, in Chapter 2, we present a brief

introduction into binocular vision and stereopsis, thus providing relevant background information to inform our understanding of the principles and human factors issues at play in stereoscopic content generation, compression, display, and viewing.

Next, Chapter 3 reviews three different approaches to stereoscopic content generation: (i) using a traditional stereo camera pair which results in separate left and right views, (ii) using a depth range camera which generates a 2-D image plus a depth map, and (iii) converting existing 2-D video material into stereoscopic 3-D by computing a depth map from the 2-D image sequences, and subsequently augmenting the depth in the 2-D image accordingly.

Chapter 4 will go into the different stereoscopic image compression techniques available in order to save valuable transmission bandwidth. As different compression methods can introduce different artifacts and image degradations, it is of clear importance to apply image compression in a way that is informed a priori by knowledge of the human perceptual system, as well as empirical formative and summative human factors evaluations.

In Chapter 5 we will briefly review the different stereoscopic display technologies available today, and discuss their advantages and disadvantages. Although many human factors issues are addressed in Chapter 3-5, the primary perspective taken in these chapters is *technology-centred*.

In Chapter 6 we review studies that have investigated viewer's responses to stereoscopic still images or image sequences, either compressed or uncompressed, from a *user-centred* perspective. That is, we focus on formal investigations that relate subjective attributes, such as perceived image quality, depth, naturalness, sharpness, eye strain, and presence, to various image, display, and viewing parameters.

Chapter 6 concludes with a first description of a draft stereoscopic image quality model, which connects the different technical parameters to viewer's subjective responses. This draft model will be used to guide future research efforts in the area of stereoscopic image evaluation, and will be fine-tuned accordingly as more empirical data will become available. An index of key-terms is added at the end of the deliverable.

Chapter 2

Principles of stereoscopic depth perception

This chapter presents an overview of the basic principles underlying human stereoscopic depth perception. In particular, we will discuss binocular disparity and stereopsis (section 2.1), the effectiveness and integration of depth cues in providing metric estimates of depth (section 2.2), and individual differences in stereoscopic ability (section 2.3).

2.1 Binocular disparity and stereopsis

When considering three-dimensional vision, perceptual psychologists are faced with an interesting paradox. How can the essentially two-dimensional mosaic of retinal receptors curved around the back of an eyeball give rise to the perception of a three-dimensional world? In addressing this problem, psychologists have adopted the empiricist notion that complex ideas have to be built up through the association of simpler ones. Therefore, it is assumed that the third dimension is reconstructed by the brain through associating visual sources of information, often called *cues*, with information from other modalities, most notably touch (Bruce et al., 1996).

Complex, natural scenes contain a wide variety of visual cues to depth. Our visual system utilizes monocularly available information such as accommodation, occlusion, linear and aerial perspective, relative size, relative density, and motion parallax, as well as the binocular depth cues of vergence and disparity to construct a perception of depth. The effectiveness of monocular cues is illustrated by the fact that we can close one eye and still have a considerable appreciation of depth. On the other hand, random-dot stereograms or Julesz patterns (Julesz, 1971) demonstrate that in the absence of consistent monocular information, binocular disparity alone provides the visual system with enough information to extract depth information.

Binocular disparity is available because the human eyes are horizontally separated (on average approximately 6.5 cm) which provides each eye with a unique viewpoint unto the world (see figure 2.1). This horizontal separation causes an interocular difference in the relative projections of monocular images onto the left and right retina. When points from one eye's view are matched to corresponding points in the other eye's view, the retinal point to point disparity variation across the

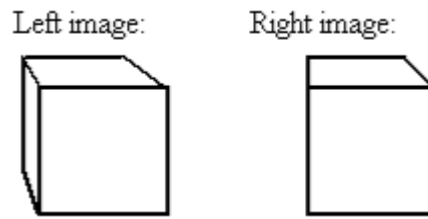


Figure 2.1: Different perspectives from the left and right eye onto an object

image provides information about the relative distances of objects to the observer and the depth structure of objects and environments.

The line that can be drawn through all points in space that stimulate corresponding retinal points for a given degree of convergence is called the *horopter* (see figure 2.2). The theoretical horopter is a circle, known as the *Vieth-Muller circle*, which passes through the fixation point and the nodal points of both eyes. Experimental work has shown that the empirical horopter lies slightly behind the theoretical horopter, although no satisfactory explanation has yet been provided to account for this fact. However, differences between the theoretical and the empirically measured horopter are small and can usually be ignored for practical purposes (Palmer, 1999).

Points which are not on the horopter will have a retinal disparity. Disparities in front of the horopter are said to be *crossed* and disparities behind the horopter *uncrossed*. As long as the disparities do not exceed a certain magnitude, the two separate viewpoints are merged into a single percept (i.e. *fusion*). The small region around the horopter within which disparities are fused is called *Panum's fusional area* (see figure 2.2).

If disparity is large the images will not fuse and double images will be seen, a phenomenon that is known as *diplopia*. The largest disparity at which fusion can occur is dependent on a number of factors. Duwaer and van den Brink (1981) showed that the diplopia threshold is dependent on the subject tested, the amount of training a subject received, the criterion used for diplopia (unequivocal 'singleness' of vision vs. unequivocal 'doubleness' of vision), and the conspicuousness of the disparity.

Although limits vary somewhat across studies, some representative disparity limits for binocular fusion can be given. For small sized stimuli (i.e. smaller than 15 minutes of arc) the range of disparity for the foveal area is about ± 10 minutes of arc, while at 6 degrees eccentricity the range increases to around ± 20 minutes of arc. For large stimuli (1.0 - 6.6 degrees) the range for the foveal region is about ± 20 minutes of arc, i.e. two times the range for smaller stimuli (Patterson and Martin, 1992). However, most individuals have an appreciation of depth beyond the diplopia threshold, i.e. the region where single vision has been lost. Up to 2 degrees of overall disparity between two images is tolerated before the sensation of depth is lost (Howard and Rogers, 1995).

In normal vision, the oculomotor mechanisms, i.e. accommodation and convergence, work in concert with stereopsis. Thus, when viewing an object the eyes converge on it, so that the disparities providing depth information about the object and its environment fall within Panum's fusional area. The eye's lens automatically focuses (accommodation) on the currently fixated object, making it stand out against its surroundings. Thus, double images in front or behind the plane of fixation tend to be out of focus and will 'disappear' in increasing optical blur. Accommodation and convergence

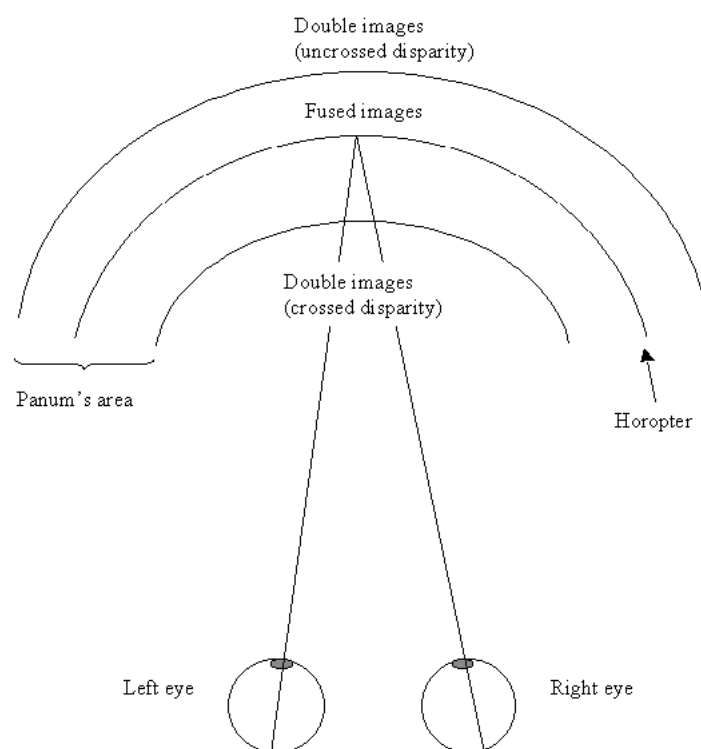


Figure 2.2: Panum's fusional area. Points within Panum's area are fused perceptually into a single image. Points that are closer or farther produce double images of crossed or uncrossed disparity (figure adapted from Palmer (1999), page 209).

operate in a closely coupled fashion, which implies that under natural viewing conditions all objects which are in focus are also fixated upon, thus bringing them within the limits of fusion. However, most of the currently available stereoscopic display techniques do not support the linkage between accommodation and convergence, forcing the observer to focus at a fixed distance (i.e. on the screen plane where the image is sharpest) irrespective of the fixation distance (i.e. where the object is located in depth according to the disparity information). This decoupling of accommodation and convergence has been suggested as a potential cause of visual strain. This topic will be addressed in more detail in Chapter 6 of this deliverable.

Stereoscopic vision greatly enhances our ability to discriminate differences in depth, particularly at close distances. Stereoscopic sensitivity is remarkably good at regions close to the horopter: people can detect a disparity of around 2 seconds of arc. Because such stereoacuties are significantly smaller than the smallest photoreceptors in the human retina, which are approximately 25-30 seconds of arc in diameter, stereoscopic acuity is known as a *hyperacuity*. Better stereo-acuity performance is reported for crossed than for uncrossed disparities (Woo and Sillanpaa, 1979). The only monocular cue that provides a similar degree of depth resolution is motion parallax created by moving the head from side to side (Howard and Rogers, 1995). In fact, monocular motion parallax and binocular disparity are closely related, since temporally separated successive views can in principle provide the same information to the visual system as spatially separated views (Rogers and Graham, 1982). Estimates of the effective range of stereopsis vary across the literature, but it is clear that stereoscopic information becomes less effective as distance increases and retinal disparities become smaller. For distances beyond 30 meters disparities become negligible.

A mechanism that is of particular interest to stereoscopic coding (see Chapter 4) is how the different inputs to the two eyes are combined to form a single percept. As was discussed earlier, with similar images presented to each eye, fusion will occur, as long as the disparities fall within certain boundaries. There is said to be *binocular summation* when visual detection or discrimination is performed better with two eyes than with one eye. However, if the stimuli presented to the two eyes are very different (e.g. large differences in pattern direction, contour, contrast, illumination, etc.), the most common percepts are binocular mixture, binocular rivalry and suppression, and binocular luster. Mixture typically occurs for instance when a uniform field in one eye is combined with a detailed stimulus in the corresponding part of the other eye. When corresponding parts of the two retina's receive very different high contrast images, then *binocular rivalry* may occur. In a way, this is the opposite of fusion, as the two monocular images will alternate repetitively, in whole or in part, with the unseen portion somehow suppressed. The stimulus seen at any given time is called the *dominant* stimulus. *Binocular luster* occurs in areas of uniform illumination in which the luminance or color is different in the two eyes. In this response, the images are stable and fused, yet appear to be shimmering or lustrous and cannot be localised in depth.

2.2 Cue combination in depth perception

Traditionally, monocular and binocular depth cues that are important in achieving a representation of distance, depth and spatial structure have been studied intensively in isolation. This research was carried out under the assumption that depth is processed in separate modules that correspond to different sources of three-dimensional information. However, it is also interesting to find out how

these modules, if independent, might be integrated to provide a coherent 3-D percept (Johnston et al., 1993; Landy et al., 1995; Landy and Brenner, 2001). The relative efficiency and robustness of our perception of the natural visual world makes it clear that we integrate multiple sources of information into a single percept. Further interest in cue integration has been stimulated by recent developments in computer vision, where shape-from-X algorithms (shape from shading, texture, stereo, motion, contour, etc.) can become more robust when processing cues in combination, and guidance has been sought in how biological vision accomplishes this feat.

In any parallel cue system, cues may act in concert with one another or can be in conflict. There are a number of ways in which different sources of information may combine (Howard and Rogers, 1995):

- *Cue averaging and summation*: for nonintensive sensory dimensions (e.g. direction, orientation), the most efficient way to combine cues is *weighted linear combination*. That is, the independent depth estimates from each cue or depth module are linearly combined with differential weights assigned to each cue. This form of interaction has been demonstrated experimentally on numerous occasions (van der Meer, 1979; Bruno and Cutting, 1988; Johnston et al., 1993, 1994; Frisby et al., 1996). For intensive sensory dimensions (e.g. distance), a summation of cues may yield a more accurate perception, if judgements based on individual cues tend to be underestimations. With reasonably accurate individual cues however, summation will lead to a large overestimation, and taking the mean rather than the sum will then be better.
- *Cue dominance*: judgements are based on only one cue, where the other cue is being suppressed when in conflict. An example of such a situation in the context of stereoscopic displays is the screen edge effect when a stereoscopic image is presented in front of the screen plane. The occlusion from the screen border will dominate the depth percept, and make the image seem to curve backwards at its edges.
- *Cue dissociation*: each cue may be interpreted as arising from a different object. For example, when the spatial separation of signals to the auditory and visual system of one object exceeds a certain angle, two objects may be perceived instead of one, one being visual and the other auditory. A well-known instance is seeing a jet airplane fly at a different location overhead from which the sound seems to originate.
- *Cue reinterpretation*: one of the cues may be interpreted differently after combination to render it compatible with the other. An example of such a process is the *kinetic depth effect*, i.e. when the silhouette of a rotating object, such as a bent piece of wire, appears three-dimensional even without the disparity cue, yet appears two-dimensional when the motion stops.
- *Cue disambiguation*: this could be regarded as a special case of cue reinterpretation, where the sign of a cue may be ambiguous (e.g. whether the object is in front or behind the fixated object), and requires supplementary information from another cue to be interpreted. An example would be image blur, which provides information about the distance of an object from a fixated object, without indicating which is nearer. Other cues, such as occlusion, familiar size, or linear perspective may act as disambiguators.

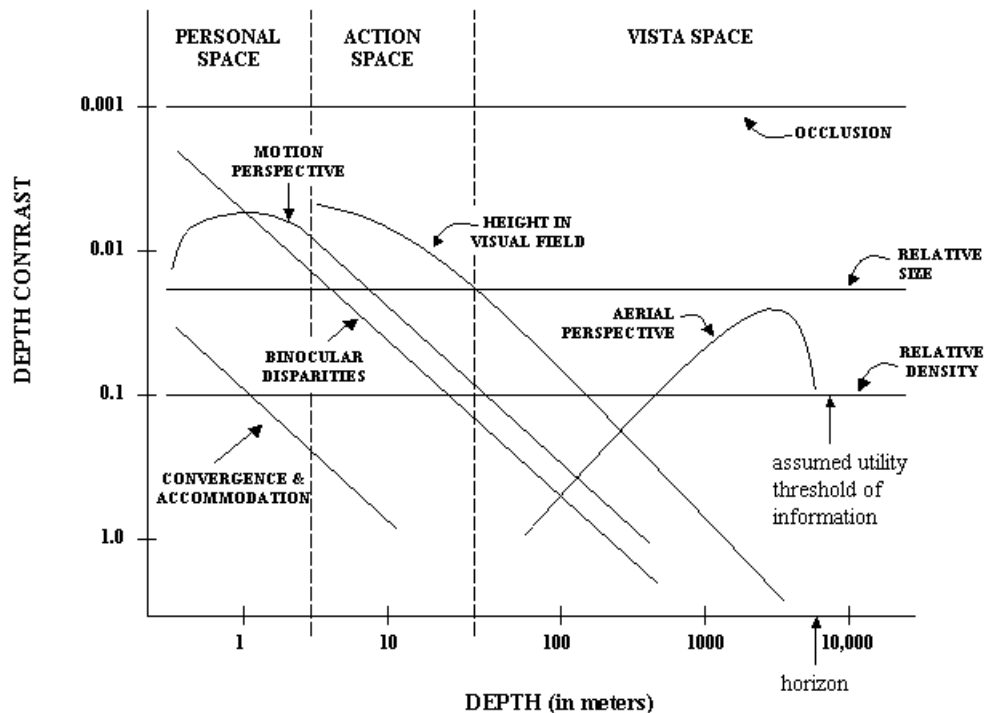


Figure 2.3: The relative depth contrast of different depth cues (figure adapted from Cutting and Vishton (1995))

Generally speaking, the more *consistent* cues, the better an accurate depth percept can be established. Cutting and Vishton (1995) provide a thorough discussion of the relative information potency of depth cues at various distances (see figure 2.3). Each cue on its own is an ambiguous indicator of distance, layout and surface structure. However, by combining several sources of information this ambiguity may be reduced, even to the point of near-metric accuracy.

2.3 Individual differences

The mechanism that underlies stereopsis is extremely precise, but also somewhat fragile. Visual disorders in early childhood, even if only temporary, may result in *stereoblindness*, which is estimated to affect 5-10% of the population. Richards (1970) did a survey of 150 members of the student community at MIT and found that 4% of the population was unable to use the depth cue offered by disparity and 10% had great difficulty deciding the direction of a hidden Julesz figure (i.e. a random-dot stereogram) relative to the background.

More recent studies have shown that the performance of observers on tests of stereoanomaly depends to a large extent on the duration the observer is allowed to look at the test figures. The

proportion of the population unable to perform a depth discrimination task decreases with increasing display duration (Patterson and Fox, 1984; Tam and Stelmach, 1998a).

The single most common cause of stereoblindness is *strabismus*, the misalignment of the two eyes. If this condition is surgically corrected at an early age, stereoscopic vision may develop normally. Stereopsis also fails to develop when children have good vision in only one eye, due to monocular nearsightedness or a cataract. Obviously, stereoblind individuals can still perceive depth, but are restricted to information sources other than binocular disparity.

It is well-known that visual abilities change with age. The major age differences in visual functioning are the result of two types of changes in the structure of the eye. The first type of change is related to the transmissiveness and accommodative power of the lens that begins to manifest itself between the ages of 35 and 45. This affects binocular depth perception, sensitivity to glare, and color sensitivity. The second type of change occurs in the retina and the nervous system, and usually starts to occur between 55 and 65 years of age. This affects the size of the visual field, sensitivity to low light levels, and sensitivity to flicker (Hayslip Jr. and Panek, 1989).

A recent study by Norman et al. (2000) demonstrated that older adults were less sensitive than younger adults to perceiving stereoscopic depth, in particular when image disparity was higher. Overall however, older adults performed to 75% of their expected depth intervals, demonstrating that their stereoscopic abilities have largely been preserved during the process of aging.

Chapter 3

Content generation

3.1 Introduction

In this chapter we will present an overview of three different approaches to producing stereoscopic content:

1. using a stereo camera pair which results in separate left and right views (section 3.2),
2. using a depth range camera which generates a 2-D image plus a depth map (section 3.3), and
3. converting existing 2-D video material into stereoscopic 3-D by computing a depth map from the 2-D image sequences, and subsequently augmenting the depth in the 2-D image accordingly (section 3.4).

Sections 3.3 and 3.4 describe the content generation approaches taken in the IST ATTEST project.

3.2 Stereoscopic dual-camera video production

Stereoscopic video can be produced by using a stereoscopic camera, consisting of a co-planar configuration of two separate, monoscopic cameras, each corresponding to one eye's view. The aim of stereoscopic video is to capture real world scenes in such a way that viewing them stereoscopically accurately simulates the information the viewer would receive in the real life situation. To achieve this, the stereoscopic filming process must be carefully planned and executed. Stereoscopic parameters need to be calculated and adjusted such that they accurately capture the disparities a viewer would receive from the actual scene¹ and do not cause any visual discomfort. The resulting three-dimensional scene must be free from visual errors and potentially conflicting information. Care needs to be taken to assure that the two stereoscopic images are calibrated in terms of contrast, brightness, colour, and sharpness of focus, and are also geometrically calibrated, i.e. without rotation, vertical

¹Although, as we shall see in Chapter 6, slight exaggerations of what is deemed to be natural can sometimes be preferred

displacements, etc. Perceptual issues that may occur as a consequence of changes in camera settings are discussed in Chapter 6.

Depending on the stereoscopic encoding format chosen, the left and right video signals can be kept separate through the entire signal path (i.e. image pickup, encoding/storage, and display) or can be stored on a single recording device using field-sequential or side-by-side 3-D video formats Woods et al. (1996).

In the next section, we will discuss the mathematical basis for calculating the horizontal disparities in a dual-camera set-up, both for parallel and converging cameras. Subsequently we will describe some of the viewing factors that have an impact on stereoscopic image perception, as screen size, viewing distance, and the viewer's eye separation all influence the effective visual disparities.

3.2.1 Stereoscopic video geometry

The mathematical basis for stereoscopic image formation was first described by Rule (1941). In the 1950s, the seminal paper by Spottiswoode et al. (1952) extended this work considerably, introducing new concepts such as the *nearness factor*², which facilitated the communication between the film director and the stereo-technician. Various accounts of the geometry of stereoscopic camera and display systems have been published since (e.g. Gonzalez and Woods (1992), Woods et al. (1993), Franich (1996)), both for the parallel camera configuration, where the optical axes of both monoscopic cameras run parallel, and for the converging camera configuration, where the optical axes of the cameras intersect at a convergence point. The equations by which the horizontal disparities can be calculated for both types of camera configurations are discussed next, based on the calculations by Gonzalez and Woods (1992) and Franich (1996).

3.2.2 Parallel cameras

Figure 3.1 illustrates the parallel camera set-up, where the optical axes of both cameras run parallel, equivalent to viewing a point at infinity. Included in the figure are the video capture variables that have an impact on the horizontal (left-right) image disparity captured at the camera's imaging sensors. The magnitude of disparity is the distance (in metres) between X-coordinates of homologous points in the left and right image, if the two camera planes were superimposed on each other.

As can be seen from the figure, a real-world point w , with coordinates X , Y , and Z , will be projected onto the left and right camera imaging sensors, and the horizontal disparity will be scaled by the following factors: the camera base separation, or *baseline*, B , the focal length of the lenses of the cameras, λ , and the distance from the cameras to the real world point, Z_w . The image capture process for each of the two cameras consists of a translation along the X -axis followed by a perspective transformation (Franich, 1996; Gonzalez and Woods, 1992).

The camera projection coordinates (left camera: x_l, y_l ; right camera: x_r, y_r) of real world point $w(X, Y, Z)$ are given by:

²The nearness factor is the ratio of the observer to screen distance to the observer to a point on the screen



$$y_r(Y, Z) = \lambda \frac{Y}{\lambda - Z} \quad (3.4)$$

As one can determine by comparing equations 3.2 and 3.4, the real world point (X,Y,Z) is projected onto the same Y coordinate for both left and right cameras. Stated differently, when properly

aligned, no vertical image displacement occurs with a parallel camera configuration. The horizontal disparity, $d_{h,p}(Z)$, can be obtained by subtracting x_l , the X coordinate of the left camera projection, from x_r , the X coordinate of the right camera projection:

$$d_{h,p}(Z) = x_r(X, Z) - x_l(X, Z) \quad (3.5)$$

By substituting 3.1 and 3.3 into 3.5 we obtain the following equation:

$$d_{h,p}(Z) = \lambda \frac{-B}{\lambda - Z} \quad (3.6)$$

From this equation we can deduce that the left-right camera plate disparity will increase as the camera base separation, B , increases, and as the focal length, λ , increases. The horizontal disparity will decrease as the distance to the real-world point, Z , increases. For points infinitely far away along the Z axis disparity becomes equal to zero.

3.2.3 Converging cameras

For a converging (or toed-in) camera set-up the optical axes of both cameras intersect at a convergence point, scaled by the camera base separation (B) and the convergence angle (β) of the camera imaging planes, such that:

$$Z_{conv} = \frac{B}{2 \tan(\beta)} \quad (3.7)$$

The converging camera set-up is illustrated in figure 3.2. The projection onto the camera imaging plane of a real world point (X, Y, Z) is slightly more complex than in the parallel set-up (which actually is a special case of the converging set-up where β equals 0). As the camera axes are no longer parallel to the Z -axis, both a horizontal translation (along the X -axis) and a rotation (around the Y -axis) are required, followed by a perspective projection (Franich, 1996). For the left camera, the horizontal translation is as follows:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \xrightarrow{\text{translation}} \begin{bmatrix} X + \frac{B}{2} \\ Y \\ Z \end{bmatrix} \quad (3.8)$$

A rotation of the translated point around the Y -axis by an angle of $+\beta$ gives:

$$\begin{bmatrix} X + \frac{B}{2} \\ Y \\ Z \end{bmatrix} \xrightarrow{\text{rotation}} \begin{bmatrix} \cos(\beta)(X + \frac{B}{2}) - \sin(\beta)Z \\ Y \\ \sin(\beta)(X + \frac{B}{2}) + \cos(\beta)Z \end{bmatrix} \quad (3.9)$$

Finally, perspective projection (Franich, 1996; Gonzalez and Woods, 1992) gives the camera plane coordinates x_l and y_l for the left camera:

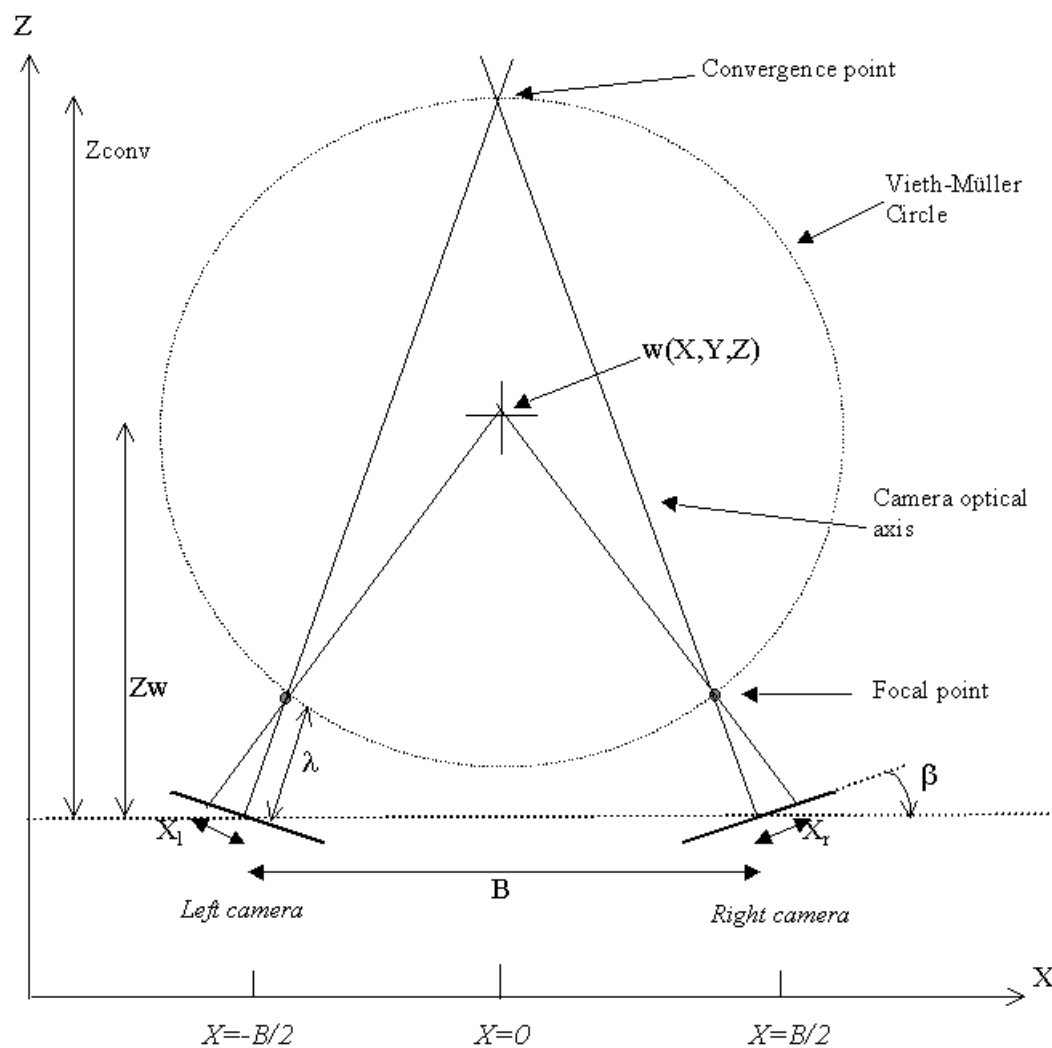


Figure 3.2: Converging camera geometry

$$x_l(X, Z) = \lambda \frac{\cos(\beta)(X + \frac{B}{2}) - \sin(\beta)Z}{\lambda - \sin(\beta)(X + \frac{B}{2}) - \cos(\beta)Z} \quad (3.10)$$

$$y_l(X, Y, Z) = \lambda \frac{Y}{\lambda - \sin(\beta)(X + \frac{B}{2}) - \cos(\beta)Z} \quad (3.11)$$

Similarly, the camera plane coordinates for the right camera can be found:

$$x_r(X, Z) = \lambda \frac{\cos(\beta)(X - \frac{B}{2}) + \sin(\beta)Z}{\lambda + \sin(\beta)(X - \frac{B}{2}) - \cos(\beta)Z} \quad (3.12)$$

$$y_r(X, Y, Z) = \lambda \frac{Y}{\lambda + \sin(\beta)(X - \frac{B}{2}) - \cos(\beta)Z} \quad (3.13)$$

By subtracting the left camera coordinates from the right camera coordinates it is possible to determine the horizontal disparity for the converging camera configuration. For $\beta=0$, the converging camera configuration reverts to the parallel configuration, and the mathematical expression for horizontal disparity can be simplified to that described in expression 3.6.

For the converging camera set-up there is also a vertical disparity component, which can be calculated separately (see Franich (1996)). This vertical component is equal to zero in the plane defined by $X=0$ as well as for the plane defined by $Y=0$. It is largest at the corners of the image. Vertical parallax causes what is known as *keystone distortion*, a depth plane curvature which makes objects at the corner of the stereoscopic image appear further away than objects at the centre of the image (Woods et al., 1993). When properly calibrated, the parallel camera set-up does not produce vertical disparities, and as a consequence no keystone distortion will occur.

With the stereoscopic formulae discussed here it is possible to determine precisely the amount of disparity captured in the filming process, translating real-world object space (X, Y, Z) into CCD³ (or camera plane) coordinates (x_l, y_l) and (x_r, y_r) . When stereoscopic images are presented on a display, the CCD coordinates need to be transformed into screen coordinates. This is achieved by multiplying the CCD coordinates by the screen magnification factor M , which is the ratio of the horizontal screen width to the camera sensor width⁴. The subsequent perception of the image space by the viewer is dependent on a number of viewing factors, discussed next.

Stereoscopic viewing factors

Figure 3.3 illustrates the viewing parameters that play a role in the perception of stereoscopic images. The disparities presented to the observer are scaled by the size of the screen, the distance of the observer to the screen, and the observer's eye separation or inter-pupillary distance (IPD). Variations in IPD for adult males range from 5.77cm to 6.96cm, with a median of 6.32cm (5th to 95th percentile range, Woodson (1981)). Given this range of IPDs, it is advisable to produce and

³Charge Coupled Device - the camera imaging sensor

⁴In fact, the screen parallax values are of opposite sign to the CCD disparity values as the projection onto the camera's image plane occurs in mirror image through the lens

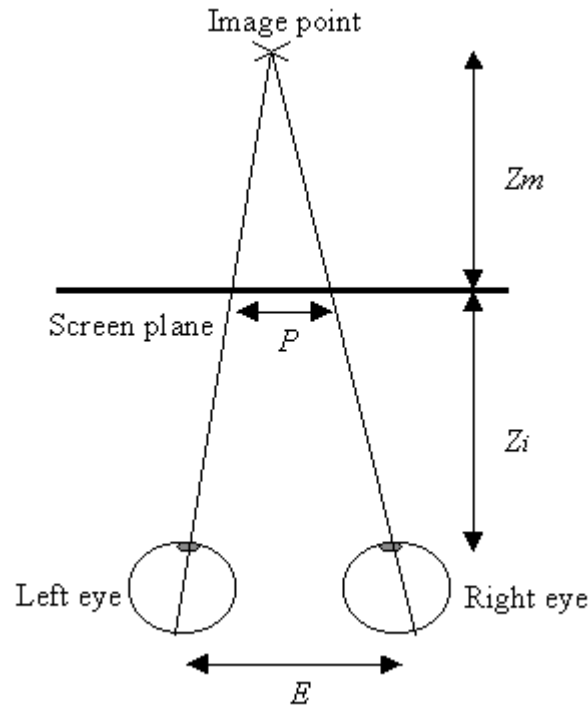


Figure 3.3: Stereoscopic viewing factors

display stereoscopic video in such a way that people with small eye separations can still comfortably fuse the stereoscopic scene. There is no point in increasing the on-screen parallax (P) beyond eye separation (E). When $P=E$ the observer's eyes are looking straight ahead, which corresponds to looking at an image infinitely far away. If the eyes need to diverge to fuse an image, this requires an unusual muscular effort which may cause discomfort.

As is illustrated in figure 3.3, the observed depth Z_{obs} of an image point projected with screen parallax P is equal to the sum of the distance between the observer and the screen, Z_i , and an offset term, Z_m . Through the use of similar triangles we obtain the following equation:

$$\frac{P}{Z_m} = \frac{E}{Z_i + Z_m} \quad (3.14)$$

From this expression Z_m can be deduced:

$$Z_m = \frac{PZ_i}{E - P} \quad (3.15)$$

Since observed depth is equal to $Z_i + Z_m$, we can express Z_{obs} as follows:

$$Z_{obs} = Z_i + \frac{PZ_i}{E - P} \quad (3.16)$$

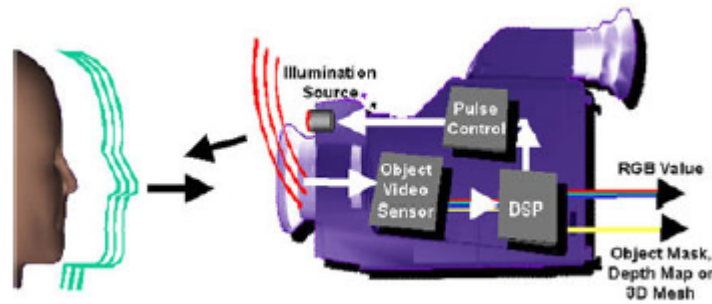


Figure 3.4: ATTEST 3-D camera

This expression can be reduced to:

$$Z_{obs} = Z_i \frac{E}{E - P} \quad (3.17)$$

Based on this equation it follows that as the observer moves closer to the screen (i.e. a decrease of Z_i), perceived depth will decrease. A reduction of screen size (and thus on-screen parallax) at a given viewing distance would have a similar effect. When the on-screen parallax P is equal to zero, the perceived depth is equal to the screen distance Z_i . It should be noted that as P becomes more negative (i.e. eventually projecting crossed disparities), the viewer will increasingly have to look cross-eyed and binocular fusion will become uncomfortable.

3.3 Depth range camera

It is clear, that the success of any future 3-D TV broadcast system will depend to a great extent on the timely availability of sufficient exciting and/or interesting 3D-video material. Therefore, methods for the fast and cheap generation of suitable 3D-video content must be developed. For the generation of new 3D-video content, ATTEST will convert an existing IR range camera, the so-called ZcamTM, into a full 3-D camera. The ZcamTM works like an infrared LIDAR (laser radar without mechanical scanner) system integrated into a traditional video capture device, and is designed as an add-on to an existing broadcast camera. The concept is based on generating a light wall. As the light wall hits the objects it is reflected back towards the camera carrying an imprint of the objects. The information from the imprint can be used to reconstruct a depth map. ATTEST will improve the original ZcamTM to deliver a high depth and pixel resolution 3-D camera, fit for indoor use. The output of the ATTEST 3-D camera will consist of two separate video streams: broadcast quality RGB and associated depth with pixel resolution accuracy. The camera is depicted in figure 3.4.

This will require a breakthrough in depth and pixel resolution, which can be achieved with solid-state shutters, developed by 3DV systems. First test with new shutter systems show significant improvement compared with the existing shutters. An infrared transmitter with limited frequency and an opening angle of 40 degrees to reduce the infrared background intensity will be developed and implemented. An important issue is the depth resolution of the 3-D camera. This depends on

the size of the measured volume (dynamic range) and the bit-depth of the CCD used. The redesign of the depth camera requires adapted optics, mechanics and electronics.

3.4 2D-3D video conversion

It seems clear, that the need for 3D-video content can only partially be satisfied with newly recorded material. Therefore, ATTEST will also develop depth reconstruction algorithms that can be used to convert existing 2D-video material into 3-D. Conversion of existing 2-D video material is a challenging task, because of problems with pixel accurate automatic video segmentation. The conversion of still and motion cues into depth measures for the segments is achieved by motion estimation and segmentation. Problems occur with the conversion of existing 2-D content because the scenes were shot using 2-D filming conventions (pans, zooms, cuts, etc.) For example, a sudden change from a wide-angle shot to close-up will also introduce significant changes in image parallax. In these situations a perceptually acceptable depth interpretation will be provided.

Conversion tools will be provided for use at the broadcaster and content provider side. These will primarily be used to convert popular movies and impressive documentaries. As there aren't any real-time constraints in this case, all available video data can be used in the computations and 3-D information can be integrated over a whole shoot, resulting in high quality 3-D reconstructions. Therefore, the ATTEST approach will build further upon techniques developed for 3-D modelling from image sequences. For the consumer side, ATTEST will develop on-line conversion methods (e.g. for processing in a set-top box), that will allow the viewer to augment any suitable, incoming 2-D broadcast to 3-D. In this case, computations can only be based on video frames that have already been received.

The aim is a 3-D video format (both for the new recordings and for the 2D-3D conversion) that can be distributed over existing broadcast channels, within the existing digital TV standards.

Chapter 4

Coding of 3-D imagery

The basis of image compression is to transform images into less bit-intensive representations, whereby resources like channel capacity and storage can be used more efficiently. During the years much effort has been spent on realizing digital image compression for two-dimensional images and video sequences. The same techniques as in conventional compression can be applied independently on the left-eye and right-eye view of a stereo image pair. However, higher compression ratios can be achieved if the high correlation between both views is exploited. In this chapter we review various approaches towards stereoscopic image compression and their influence on the perceived image quality. We will discuss how coding efficiency can be improved by removing redundant image data (section 4.1), various strategies of stereoscopic image compression (section 4.2), and the effect of coding distortions on the perceived image quality (section 4.3).

4.1 Image data redundancy

The basic principle to achieve a more efficient representation of a stereoscopic image is to exploit redundant image information. Since stereoscopic images are usually represented as two matrices of pixel values (e.g. two images taken from slightly different viewpoints) we describe three types of image redundancy that can be distinguished in the pixel domain (i) coding redundancy, (ii) inter-pixel redundancy and (iii) psycho-visual redundancy (Wandell, 1995; Gonzalez and Woods, 1992).

Coding redundancy can be exploited by, for example, recoding the source information (e.g. gray-scale pixel values) such that the most frequently occurring source values are assigned less bits than the less frequently occurring source symbols. Examples of entropy coding are Huffman coding and arithmetic coding (Gonzalez and Woods, 1992). On average this reduces the number of bits needed to describe an image.

Inter-pixel redundancy is removed by making use of the correlation between pixel values. This correlation can be (i) between adjacent pixel values within a frame (intra-frame redundancy), (ii) between pixel values in successive frames (inter-frame redundancy) and (iii) between pixel values in the left and right view (inter-view redundancy). For instance, within a frame the gray-scale values change gradually from one pixel to the other. Therefore the value of any pixel can be predicted from the values of its neighbors. Especially in stereoscopic video sequences the correlation of pixel values

in succeeding frames or multiple-views can be exploited by predictive coding (e.g. lossy predictive coding by differential pulse code modulation, DPCM) to achieve compression.

Psycho-visual redundant data is image information that can be discarded without being noticeable. Thus the recovered coded image can be perceptually the same as the original image although the physical signals are different (Gonzalez and Woods, 1992). Psycho-visual studies showed that the human visual system does not respond with the same sensitivity to all visual information. For instance, spatio-temporal contrast sensitivity and contrast masking are properties of the human visual system that can be used to remove perceptually redundant data in monoscopic and stereoscopic images (Winkler, 1999). Additionally, typical properties of binocular vision, e.g. binocular rivalry (see Chapter 2) and disparity sensitivity, can be used to exploit psycho-visual redundant data in stereo images (Howard and Rogers, 1995; Bradshaw and Rogers, 1999).

The properties of the human visual system can be incorporated in compression schemes most easily with transform coding techniques such as the discrete cosine transform or the wavelet transform. For example, the discrete cosine transform (DCT), transforms images from the spatial or pixel domain into the frequency domain. The transform coefficients are products of cosines in two orientations at different spatial frequencies. Although the decomposition is not the same as assumed in the human visual system, the understanding is that high spatial frequencies can be quantized without losing much image quality. It is mainly the quantization process which achieves compression. Other transform coding methods, which are modelled after the properties of the human visual system, use a pyramid decomposition of the image to achieve a higher compression ratio by quantization of the error images. Another advantage of transform coding techniques, such as the DCT, is that an image can be described more efficiently, in terms of bits, by transform coefficients than by gray-scale pixel values. Transform coding is therefore applied in most monoscopic and stereoscopic compression schemes.

In general two compression approaches can be distinguished: lossless and lossy compression. Lossless compression algorithms preserve all image information and employ for instance transform coding, lossless predictive coding and entropy coding of the transform coefficients. The defining feature of lossless compression is that the compression process is error-free and reversible, thus the original signal can be recovered. Through lossy compression, on the other hand, some image information is lost. Lossy compression schemes incorporate for example transform coding, DPCM, quantization and entropy coding of the transform coefficients. In this case the original signal can not be recovered and lossy coding is thus a non-reversible process.

Two forms of lossy compression can be distinguished: perceptually lossless coding, where the coded image is perceptually indistinguishable from the original, and perceptually lossy coding, where the coded image visibly differs from the original. In the former case, the image represented with the smallest number of bits is assumed to contain only the information that a human perceives (Watson, 1987).

In present-day applications, removing also non-redundant information seems unavoidable for maximum utilization of the bandwidth. The effects of perceptually lossy compression of broadcast programs for 2-D TV and in particular 3-D TV are often badly understood (Falkus, 1996). A definition of the relationship between the physical coding parameters and the perceived image quality would thus be a valuable contribution to guarantee also for compressed television programs a good picture quality.

4.2 3-D compression schemes

In this section we discuss two stereoscopic compression approaches: (i) disparity and depth based coding (section 4.2.1), and (ii) mixed resolution coding (section 4.2.2). In section 4.2.3 multiview compression is discussed.

4.2.1 Disparity and depth based coding

Most 3-D compression schemes are developed for stereoscopic pictures or video that consist of two views taken from a slightly different perspective of a three-dimensional scene. The image information in both perspective views correlates highly. The principle of compression schemes utilizing disparity estimation is to exploit this correlation between views by predicting one view (target) from the other view (reference). Hence, the reference view is in general coded with a traditional 2-D compression method whereas the target view can be represented by disparity vectors.

Disparity estimation in coding schemes is analog to the correspondence problem in theories of stereopsis (see Chapter 2), namely to cross-correlate corresponding image points between the left and right-eye view. Two basic categories of stereoscopic coding schemes using disparity estimation can be distinguished: intensity based methods and feature based methods. The former determine the disparity on the basis of corresponding image intensities while the latter methods use image features such as edges or objects (Strintzis and Malassiotis, 1998; Aydinoglu and Hayes, 1998; Naemura et al., 1999; Jiang and Edirisinghe, 2002).

Intensity based methods and in particular the block-based disparity estimation approach are mostly used in stereoscopic compression schemes. The principle of block-based disparity estimation is to divide one of the views (target) into non-overlapping blocks with a fixed size of $N \times N$ pixels (e.g. 8×8 pixels). For each $N \times N$ pixel block the most similar block in the other view (reference) is determined by means of an error criterion that expresses the best corresponding match. The accuracy of this best corresponding match depends on the used statistical measure, expressing the matching error of a $N \times N$ pixel block. For instance with the mean squared error (MSE) better matches can be deduced than with the mean absolute error (MAE) (An et al., 2001). The most similar block is the disparity compensated prediction and the displacement is the blocks disparity. Though most algorithms use non-overlapping blocks with a fixed size of $N \times N$ pixels, coding improvement was reported by introducing variable block sizes (Seferidis and Papadimitriou, 1993; Sethuraman et al., 1995; Aydinoglu and Hayes, 1996). Xu et al. (2002) propose unequally sized overlapping blocks to avoid discontinuity in estimated depth at block-borders and to obtain a smoother disparity field.

Most disparity estimation techniques are applied on stereoscopic image pairs captured by a parallel camera configuration. This facilitates the disparity estimation process since only displacements in the horizontal direction need to be estimated. A disparity estimation approach that incorporates the perspective distortions is suggested by Seferidis and Papadimitriou (1993). Apart from that, it is often assumed that both cameras are calibrated identically. Thus the intensity of pixel values in both views are assumed to be comparable. Possible intensity differences between the two views are taken into consideration in the disparity-compensated transform-domain predictive coding (DCTDP) approach (Perkins, 1992).

One of the drawbacks of block-based disparity estimation is that occluded regions can not be represented by the reference image and disparity vectors alone. No true matching block can be identified for areas (e.g. objects or image borders) in the target image that are occluded in the reference image. In order to avoid unreliable disparity estimates for occluded regions a similarity threshold can be used. If the blocks do not match according to the similarity threshold, the pixel block of the target image is directly used instead of its disparity estimation (Dinstein et al., 1988). Other effective approaches to handle occlusion are for instance based on quadtree decomposition (Seferidis and Papadimitriou, 1993; Sethuraman et al., 1995; Aydinoglu and Hayes, 1996). Better matches can be deduced by decreasing the block-size if the matching error is too large. The most straight forward way to avoid occlusion problems is to code the reference image together with the disparity vector and a residual image. The residual error image is the obtained difference between the original target and the recovered disparity compensated target image (Yamaguchi et al., 1989). In this case, coding improvement can be gained by quantization of the residual image.

In block-based disparity estimation it is assumed that the disparity is uniformly distributed in a $N \times N$ pixel block. However for natural scenes this is not necessarily true. Moreover, the pixel block in the target image that is matched to a pixel block in the reference image does not necessarily represent matching image points and is thus not capable to reflect the true disparity. These issues are avoided in object-based disparity estimation algorithms. In this case, features (e.g. edges or objects) in the target image are matched to similar features in the reference image in order to deduce the disparity vector. However, complex analyses are needed to identify the objects in a scene and estimate the disparity vector. Since object-based techniques introduce less coding errors in the reconstructed image they are particularly suited to achieve compression at low bit-rates (Aizawa and Huang, 1995; Strintzis and Malassiotis, 1998). A hybrid method to estimate the disparity at an object-based level but code the image at a block-based level is suggested by Jiang and Edirisinghe (2002).

The above described disparity estimation methods are deduced from pictures or video captured with passive range finding techniques. If on the other hand active range finding techniques are used (see section 3.3), a dense depth map is provided instantly. Depth maps can be used to identify regions of interest (ROI) that can be coded at suitable resolutions to achieve the best image quality for a given bit-rate (Grau et al., 2001). This makes it possible that important regions can be coded at a higher bit-rate than less important image regions. One of the drawbacks of active depth range finding techniques is that only a single view is captured. The second view is reconstructed from the captured image and depth information. However, occluded areas can not be recovered from this information alone (Op de Beeck and Redert, 2001).

Both passive and active range-finding techniques often use traditional 2-D compression methods to code the reference view independently. Block-based DCT transform coding such as used in the standards JPEG and MPEG-2 are often applied (Dinstein et al., 1988; Perkins, 1992; Ziegler et al., 1998). Woo and Ortege (1999) describe a stereoscopic compression scheme for blockwise dependent bit-allocation that outperforms regular block-based coding schemes. The quantization scale can differ for each 8×8 block of DCT-coefficients whereas in e.g. baseline JPEG-coding the same quantization scale is applied on each block. However, comparable or better compression with less annoying artifacts can be achieved by wavelet based coding (Xu et al., 2002). For most disparity estimation based coding methods the reconstruction quality of the target image depends also on the image quality of the reference image. Thus, if the reference image is badly degraded by the 2-D compression technique the distortions are also introduced in the disparity compensated target image.

Compression schemes using disparity or depth information (depth-annotated images) can be quite easily incorporated into MPEG-2 broadcast standards. Stereoscopic coding tools compatible with MPEG-2 were investigated in the European project DISTIMA (Ziegler et al., 1998). Two approaches were investigated: (i) separate MPEG-2 coding of each view, and (ii) MPEG-2 based coding with integrated disparity estimation. The latter approach, resulted in a disparity estimation tool adapted from the MPEG-2 motion compensation technique. The left view is coded with the traditional (i) Intra (I) pictures, (ii) Predictive (P) pictures and (iii) Bidirectional (B) pictures. The right view consists of pictures containing only disparity compensation and pictures with both disparity and forward motion compensation. The right view is only predicted from the I- and P-pictures in the left view. The image quality of stereoscopic MPEG-2 coded video sequences can be improved by using a perceptual adaptive quantization technique (Tseng and Anastassiou, 1995a). The quantization parameters are determined by predictors based on binocular masking properties. The quantization scale depends on (i) predictions that are visibly intolerant, (ii) visibility of blockiness, (iii) occluded regions, and (iv) spatial activity.

Within the ATTEST project, an MPEG coding syntax that can handle depth-annotated images will be explored. Such a data format should be flexible enough to support different display types as well as interactivity with the user (e.g. adaptable depth). The goal is to achieve a video data format that is compatible with traditional coding standards (MPEG-2/4/7) and 2-D TV-sets as well as suited for novel 3-D TV applications. It is proposed that the stereoscopic video format consists of a regular 2-D sequence with synchronized depth information as an additional enhancement layer. The data-format should be flexible enough such that also depth cues like occlusion and motion parallax can be incorporated (Fehn et al., 2002).

4.2.2 Mixed resolution coding

A different class of stereoscopic coding schemes are based on theories of binocular suppression (see Chapter 2. The assumption that the final percept is dominated by the high quality component of a stereo pair is exploited to achieve compression. Thus, when one view of the stereo pair is of high image quality the other view can be degraded without introducing visible image degradations in the binocular percept. This approach is in particular interesting for appreciation oriented applications, such as television programmes, where the appreciation of depth is important. Mixed quality coding is probably less suited for performance oriented application requiring accurate detailed depth perception (e.g. like remote-controlled navigation).

Perkins (1992) introduced the mixed resolution coding concept. One view of the stereo image pair is represented as a high-resolution image while the other view is reproduced at a low image resolution. Compression is achieved by subsampling the image at the coding stage and thereby reducing the high spatial frequencies. For instance, in comparison with the high-resolution view only 6% more bit-rate is needed to represent the low-resolution view if a subsample factor of 4 is used. Subsampling with a 4x4 low-pass filter at the coding stage in combination with a bilinear interpolation filter at the reconstruction stage in the decoder resulted in the best image quality. Ideal low-pass filtering or Hamming window filtering in combination with sinc function interpolation at the decoding stage resulted in less pleasing images.

Reynolds and Kenyon (1996) uses a wavelet transform of the left and right-eye image to obtain two independent multi-resolution images. One view represents a detailed high-resolution image while

the other view is a lower resolution image. The low-resolution image is obtained by selecting the low spatial frequencies which give a coarse description of the disparity information. Additionally, the low-resolution view contains (i) the vertical edges at low spatial frequencies, (ii) horizontal edges at high spatial frequencies, and (iii) intermediate frequencies. This additional image information reduces the suppressed information in the low-resolution image and thereby enhances the depth percept.

Mixed resolution coding assumes that if one view is a high quality image the other view can be degraded without affecting the binocular percept. However, the perceived stereoscopic image quality of such an asymmetric coded stereo pair depends on the impact of coding distortions with different appearances (e.g. blockiness and blurring). In section 4.3.1 we discuss the binocular combination rule of different coding distortions in asymmetric coded images.

4.2.3 Multiview coding

Different applications can be distinguished for stereoscopic multiview pictures or sequences. The first application is a multiviewer scenario where several viewers look at the same display from different viewing angles. A correct yet different stereoscopic perspective of the same scene is provided to each viewer. In the second application, motion parallax is supported. Head movements of a single observer are tracked and the presented view-point will change accordingly. Additionally, except for stereoscopic display systems conventional 2-D TV can also support multiview applications. On a traditional 2-D TV viewers can experience depth through motion parallax. The goal of the ATTEST project is to define a coding syntax that supports different kinds of viewing conditions (Op de Beeck et al., 2002).

The properties (e.g. type and number of views) of a multiview stereoscopic picture depends on the viewing condition. If several viewers look at the same scene from a different viewpoint, a predefined number of stereoscopic views needs to be transmitted, one for each viewer. Whereas, if motion parallax is supported for a single viewer, the views depend on the head movements of the viewer and a few hundred views may be needed to support head movements in a range of 50 cm (Aydinoglu and Hayes, 1994).

The large amount of data inherent to multiview systems can be reduced by exploiting the inter-view redundancy. A major reduction of the data can be obtained by compressing the original multiviews into a smaller number of key views with sufficient disparity information. At the receiver side, intermediate views can be reconstructed from these sparse number of key views and the disparity information (Naemura et al., 1999).

The image quality of intermediate views depends among other things on whether occluded areas can be recovered from (i) the coded key views, and (ii) the density of the disparity values. In Tseng and Anastassiou (1995b), the occlusion problem is tackled by coding the extreme left and right images as well as a third image representing all objects that are visible in intermediate frames but not in the extreme left and right image. In combination with dense disparity maps, intermediate views can be reconstructed. In Siegel et al. (1997) a method is proposed to select key views with minimal occlusion and an interpolation method to generate intermediate views from a sparse disparity map. Different techniques to obtain a sparse depth or density map and interpolation methods to reconstruct an intermediate view are described in Tzovaras et al. (1998) and Chang and Lie (2000).

4.3 Evaluation of compression schemes

The performance of compression schemes is mainly dictated by two dependent factors, the achieved bit reduction and the perceived image quality. The image quality is often evaluated by means of objective fidelity criteria such as the peak-signal-to-noise ratio (PSNR) or the root-mean-squared error (RMSE). Although, these measures give an indication of the difference between the original and the coded images, the results do not necessarily correspond to the perceived image quality as obtained by subjective testing.

In the case of perceptually lossless compression, subjective testing is needed to verify that the coding scheme is tuned such that the reduced information is not visible. On the other hand, for perceptually lossy image compression the defining feature is that visible image degradations are permitted. In this case, subjective tests are needed to establish the relation between the introduced coding distortions (inherent to the coding approach) and the perceived image quality. Since each coding method introduces typical distortions, the differences in appearance of these distortions may cause differences in image quality appreciation.

In section 4.3.1, we discuss a set of conventional coding distortions that also occur in stereoscopic images. In section 4.3.2 artifacts specific to stereoscopic coding are discussed.

4.3.1 Conventional coding artifacts

Coding distortions introduced by traditional coders have been studied extensively. Physical characteristics and perceptual effects of coding distortions, determined by the codec's design choices and parameters (e.g. motion compensation (MC), differential pulse code modulation (DPCM) and quantization of block-transformed DCT-coefficients), are discussed in depth by Yuen and Wu (1998) and Wu et al. (1996). Traditional 2-D compression techniques are often instantly applied on stereoscopic image material or incorporated in stereoscopic compression algorithms (e.g. in combination with disparity compensation (DC)). Hence, a number of stereoscopic coding distortions can have the same physical characteristics as those introduced by monoscopic compression algorithms. However, human viewers are in general more tolerant towards coding impaired stereoscopic video than to non-stereoscopic video containing the same type and degree of image degradations (Schertz, 1992; Chen et al., 1998).

The relationship between the physical characteristics of distortions and the type and degree of the perceived impairments is relevant to engineers to be able to tune and optimize compression algorithms such that not only the achieved bit-rate can be controlled but also the impact of coding on the perceived image quality can be determined. The physical characteristics of the conventional coding artifacts blockiness, blurring, jerkiness and ghosting, and experimental results related to these perceived impairments in stereoscopic images and sequences are discussed next.

- Blockiness is caused by coarse quantization of block-transformed DCT-coefficients. The independent coding of $N \times N$ pixel blocks gives rise to different characteristics and degrees of coding errors between adjacent blocks. These different coding errors manifest as discontinuities at neighboring block boundaries. The sudden intensity changes are most conspicuous in uniform regions and primarily caused by the coarse quantization of the lower-order DCT-coefficient. In

textured areas the coarse quantization or absence of higher-order DCT-coefficients results in a smooth reconstructed $N \times N$ pixel block and thereby induce to a large extent the blocking effect. As a matter of fact, the blocking effect coincides with a collection of image distortions (e.g. false edges, motion compensation mismatch and mosquito noise). The physical characteristics of these related image degradations and their relationship with the blocking effect are discussed in depth by Yuen and Wu (1998).

- Blurring is caused by coarse quantization of the higher-order DCT-coefficients that can be related to low-pass filtering. This suppression of high spatial activity in an image results in a smooth percept that suffers from loss of spatial detail and reduced sharpness of object edges and textured areas. A moderate blurring effect can be caused by bidirectionally predicted macroblocks. The content of a bidirectional predicted macroblock is averaged by the interpolation of the backward and forward predictions. Where coding of the luminance information may lead to blurring, coarse quantized higher-order DCT-coefficients derived from the chrominance domain result in color bleeding. This is perceived as color smearing at the borders of regions with contrasting chrominance values (Yuen and Wu, 1998).
- Jerkiness can be caused by temporal resolution reduction such as with drop-and-repeat frames (Stelmach et al., 2000b). Additionally, jerkiness can also be caused by transmission delays and thus attributed to the codecs buffer abilities (Yuen and Wu, 1998). The temporal sampling rate depends on the degree of motion in a scene and the ability of the eye to distinguish fast changing objects. If the temporal sample rate is too low then the smoothness of motion in a video sequence is disrupted and perceived as jerky motion. Temporal resolution reduction is permitted during approximately 66 msec after a scene change. Due to the reduced ability of the eye to perceive temporal details after a scene cut, the temporal resolution can be reduced for a moment without being noticeable (Haskell et al., 1997).
- Ghosting or double contouring is caused by temporal low-pass filtering such as averaging two consecutive frames of the original video. It is perceived as a blurred residue dragged along fast moving objects. The principle of temporal filtering is that the image information in successive frames is highly correlated and thus a more efficient representation can be obtained by averaging consecutive frames. However, if a scene contains rapidly moving objects the correlation between frames diminishes and the filtering produces a percept of the uncorrelated image information resulting in double contouring.

The effects of image degradations have been studied for symmetric and asymmetric processed stereo images and sequences. Symmetric processing implies that the left- and right-eye image are processed identically. Thus the same degree of distortion is introduced into both images of a stereo pair. On the other hand, a stereo pair is processed asymmetrically if the left- and right-eye image contain a different degree of distortion. For example, the left-eye image is the source while the right-eye image is low-pass filtered. The main research topics in psycho-visual studies on image distortions with regard to 3-D TV have been:

- the tolerance of observers to coding distorted stereo images in comparison to distorted non-stereo images, and

- what binocular combination rule is applied for monocular inputs containing a different level of degradation.

Human observers seem more tolerant to coding distortions in stereo than in non-stereo sequences. Schertz (1992) showed that observers prefer impaired DCT-coded stereo sequences over the monoscopic originals, even though the perceived impairment was rated as perceptible and slightly annoying. The experiments of Chen et al. (1998) support these results for JPEG-coded stereo and monoscopic images. The authors showed that human observers experience an impaired JPEG-coded stereo pair as containing less (annoying) impairment than the equally impaired JPEG-coded monoscopic image.

Sharpness ratings of spatial low-pass filtered stereo and non-stereo sequences gave the same indication. Observers seem more tolerant to blurred stereo than blurred monoscopic sequences. Symmetrically low-pass filtered stereo sequences (the same level of spatial low-pass filtering is applied to the left- and the right-eye image) are perceived as sharper than equally processed non-stereo sequences (Berthold, 1997). On the other hand, Tam et al. (1998) showed that symmetrically MPEG-2 coded stereo sequences are perceived equally or less sharp than their non-stereo versions. The authors argue that this can be caused by the interaction between disparity, image resolution and image size. Stereoscopic sequences that were rated less sharp than the non-stereo ones contained excessive disparity. Hence, it may have been difficult for the observers to fuse the images or track the objects.

One of the theories in stereoscopic compression is that the binocular percept is dominated by the high image quality component. This suggests that with asymmetric compression one stereo component can be degraded while the other component is maintained at the desired image quality. Stelmach et al. (2000b) showed that the perceived image quality of an asymmetric spatial low-pass filtered stereo sequence is dominated by the high quality component. However, the perceived image quality of asymmetric MPEG-2 coded sequences is approximately the average image quality of the monocular sequences. Moreover, low-pass filtering and MPEG-2 coding seem to have an independent effect on the perceived image quality. This was demonstrated for stereo sequences where one component is the original sequence and the other component is spatial low-pass filtered and MPEG-2 coded (Stelmach et al., 2000a).

Blur and blockiness are typically induced by spatial low-pass filtering and MPEG-2 coding, respectively. The binocular combination of these impairments was studied by Meegan et al. (2001). If one component of the stereo pair was blurred while the other component was the source image, the binocular sharpness percept was dominated by the high quality component (source). Thus the source image is assigned a higher weight in the binocular combination than the degraded component. This implies that when one image is compressed by removing high spatial frequencies and the other image remains of good image quality the final binocular percept is hardly affected. Different results were found for the weighting of blockiness artifacts. In this case the degree of blockiness perceived in the binocular percept is approximately the average of the perceived blockiness in the two monoscopic components. Thus in this case the high quality image (source) does not dominate the final percept. It can be concluded that the type of coding artifact and thereby the type of introduced impairment are processed differently by the human visual system.

The temporal coding distortions ghosting and jerkiness, induced by the averaging of two frames and drop-and-repeat frames, respectively, have the same effect on stereo and on non-stereo sequences. The perceived image quality decreases with the same amount in asymmetrically processed stereo

sequences (one view is the source and the other view is temporal filtered) as in non-stereo sequences if the coding distortions increase (Stelmach et al., 2000b). Therefore it can be said that the temporal filtering is not attenuated by the high quality sequence.

Several studies showed that the human visual system is more tolerant towards coding artifacts (in particular blurring) appearing in stereoscopic images than in monoscopic images. Therefore the relationships between the bit-rate of an image and the perceived degree of impairments as determined for traditional coding artifacts are not always applicable on the stereoscopic percept. Especially, the binocular combination of coding artifacts is of interest and can be used to improve the coding gain.

4.3.2 Artifacts specific to stereoscopic coding

First, we review psycho-visual studies that were carried out to investigate the effect of conventional coding artifacts in stereoscopic images on the perceived sensation of depth and the relative depth. Next, we review results on a conspicuous typical depth distortion (cardboard effect) resulting from coarse quantization of the disparity values.

Several studies demonstrated that the effect of conventional coding distortions on the sensation of depth in stereoscopic images seems negligible. In Chen et al. (1998) the left and the right view images of a stereo pair were independently JPEG coded. It was shown that even though the introduced JPEG-coding distortions were perceptible and ranged from slightly annoying to annoying, for a compression ratio between 1 and 100, the sensation of depth was hardly affected. In Tam et al. (1998), MPEG-2 coded monoscopic and stereoscopic sequences were used to assess the perceived sensation of depth. In the stereoscopic case both views were independently MPEG-2 coded. The same decrease in perceived sensation of depth was found for the MPEG-2 coded monoscopic sequences (consisting of only monoscopic depth cues such as size, occlusion and perspective) and the MPEG-2 coded stereoscopic sequences (with additional binocular parallax). Also for spatial and temporal filtering the depth percept was hardly affected by the introduced coding distortions such as blur, jerkiness and double contouring (Stelmach et al., 2000b).

The effect of coding distortions, induced by conventional compression in combination with disparity estimation, on the perceived relative depth seems to depend on the experimental paradigm or the applied compression method. In Dinstein et al. (1988) observer's response time and accuracy measurements showed that the perceived relative depth is not significantly different in compressed and uncompressed images. Although the measured reaction times indicate that the perceptual processing for compressed images is slightly slower than for uncompressed images. On the other hand, Perkins (1992) showed, by means of a matching paradigm, that objects are perceived further away in compressed images than in uncompressed images. The author argues that the loss of fine spatial detail causes blurred edges and thereby causes ambiguity in edge locations. Consequently, observers perceive a scene with lack of spatial detail and ambiguous disparity information which is in the brain associated with more distant objects (i.e. atmospheric perspective cues).

A typical stereoscopic coding distortion affecting the perceived depth is the cardboard effect. A cardboard effect can be caused by coarse quantization of disparity or depth values. This results in a depth percept whereby the stereoscopic scene is divided into planes at different distances and as a consequence objects appear flat and projected onto the different depth planes. The effect can be compared to the scenery in a theatre (Schertz, 1992). The discontinuous depth and hence the

flattening of objects in a scene evokes an unnatural depth percept. The same spatial depth distortion can also be produced by shooting conditions of a three-dimensional scene (see Chapter 6). Moreover, quantization of disparity values can also tear up an object such that it is represented in several depth planes and perceived as a disjoined object. Temporal discontinuous depth mismatches can occur if an object or parts of an object are assigned to different depth layers in time and result in a flickering depth percept.

The ability of humans to perceive depth decreases exponentially with increasing distance from the convergence plane. Therefore, Schertz (1992) suggests to vary the disparity resolution by means of an exponential quantization function. The cardboard effect, caused by coarse quantization, can be reduced by low-pass filtering of the disparity values at the receiver side. By means of subjective tests the author showed that no impairments were perceived if approximately 24 disparity values were used. Depending on the scene content, perceptible but not annoying impairment ratings were obtained when the number of disparity values were reduced to eight or ten. Moreover, if objects were torn apart by the quantization of their disparity values the impairments in a scene were judged as more annoying than if the objects are reunited by reducing the depth resolution.

Chapter 5

Stereoscopic display techniques

5.1 Introduction

Stereoscopic display techniques are based on the principle of taking two images and displaying them in such a way that the left view is only seen by the left eye, and the right view only seen by the right eye. There are a number of ways of achieving this (Pastoor and Wöpking, 1997; Sexton and Surman, 1999). Stereoscopic displays can be categorized based on the technique used to channel the right and left images to the appropriate eyes (see Table 5.1). A distinguishing feature in this regard is whether the display method requires a viewing aid (e.g. glasses) to separate the right and left eye images. Stereoscopic displays that do not require such a viewing aid are known as *autostereoscopic* displays. They have the eye-addressing techniques completely integrated in the display itself. Other distinguishing features are whether the display is suitable for more than one viewer (i.e. allows for more than one geometrically correct viewpoint), and whether look-around capabilities are supported, a feature inherent to a number of autostereoscopic displays (e.g. holographic or volumetric displays), but which requires some kind of head-tracking when implemented in most other stereoscopic and autostereoscopic displays.

5.2 Stereoscopic displays

Stereoscopic displays (i.e. those using viewing aids) can be *time-parallel*, with both left and right eye views appearing simultaneously, or *time-multiplexed*, where the left and right eye views are shown in rapid alternation and synchronized with a liquid crystal (LC) shuttering system which opens in turn for one eye, while occluding the other eye. Time-multiplexed systems make use of the fact that the human visual system is capable of integrating the constituents of a stereo pair across a time-lag of up to 50 ms (Pastoor and Wöpking, 1997). Because of the rapid alternation of right and left eye images, the display will need to run at twice the frame rate of the original sequence material. For example, to display a 60 Hz sequence a 120 Hz display is required. When image extinction is not fast enough, due to the persistence of CRT phosphors, *cross-talk* may occur. This is an imperfect separation of the left and right eye views which can be perceptually very annoying. An example of a time-multiplexed system is the StereoGraphics ChrystalEyes system, which has become popular for

Table 5.1: Classification of stereoscopic display techniques (adapted from Pastoor & Wöpking (1997), page 101)

<i>Principle of eye addressing</i>		<i>Number of different views</i>	<i>Eye-point dependent perspective</i>
Aided viewing (stereoscopic)	Color multiplex, polarization multiplex, time multiplex, location multiplex	Two	Optional (for a single observer)
Free viewing (autostereoscopic)	Direction multiplex (e.g. by refraction, or occlusion)	\geq Two	Optional (for multiple observers)
	Volumetric displays, electro-holography	Unlimited	Inherent (for multiple observers)

use with stereoscopic desktop workstations, and is also frequently used for large projection spaces, such as the CAVE (Cruz-Neira et al., 1993).

For time-parallel stereoscopic displays several multiplexing methods have been used, based on either color, polarization or location. In color-multiplexed, or *anaglyph*, displays the left and right eye images are filtered with near-complementary colors (red and green for Europe, red and blue for the USA). The observer is required to wear color-filter glasses to separate the images. This well-known and inexpensive method has been used for stereoscopic cinema and television, and is still popular for viewing stereoscopic images in print (magazines, etc.), since the approach readily lends itself to the production of hard copies. A serious limitation of this method is that color information is lost since it is used as a selection mechanism. Only limited color rendition is possible through the mechanism of binocular color mixture.

Polarization-multiplexed displays separate left and right eye images by means of polarized light. Left and right output channels (monitors or projectors) are covered by orthogonally oriented filters, using either linear or circular polarization. The observer needs to wear polarized glasses to separate the different views again. This system offers good quality stereoscopic imagery, with full color rendition at full resolution, and very little cross-talk in the stereo pairs¹ (Pastoor and Wöpking, 1997). It is the system most commonly used in stereoscopic cinemas today. The most significant drawback of this kind of system is the loss of light output from using the polarizing filters. In addition, with linearly polarized systems, some amount of ghosting or cross-talk may occur when the viewer's head is not in the correct position, e.g. by tilting it. This is not a problem with circular polarization, however at the expense of an even higher loss of light intensity.

Perhaps the best known and certainly the oldest method for displaying stereoscopic images is through location multiplexing. This means that the two views are generated separately and are subsequently relayed to the appropriate eyes through separate channels. The Wheatstone and Brewster stereoscopes, discussed earlier, are historical examples of this type of display. The principle of

¹This is usually less than 0.1 % with linear filters.

location multiplexing is today being used in stereoscopic head-mounted or BOOM displays, as well as in their popular low-tech sibling, the View-Master stereoscopic viewer.

An interesting stereoscopic effect, based on artificially induced disparity, is known as the *Pulfrich effect*. Based on observations by the German astronomer Carl Pulfrich in the 1920s, the effect can be observed with a simple pendulum swinging to and fro in the frontal plane. When one eye is covered with a darkened filter (such as a sunglass) the perceived trajectory will describe an ellipse, rather than the actual linear movement. As an explanation Pulfrich suggested that the response time of the eye becomes slower at lower light intensities, much like the shutter speed of a camera in low light. Therefore, when observing an object in (linear) motion, the uncovered eye registers the object in one position, whereas the eye covered with the filter "sees" it where it was previously, thus providing illusory disparity information.

This effect can be used as a very cheap and simple means to obtain a limited form of 3-D TV, with the added advantage that the image looks totally normal when viewed without glasses. Any image that displays a constant lateral movement, either natural movement of objects in the image or movement induced by lateral camera motion, can display the Pulfrich effect. However, the fact that lateral motion is required to generate the stereoscopic illusion also signals the main limitation of this technique. Stationary shots do not exhibit Pulfrich depth. Moreover, when objects suddenly start or stop moving in an image, or move in a non-constant fashion, this will affect the depth percept in a noticeable and unnatural way. In fact, a correct stereo effect will only occur when the amount of constant lateral movement during a time interval equal to the difference in integration time between the two eyes is approximately the same as the inter-ocular distance of the viewer's eyes. Given these limitations the Pulfrich effect is clearly unacceptable and impractical for serious TV applications².

5.3 Autostereoscopic displays

Autostereoscopic displays (i.e. those not requiring any viewing aids) can be *direction multiplexed*, whereby different two-dimensional images are projected across the viewing field, *holographic*, where the image is formed by wavefront reconstruction, or *volumetric*, where the image is formed within a physical volume of space. In addition, free-viewing stereoscopic images by fusing them without any viewing aid, either parallel or cross-eyed, can be regarded as another form of autostereoscopic display.

Direction-multiplexed displays apply optical principles such as diffraction, refraction, reflection and occlusion to direct the light from the different perspective views to the appropriate eye (Pastoor and Wöpking, 1997). Historically, the two most dominant autostereoscopic techniques are based on *parallax barriers* and *lenticular arrays*, and they are still popular today (Sexton and Surman, 1999).

Parallax barrier displays are based on the principle of occlusion, i.e. part of the image is hidden from one eye but visible to the other eye. The oldest-known display of this type was proposed in the early 20th century and consisted of a fine vertical slit plate or grid superimposed upon a photograph which consisted of vertical columns or stripes, alternating right and left eye views. At the right

²Provided that there is constant lateral motion in the image, two temporally segregated frames from a monoscopic film or video can be used to calculate a spatial disparity map of the original scene. This provides a limited way to convert monoscopic video to quasi-stereoscopic video.

viewing distance and angle, one eye can only see the appropriate view, as the other view is occluded by the barrier effect of the vertical slits. Different implementations of this principle are available, including parallax illumination displays (where the opaque barriers are placed behind the image screen) and moving slit displays (using time-sequential instead of stationary slits).

Lenticular systems are based on the principle of refraction. Instead of using a vertical grating as with parallax barrier displays, an array (or sheet) of vertically oriented cylindrical lenses is placed in front of columns of pixels alternately representing parts of the left and right eye view. Through refraction, the light of each image point is emitted in a specific direction in the horizontal plane. This technique is well-known from 3-D picture postcards and photographic prints, characterized by a convoluted plastic sheet on the surface of the image. Both lenticular and parallax barrier systems require the precise alignment of the picture splitter (vertical stripes or lenses) with the vertical left-right image strips. This requires displays with a very stable position of the picture elements, thus favoring LCDs or plasma displays over CRT displays.

An important drawback of many direction-multiplexed autostereoscopic systems (including lenticular sheet or parallax barrier systems) is the fact that only under a limited horizontal viewing angle the picture will be perceived correctly. A lenticular sheet system developed at Philips Research Labs in the U.K. addresses this issue by presenting a series of discrete views across the viewing field. Interestingly, their multiview display uses a slanted lenticular sheet to eliminate the picket fence effect (i.e. the appearance of vertical banding in the image due to the black masks between columns of pixels in the LCD) and reduce image flipping (i.e. the noticeable transition between two viewing zones). The slanted lenticular hides the black mask image by spreading it out and softens the transition between one view and the next (van Berkel and Clarke, 1997). Each view in a multiview system is displayed at the expense of image resolution, thus restricting the number of views that can be displayed at an acceptable image quality. The arrival of high resolution flat panel displays have made multiview systems become more feasible however (Sexton and Surman, 1999).

Volumetric displays tend not to rely upon flat displays, but generate an image that occupies a limited volume in space. A three-dimensional image is divided into a very large number of planes (which is why these displays are also referred to as *multiplanar*) and image points of these planes can be plotted in rapid succession onto a specially shaped rotating mirror. An example of such a system is the Texas Instruments "Omniview" system, which uses three laser beams (red, green, and blue) projected onto a rotating helix mirror.

Alternatively, a multiplanar image can be displayed using the "varifocal mirror" technique, which requires a vibrating mirror driven by an acoustical woofer. The varifocal mirror is pointed at a CRT which presents different depth layers time-sequentially. As rendition of the depth layers is synchronized with changes in the focal length of the mirror, different slices of a 3-D volume are created in planes of varying distance. Volumetric techniques allow for multiple viewpoints. Displayed objects tend to look transparent however, as if made in glass, which limits the use of such displays for images with high information content. Applications have mainly been limited to fields where the objects of interest can be represented by fairly simple shapes, such as wireframes.

Electro-holographic displays have sometimes been presented as the ideal free-viewing 3-D technique, and have received considerable attention over the past years from research groups in Japan, the United States and the United Kingdom. For most practical purposes, an important drawback is the fact that coherent light is required during image recording, which means that holograms

cannot be recorded with natural (incoherent) lighting. Also, the amount of data contained in a hologram is enormous³, requiring specific data compression techniques. At present, only very small and monochromatic displays are feasible for video-based holography (Pastoor and Wöpking, 1997).

For television viewing, Smith and Dumbreck (1988) deem holography as unsuitable, not just on grounds of technical feasibility, but also from the point of view of artistic direction. They argue that exploration of a 3-D scene should be carried out by the director at the time of production, and not by the individual viewer inspecting a hologram from different angles. Although it is certainly true that holographic filmmaking would require a different production grammar from traditional, or even 3-D, filmmaking, Smith and Dumbreck (1988) seem to overlook an important benefit of holography. Since holograms allow inspection from multiple viewpoints, they will also allow multiple viewers to enjoy the imagery at the same time, a feature that is clearly required for any successful 3-D television system, as will be discussed in the next section.

5.4 Preliminary requirements of a 3-D TV system

Although stereoscopic displays are widely being used for professional applications and in cinema, their application to home entertainment, and in particular TV, has lagged behind. As was argued in Chapter 1, the widespread introduction and acceptance of digital broadcasting makes the transmission of a stereoscopic signal increasingly feasible. Proponents of 3-D TV have argued that 3-D TV will bring the viewer a wholly new experience, a "fundamental change in the character of the image, not just an enhancement of quality" (Smith and Dumbreck (1988), p.10).

It is a widely held belief that 3-D television should be autostereoscopic. The main reason for this is that the need to wear glasses is unacceptable in a home situation where television is usually watched casually, i.e. with many interruptions for telephone calls, conversations, making a sandwich, or engaging in other activities with TV as a simultaneous background activity. Having to take a pair of spectacles on and off constantly is a nuisance.

Most current autostereoscopic displays, on the other hand, tend to restrict the viewer to a fairly rigid viewing position, in terms of the angle under which the stereoscopic imagery can be perceived without geometrical distortions or substantial artifacts (e.g. cross-talk or picket-fence effects). Stereoscopic TV should be able to provide good quality stereo pictures to multiple viewers who are free to move throughout the room. Current developments in multiview autostereoscopic displays (see e.g. van Berkel and Clarke (1997), Travis (1990)) provide hope that such a system may be feasible in the not too distant future.

In addition to the requirements mentioned above, any stereoscopic system should also be able to display monoscopic images without problem. Other important considerations include cost, picture quality⁴, and size⁵ of the stereoscopic television system. Subjective quality requirements of stereoscopic systems will be discussed in detail in the next chapter.

³The source rate is estimated to exceed 10^{12} bit/sec (Pastoor and Wöpking, 1997)

⁴Stereoscopic (in particular multiview) displays sacrifice the display's spatial or temporal resolution in order to display the appropriate views.

⁵Many of the systems proposed today require a housing size that is not acceptable for domestic settings.

Chapter 6

Human factors

There is much evidence that 3-D TV will have the potential to establish a future market in the field of digital broadcasting and display developments. Subjective tests have demonstrated the more appealing impact of 3-D material compared to conventional 2-D display techniques. Many studies are available that have demonstrated the relative added value of stereoscopic over monoscopic TV images, in terms of overall quality (IJsselstein et al., 2000b; Chen et al., 1998; Schertz, 1992; Stelmach and Tam, 1998; Berthold, 1997). Additionally, the new display techniques enable compelling reproductions of reality giving the viewers a subjective sensation of 'being there'. Freeman and Avons (2000) stated that conventional assessment criteria (such as image quality) cannot fully describe the impact on the viewer of these new types of displaying techniques. The sensation of presence during the presentation of stereoscopic video material and monoscopic material was studied by means of a focus group. Subjects watching 3-D TV reported a higher sensation of 'being there' than during monoscopic presentations of the same sequence.

In this chapter we will focus on human factors research and its contribution to specifying appreciation criteria of 3-D image systems. In section 6.1 we describe subjective assessment methods of stereoscopic television. In section 6.2 we will discuss conventional subjective attributes, such as image quality, sharpness, naturalness, but we will also cover attributes that are specifically related to 3-D TV evaluation (eye strain, depth and presence). In section 6.3, we review image quality modelling approaches for conventional 2-D images and ATTEST's quality model for stereoscopic image systems will be introduced.

6.1 Subjective assessment methods

Various methods have been used for subjective assessment of monoscopic and stereoscopic television pictures. In this section we will discuss (i) an explorative method, and (ii) methods to obtain direct ratings for particular subjective attributes such as image quality, sharpness and depth.

6.1.1 Explorative study: focus groups

By means of focus groups no direct subjective ratings are acquired but instead the unbiased humans attitude, feelings and reactions towards 3-D TV are explored. Naive subjects participate in small

discussion groups to express and share their experience of the viewed stereoscopic image system. The experiment leader moderates the discussion and gives guidance if necessary. Thus, focus groups can be used to (i) collect unbiased viewers's descriptions of the sensations evoked by a stereoscopic image system, (ii) investigate the added value of new image systems, e.g. 3-D TV, 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.

In Freeman and Avons (2000) focus groups were used to explore viewers' reactions to conventional 2-D TV and novel 3-D TV. The results showed that viewers report, with respect to stereoscopic sequences, a sense of "being there" before this concept was raised by the moderator. Furthermore, this feeling of "being there" was related to attributes such as realism, naturalness, and involvement. A second aim of the focus group was to identify programme types suited for 3-D TV. In general, the subjects prefer action movies and live events such as sports, theatre and concerts. Programme types such as news, soaps, documentaries and talk shows were thought of as inappropriate for 3-D TV. Moreover, subjects indicated that they would like to decide on a program-by-program basis whether they want to watch it in 3-D or 2-D.

6.1.2 Direct ratings of subjective image quality

Several experimental paradigms can be used to measure and quantify image quality of images and sequences. Perceptual image quality is expressed as a gradation of subjective impressions of how well the image information is transmitted to an observer. The observer's criterion of good transmission of image information depends on the application. 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. Medical diagnosing for instance is facilitated by MRI or CT images. The purpose of such images is to give accurate information. Therefore if a lesion can be detected by means of a noisy MRI image the image quality satisfies the purpose.

In appreciation oriented applications, such as stereoscopic television, the goal is to generate images that are as "pleasing" as possible. The emphasis is on visual comfort associated with the images. For instance, it is strenuous to watch a television program containing excessive binocular disparities. Watching programs that induce diplopia requires a great deal of effort and viewers experience this as unpleasant.

Appreciation oriented subjective assessment of stereoscopic television pictures is described in the ITU-R BT.1438 recommendation (ITU, 2000b). The subjective assessment methods are adopted from the ITU-R BT.500-10 recommendation for conventional monoscopic television (ITU, 2000a). The proposed assessment methods are used to measure overall perceived impairment or image quality of degraded still images and image sequences. The same experimental paradigms can be applied to obtain ratings of the perceived strengths or sensation of attributes such as sharpness, depth, naturalness, or presence. In general three different approaches are proposed: the double-stimulus-continuous-quality-scale method (DSCQS), single-stimulus methods and stimulus-comparison methods.

In DSCQS observers assess the overall image quality for a series of image pairs. Each pair consists of an unimpaired image (reference) and an impaired image (test). For both images (reference and

Table 6.1: ITU-R BT.500-10 recommendation rating scales

single stimulus quality scale	DSIS and single stimulus impairment scale	comparison scale
5 excellent	5 imperceptible	-3 much worse
4 good	4 perceptible but not annoying	-2 worse
3 fair	3 slightly annoying	-1 slightly worse
2 poor	2 annoying	0 the same
1 bad	1 very annoying	1 slightly better
		2 better
		3 much better

test) observers assess the overall picture quality separately. Eventually the DSCQS assessment results are the difference in scores between the reference and test image.

In single-stimulus scaling the overall picture quality of each image in the stimulus set is assessed individually. In stimulus-comparison scaling, again a series of image pairs is used. These image pairs can include all possible combinations of two images in the stimulus set or just a sample of all possible image pairs in order to restrict the number of observations. In this procedure, observers assign a relation between the two images for each image pair.

The same single-stimulus and stimulus-comparison methods can be used to assess impairment. In the double-stimulus-impairment scale method (DSIS) again a series of image pairs (reference and test) is presented. However, the assessors are asked to judge only the test image, "keeping in mind the reference" (ITU, 2000a).

The scaling methods impose different grading scales to assess the perceived image quality. In DSCQS, a continuous graphical scale is used to avoid forcing subjects to answer within too coarse a category. The scale is often labeled with verbal terms such as *excellent*, *good*, *fair*, *poor*, and *bad*, to guide the observer. For single-stimulus, stimulus-comparison and DSIS the usually applied rating scales such as verbal or numerical categories are given in Table 6.1. The subjects express the perceived image quality, the impairment, or the relation between two images by placing the presented stimuli in one of these categories.

Average observer's quality judgements can be obtained by a number of different analysis methods. Methods such as averaging the judgements across observers by defining a confidence interval indicating the individual differences are specified by the ITU. More complex judgment models were proposed by Torgerson (1958). Such a model underlying the rating mechanisms of observers is described in Boschman (2001).

The above described subjective assessment methods, DSCQS, single-stimulus and stimulus-comparison scaling, are used to obtain a single judgement of the overall image quality of still pictures or short video sequences of 10 seconds. An alternative assessment method, single stimulus continuous quality evaluation (SSCQE), is proposed to obtain continuous time-varying quality judgements of longer video sequences. Subjects continuously assess the picture quality by moving a hand-held slider. A subject indicates excellent image quality when the slider is positioned at the top of the grading scale, and bad image quality is indicated by moving the slider to the bottom. SSCQE is

used to assess video that contains scene-dependent and time-varying impairment, for example introduced by compression. Furthermore, television is usually watched for longer periods, so SSCQE is the most appropriate way to mimic home viewing conditions. This method of assessment has been applied by IJsselsteijn et al. (1998b) to continuously assess observers' sense of presence, depth, and naturalness in the context of 3-D TV.

6.1.3 Context effects

De Ridder (2001) pointed out that "quality judgements are affected by the judgement strategies induced by the experimental procedure". Experimental conditions such as stimulus set composition, instructions and scaling technique may influence the image quality responses. Therefore, acquisition of subjective image quality data is not as straightforward as it seems. Three studies that demonstrate the context effects are described below.

The effect of scaling technique and stimulus spacing on perceived sharpness judgements was demonstrated by de Ridder (2001). Three scaling techniques: single-stimulus, double-stimulus and comparison scaling, were evaluated for a positively and negatively skewed stimulus set. Only for comparison scaling, comparable results were obtained for both sets. This suggests that for this scaling technique the judgements are hardly affected by stimulus spacing.

The issue of influence of scaling procedure on subjective judgements was also addressed by van Dijk and Martens (1996). Single-stimulus and comparison scaling lead to different results for subjective quality evaluation between different codecs. They argued that typical distortions introduced by the different codecs can be easily identified and, consequently, subjects are inclined to use separate rating scales for each coder in single-stimulus scaling. In order to link these subjective scales, an explicit comparison between images from the different coders is required.

In Meesters (2002) it was shown that subjective quality judgements across scenes and processing methods can be biased by the imposed experimental procedure. The effect of stimulus presentation on the assessed image quality was investigated for a condition in which different classes (e.g. impairment types or scenes) could be identified in a stimulus set. The question was whether human observers link quality judgements across identifiable classes (e.g. different scenes) if they are not forced to compare these classes explicitly. The results showed that in stimulus-comparison scaling subjects seemed to use separate quality rating scales if images with different scene content were not compared explicitly. This was tested for a stimulus set containing wavelet-coded images and a stimulus set consisting of JPEG-coded images. Two experiments were also conducted with stimulus sets containing different impairment types introduced by wavelet-coding, DCTune-coding, JPEG and low-pass filtering. In each experiment a single scene was impaired by each of these four processing methods. For one scene the observers seemed to judge the perceived image quality on a single rating scale whereas for the second scene these findings could not be substantiated. This is most probably due to the fact that in the second scene, the quality-ranges of the distortions were not dissimilar enough. In general, it was concluded that observers use separate quality scales for identifiable classes of stimuli if these are not compared explicitly.

Heynderickx et al. (2002) investigated the influence of the way stimuli are presented in a double stimulus subjective test. The stimuli in a double-stimulus experiment can be presented time-sequentially, i.e. first the reference stimulus followed by the test stimulus on the same display, or

simultaneously on two neighbouring displays or on two neighbouring positions on one display. The disadvantage of a simultaneous presentation is that it is almost impossible to find two displays which are exactly the same and this is required in case of evaluating images that cover the full size of the display. Thus the ITU recommends the presentation on one display, which in most cases implies a time-sequential presentation of the stimuli. Heynderickx et al. (2002) compared both methods by determining the visibility threshold of a colour gradient across the screen. Results show that in the case of a saturation gradient there is no significant difference in visibility threshold between showing the images simultaneously or sequentially. However, this is not the case for a hue gradient. The visibility thresholds of a hue gradient are approximately 3 to 4 times higher when the images are shown sequentially instead of simultaneously. Thus, for the most critical test in a double stimulus experiment the simultaneous setup is most suitable.

Experimental methods such as just-noticeable difference (JND) measurements, matching procedures and response time have been developed to avoid context effects. Watson and Kreslake (2001) describe a visual impairment scale (with JND as unit) for visibly impaired sequences that is not liable to context effects. The authors showed that JND measurements for natural scenes correlate highly with mean opinion scores (MOS) obtained by DSCQS. Furthermore, experimental paradigms such as matching procedures and response time measurements are also not subject to context effects. In a matching procedure one stimulus is adapted until it is perceptually equal to the reference stimulus. In Meegan et al. (2001) this method was used to assess the visibility of particular coding artifacts (blur and blockiness) in monoscopic and stereoscopic images. Response time measurements are for example used by Dinstein et al. (1988). The authors compared, for a number of compressed and uncompressed stimuli, the response time subjects needed to decide whether an object is further away or not. It is assumed that if the response time is longer, the perceptual processing is more difficult. Thus, if compression interferes negatively with the depth percept, the observers' response takes longer than for uncompressed images. Nevertheless, these measures may not always be suited to measure the observers' experienced appreciation of a stimulus.

6.2 Subjective attributes of stereoscopic viewing

Perceptual attributes form the basis of the customer's quality preference or judgement, so it is of great importance to understand the customer's perception. Observers do not judge quality based on technological variables but they express a preference on the basis of what they see. In this paragraph we want to discuss the relation between subjective attributes and technical parameters important for stereoscopic images (e.g. shooting parameters, compression parameters, etc.). Relevant subjective attributes of stereoscopic images are sharpness, eye strain, depth, presence, naturalness and image quality. In the following paragraphs each attribute is discussed in relation to stereoscopic images.

6.2.1 Perceived sharpness

Sharpness is an important percept that contributes to image quality. Sharpness in stereoscopic images can be affected by several parameters e.g. camera defocus, coding, binocular disparity. In this paragraph we discuss some research related to perceived sharpness in 3-D images. One would

expect that introducing disparity information in a stereo image would sharpen the edges by reducing positional uncertainty.

Berthold (1997) and Tam et al. (1998) performed an experiment in which they compared monoscopic to stereoscopic images. In the study by Berthold (1997) the subjects reported that stereo images were perceived as sharper than non-stereo images. In the study of Tam et al. (1998) the stereo images and non-stereo images were rated equally sharp or stereo even slightly less sharp. In both studies, there is a high correlation between ratings of sharpness and image quality. This suggests that perceived image quality is highly dependent on perceived sharpness.

Stelmach et al. (2000b) investigated the effect of mixed-resolution (spatial or temporal low pass filtering) stereo video sequences on the perceived sharpness in the stereo and non-stereo condition. They concluded that spatial low pass filtering gives an acceptable sharpness. The results showed that sharpness was biased towards the image 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 where they measured the visibility of blur in asymmetric processed stereo image-pairs. The perceived sharpness of the binocular percept is dominated by the sharpest of the monocular images.

6.2.2 Depth

As discussed in Chapter 2, depth perception of displayed objects is formed by monocular cues (luminance, shading, accommodation, shadows, perspective, size, etc.) and binocular cues (vergence and retinal disparity). The use of disparity information produces a compelling sense of depth, which defines the added value of stereoscopic TV.

The perceived magnitude of depth depends upon the visual angle of the separation of the two views. Thus, the perceived depth is strongly influenced by the distance of the observer from the display, as was discussed in section 3.2.3. Gooding et al. (1991) investigated the effect of viewing distance and disparity on perceived depth. A change in viewing distance significantly influenced subjective depth. Objects viewed from greater distances reflected greater differences in depth than equivalent displayed objects viewed from smaller distances. Patterson (1997) investigated several factors that influence stereoscopic depth perception. First, he divided the stereoscopic perception of depth in two stages: (i) the establishment of binocular correspondence or disparity detection and (ii) disparity scaling for different viewing distances. Stereopsis is degraded by factors that disrupt these two stages. The perceived depth from displays with normal disparities and multiple distance cues follows the geometric calculations quite closely, but when disparities are too large, diplopia occurs causing invalid depth perception. Perceived depth increases with a larger viewing distance and with half-image separation (lateral shift between two images). This suggests that large image separations and large viewing distances may be used when large depth intervals are needed.

Motoki et al. (1995) measured the effect of image size on the sensation of depth. He concluded that perceived depth is the least responsive subjective attribute to an increase in picture size. Regardless of image size, viewers experience a strong sensation of depth from the picture with binocular parallax.

Yamanoue et al. (2000) discusses the cardboard effect that may occur as a consequence of converging stereoscopic camera set-up. His results indicate that the cardboard effect is influenced by the

binocular disparity represented by spatial thickness. The cardboard effect can be lessened by enhancing the binocular parallax. Accurate reproduction of binocular parallax (spatial thickness around 1.0) reduces not only the cardboard effect, but the puppet-theater effect as well. The puppet-theater effect will be explained in 6.2.4.

IJsselsteijn et al. (1998b) investigated the perception of depth and the naturalness of depth when viewing stereoscopic image material. As soon as binocular disparity is introduced in the 8 minutes sequence, the ratings of perceived depth increased. Naturalness of depth also increased, indicating that the presented disparity is well within optimal bounds. During the sequence, depth ratings further increased when the camera started to pan and motion parallax was provided to the subject. Motion parallax is also known to provide the visual system with a strong depth cue.

Tam et al. (1998) investigated the perceived depth for stereoscopic images and non-stereoscopic images. Viewers also rated the perceived depth of asymmetric MPEG-2 coded sequences at 6, 3, and 1 Mbits/s. The results show that the addition of disparity information to non-stereoscopic video sequences enhanced the experience of depth. The sensation of depth was higher in stereoscopic sequences at all bit rates. The greater depth perceived in the stereoscopic sequences did not depend on the visibility of MPEG-2 coding artifacts. When the bit-rate was reduced from 6 Mbits/s to 1 Mbits/s, the perceived depth dropped in equal measure for both stereo and non-stereo sequences. The visibility of coding artifacts caused an equivalent reduction in the sensation of depth for stereo and non-stereo sequences.

Stelmach et al. (2000b) investigated the effect of spatial and temporal low-pass filtering on the 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, carrying 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 result appears to be at variance with the findings of IJsselsteijn et al. (2000b), where perceived quality is largely determined by perceived depth, attenuated by 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 subjects to base their quality judgements on different image attributes than with the Stelmach et al. (2000b) study.

6.2.3 Image quality

Subjective image quality is a standard psychological criterion used to evaluate imaging systems. It is a subjective preference judgement which is widely used to compare for example coding algorithms, image processing techniques, and system configurations. Image quality is regarded as a multidimensional psychological construct (Meesters, 2002), based on several attributes. In case of stereo images no comprehensive stereoscopic image quality model had been formulated to date, yet it is likely that attributes such as depth, sharpness, colour, motion rendition, flicker, and eye strain contribute to the perceived stereoscopic image quality.

Yano and Yuyama (1991) investigated the perceived image quality of stereoscopic images compared with non-stereoscopic images. Both for still and moving pictures, the stereo images were rated

higher in image quality than the non-stereo images. Berthold (1997) performed the same experiment but with picture of lower resolution. Also in this case, subjective image quality was rated higher for stereo than for non-stereo image sequences at all levels of blur.

The storage and transmission of 3-D material involves a large amount of data. Therefore it is necessary to apply several coding techniques such as JPEG or MPEG coding to reduce the used bandwidth. Chen et al. (1998) investigated the effect of JPEG coding on the subjective image quality of stereoscopic images. The results show that a JPEG compression ratio less than 80 has little influence on the subjective image quality of the 3-D images. In contrast with the subjective image quality of the 3-D scenes, the 2-D subjective image quality decreases greatly when the compression ratio is bigger than 50. So unacceptable compression rates for still 2-D images can be acceptable for the stereoscopic version of the same scene. Two degraded 3-D image pairs can give more information than an original 2-D one. Stelmach and Tam (1998) and Tam et al. (1998) applied MPEG-2 coding in order to reduce bandwidth. They applied a different compression ratio on the left- and right-eye views of a stereoscopic sequence using MPEG-2. In the experiment of Stelmach and Tam (1998), subjects rated the overall subjective image quality of the sequence while the left-eye view was displayed at a higher resolution than the right-eye view. 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. The results of the experiment of Tam et al. (1998) give the correlation between image quality, sharpness and depth. The experimental results show a high correlation between subjective image quality and sharpness and a low correlation between image quality and perceived depth, so image quality of coded images was mainly influenced by sharpness, and less so by perceived depth. The asymmetric MPEG-2 coding results of Stelmach and Tam (1998) show resemblance with basic research on binocular vision. The human visual system weighs incoming information from both eyes when arriving at a final percept.

Subjective image quality regarding asymmetric low-pass filtering has been studied by Stelmach et al. (2000a) and Stelmach et al. (2000b). In Stelmach et al. (2000b), the response of the visual system to spatial and temporal low-pass filtering is investigated and in Stelmach et al. (2000a) the relation between spatial low-pass filtering, quantization and a combination of spatial low-pass filtering and quantization is investigated. In the study of Stelmach et al. (2000b), subjects viewed a mixed resolution stereo sequence in which the right-eye was spatially or temporally low-pass filtered. Results of spatial low-pass filtering show that the subjective image quality of the stereoscopic image is biased towards the eye with the greater spatial resolution. This means that the high spatial frequency information in the left-eye image compensated for the degraded information in the right-eye image. Temporal low-pass filtering produced stereoscopic images with poor quality, so producing mixed resolution sequences using temporal low-pass filtering is unacceptable. In the study of Stelmach et al. (2000a), the perceptual impact of asymmetric coded stereo images using spatial low-pass filtering, quantization and a combination of spatial low-pass filtering and quantization. The results show that the visual system weighs the input from the two eyes depending on the type of coding. For spatial low-pass filtering, the high-quality image dominates the percept and for quantization, the percept is roughly the average of the left- and right-eye inputs. The results of the combination of spatial low-pass filtering and quantization showed that the subjective image quality was dominated by quantization. Spatial low-pass filtering had little effect on the subjective image quality compared to quantization.

IJsselstein et al. (2000c) 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. However, quality judgements were attenuated by the eye strain ratings, thus arriving at a simple stereoscopic image quality model for uncompressed images.

6.2.4 Naturalness

As discussed in the previous section, perceived quality refers to a subjective preference scale. This scale does not necessarily correspond to the most realistic or truthful reproduction. In this context, naturalness is employed as a subjective evaluation concept to refer to the subjective fidelity of the reproduction. 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 and colleagues found a small but systematic deviation between image quality and naturalness. This deviation was interpreted to reflect the subjects' preference for more colorful but, at the same time, somewhat unnatural images (de Ridder et al., 1995; de Ridder, 1996). 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 applying the continuous assessment methodology (ITU, 2000a) to assess viewer's depth, naturalness and presence 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).

Yamanoue (1997) investigated the relation between size distortion and shooting conditions for stereoscopic images. Inconsistency between depth information by lens perspective (focal length) and by binocular parallax in a toed-in camera configuration leads to size distortion. This distortion is called the puppet theater effect, describing the tendency of objects in 3-D images to look unnaturally small. Yamanoue (1997) also concluded that an improvement of the depth sensation in the background (due to camera separation) affects the subjective size of objects. The subjective size of objects is perceived as increasingly small. Subjective evaluation tests showed that images shot with a parallel camera configuration did not cause the puppet theater effect.

In real life the perceived size of an object remains fairly constant independent of its distance and the visual angle. Any change in distance in real life results in a corresponding change in angular size. In 3-D displays both distance and angular size can be decoupled by varying the stereoscopic depth. Introducing these size distortions can make objects look unnaturally large or small. Pastoor (1993) concluded that stereoscopic images look unnaturally small when the camera distance is greater than the screen distance. On the other hand, these distortions are ruled out when the stereo parameters match real-life conditions.

Unnatural visual effects can occur in conventional stereoscopic systems because they provide a fixed pair of views so motion parallax information is missing and the perceived shape of stereoscopic objects is distorted if the viewing position deviates from the centre. These problems can be avoided using a multiview display, generating a proper perspective view corresponding to the observer's viewpoint.

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 observer's 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 3-D programs shot under the same conditions look more natural than those shot using the toed-in camera configuration at any display size.

6.2.5 Presence and enjoyment

With the rapid developments in the area of 3-D, immersive, multisensory displays, and the increased availability of transmission bandwidth, computing power and digital resources, we are able to create and experience reproductions and simulations of reality with an unprecedented sensory quality, blurring the distinction between reality and its representation. Particularly in the area of broadcast displays, recent technological advances have been aimed at improving the reproduction and scope for both sound and vision, including wide-screen, high-definition displays, immersive television, stereoscopic television, and directional audio formats.

With the increased perceptual realism and impact of these media, additional evaluation concepts and tools are needed that go beyond evaluating image or sound quality alone and are sensitive to the overall psychological impact a display has on the viewer. When aiming to evaluate the overall viewing experience, a particularly relevant user experience is that of presence (IJsselsteijn et al., 2002). When users are exposed to immersive and perceptually realistic media, they report a sense of "being there" in the scene - of becoming "part of the action". In addition to such subjective judgements, behavioral and psychophysiological responses may be provoked that are similar to those in non-mediated environments. Such responses could potentially serve as objective indicators of presence, corroborating the results of subjective assessments (IJsselsteijn et al., 2000a).

Stereoscopic displays are expected to give the viewer a heightened sense of presence. IJsselsteijn et al. (1998b) investigated the subjective feeling of presence in 3-D TV and the relationship of presence to perceived depth and image content. Subjects continuously assessed their impression of presence, sensation of depth and the perceived naturalness of depth. The image material varied greatly in terms of both content and depth cues present. IJsselsteijn et al. (1998b) found that subjective presence ratings are subject to considerable temporal variation depending on the image content and camera techniques (motion parallax) used. When depth is perceived as natural, an increase in depth may lead to an enhanced sense of presence. Observers may anchor their judgements onto the salient features of any mediated environment, such as scene cuts in the case of television, because presence is a fairly unfamiliar concept to most naive observers.

Freeman and Avons (2000) conducted a focus group experiment in which they measured presence through 3-D TV. All subjects reported a higher sensation of presence when watching 3-D image material compared to 2-D material. Freeman and Avons (2000) and IJsselsteijn et al. (1998b) concluded

that when stereoscopic images contained exaggerated depth or unnatural depth the enhancement of presence was lessened.

6.2.6 Eye strain

Many studies report a clear preference for stereoscopic images over monoscopic ones. However, viewing stereo images can be more fatiguing than a conventional two-dimensional image. Since eye strain is a potential health hazard that may impede customer acceptance and effective use of stereoscopic displays, as well as their enjoyability, it is important to have an understanding of its subjective magnitude and impact on the user.

A number of studies have investigated eye strain in relation to stereoscopic HMDs. Variable results have been obtained, both objectively and subjectively, with some studies showing clear signs of binocular stress (Mon-Williams et al., 1993), whereas other studies report only mild symptoms (Peli, 1998). Sources of the variability between studies are likely related to differences in the type of HMD under test, and differences in experimental stimuli and protocol, including the method of subjective assessment. Although a number of potential ill effects have been associated with HMD use (e.g. simulator sickness, problems with user calibration, instrument myopia, etc.), one of the potential sources of eye strain that has frequently been suggested in relation to stereoscopic displays in general is the *accommodation-vergence conflict*. As was discussed in Chapter 2, under natural viewing conditions accommodation and vergence operate in a cross-linked fashion, i.e. accommodation may produce vergence movements (i.e. accommodative vergence) and vergence may produce accommodation (i.e. vergence accommodation). Thus, fixating on an object will trigger a rough pre-adjustment of focus, which will be fine-tuned during subsequent accommodation. This reflexively yoked mechanism works well for natural viewing situations (and monoscopic displays) where the plane of focus equals the plane of fixation. For stereoscopic displays however, accommodation will be on the screen plane, where the image is sharpest, whereas fixation may be directed to an object in depth, outside the screen plane (see figure 6.1). In such a case, vergence may drive the accommodation response, which implies that observers' retinal images will be blurred, especially for stereoscopic images containing large image parallaxes.

Several studies provide evidence that vergence accommodation indeed occurs when viewing stereoscopic TV images. Inoue and Ohzu (1990) investigated the accommodation and vergence responses of the eyes to stereoscopic images. They found evidence suggesting that accommodation did not remain on the screen plane (plane of focus) but was adjusted in the direction of the plane of fixation. Hiruma (1990) and Hiruma and Fukuda (1993) reported a study in which subjects' accommodation response to stereoscopic TV images was investigated using an infrared optometer. The results showed that the accommodation response was sensitive to variations in image separation, saturating at large separations.

Changes in visual function have been shown to occur after viewing stereoscopic images for a prolonged period of time. For instance, far-to-near accommodation response times become longer after viewing stereoscopic images for a longer period of time, and the accommodative mechanism (implemented by the circular ciliary muscles that surround the lens of the eye) cannot maintain tonic state (Ohzu and Habara, 1996). The studies cited here thus provide evidence that objectively measurable changes may occur in the human visual system as a consequence of viewing stereoscopic

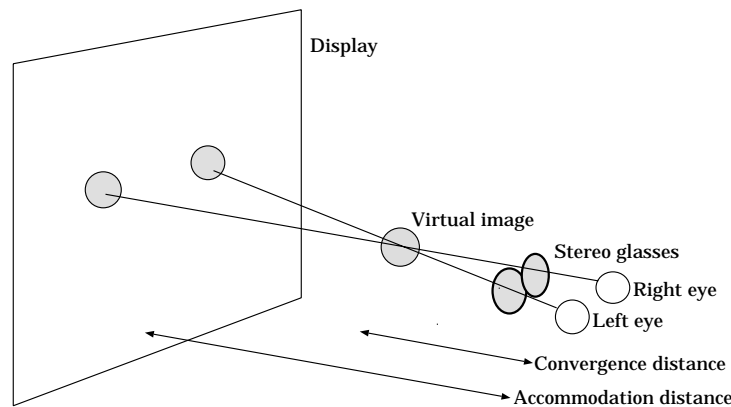


Figure 6.1: Difference between accommodation distance and vergence distance when looking at a stereoscopic display

displays. However, we cannot conclude that these changes by themselves are inherently problematic. Such changes only become an issue when they are shown to have a negative impact on function or comfort. To establish this, subjective tests are needed to investigate the magnitude of potential negative effects and their impact on the user's experience.

IJsselsteijn et al. (2000b) investigated the effect of stereoscopic filming parameters and display duration on the subjective assessment of eye strain. The experimental parameters in this design were camera separation, focal length, convergence distance and display duration. The averaged results of the eye strain ratings show a clear linear increase with increasing disparities, resulting from the filming parameters. The increase in eye strain is larger at short convergence distances because of an increase in keystone distortion at a short convergence distance. There was no significant effect of display duration for eye strain, but the different display durations were relatively short (1-15 seconds). Impact of display duration on experienced eye strain for longer sequences will be of interest to the ATTEST-project. As noted previously, image quality for uncompressed stereoscopic still images can be regarded as a combination of perceived depth, attenuated by the subjective eye strain.

Mitsuhashi (1996) found that observers experience more visual fatigue for binocular stereoscopic television than with the conventional television picture, using an objective measure known as the critical flicker frequency. Okuyama (1999) evaluated visual fatigue with visual function testing (objective evaluation) and interviews (subjective evaluation). Both evaluations show an increased visual fatigue for stereoscopic images.

Kooi and Lucassen (2001) determined the level of eye strain for a wide range of binocular image distortions. The number of unwanted distortions depends on the technique that is used to present the left and right images. When optics are used the left and right image may differ by shift, rotation, magnification, accommodation, or optical distortion. If filters are used the left and right images may differ in luminance, color, sharpness, contrast, accommodation or crosstalk. Kooi and Lucassen (2001) concluded that disparity, crosstalk and blur are the most important parameters that cause eye fatigue. However, no relationship was found between the level of eye fatigue and the visual performance of the viewer (stereopsis, visual acuity, horizontal and vertical phorias and eye dominance).

6.3 Stereoscopic image quality model

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 variables of the display system.

In section 6.3.1, a review is given of several approaches to model the perceived image quality of conventional 2-D images and video sequences. In section 6.3.2, we describe the image quality circle as proposed by Engeldrum (2000) and modifications to make this model suited for the ATTEST purposes of measuring the image quality of stereoscopic image systems.

6.3.1 Approaches towards image quality modelling

Several approaches have been proposed to obtain a quantitative measure of image quality for conventional 2-D images or sequences. For example quality models 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 drawbacks 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, 1996a; Libert and Fenimore, 1999). To relate the attribute strengths to overall image quality, different combination rules can be used (de Ridder, 1992). The visibility of the attribute strengths can be

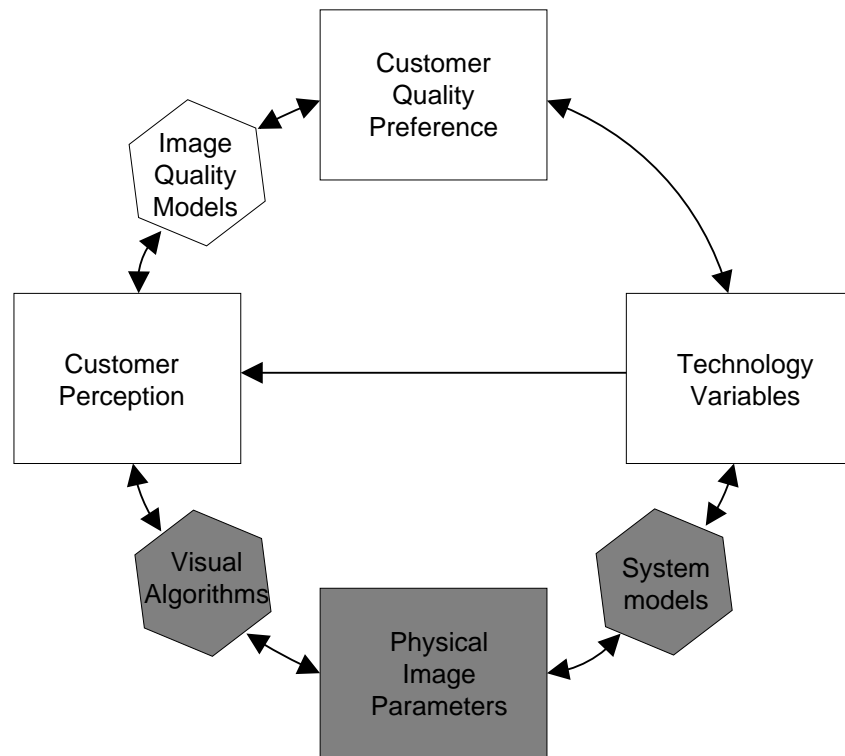


Figure 6.2: Adapted Image Quality Circle from Engeldrum (2000)

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, 1996b; Meesters, 2002).

Another current approach is to consider image quality in terms of the adequacy of the image to enable humans to interact with their environment. In this concept image quality is attributed 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 these image quality attributes usefulness and naturalness, measures of discriminability and identifiability were used (Janssen and Blommaert, 2000).

6.3.2 ATTEST's quality model for stereoscopic image systems

Engeldrum (2000) describes an image quality model as a model of subjective preferences for display systems, which helps manufacturers to implement and integrate image quality into their products. On the basis of his model we will develop a 3-D image quality model. The adapted Image Quality Circle from Engeldrum (2000) is shown in figure 6.2.

The image quality circle is a model, which helps to translate image quality in terms of the technological variables of e.g. a display system. Customer quality preference reflects the customer's opinion about image quality and the judgement they express. The technology variables are a set of parameters that are used to describe the system. Physical image parameters are the measurable properties of the display such as optical density or spectral reflectance.

In the 3-D ATTEST chain one can distinguish three categories of technology variables: variables related to the content generation (e.g. camera separation, focal length, convergence distance), variables related to the coding algorithms (e.g. Q-parameter, bit-rate reduction by quantization) and variables related to the display system and viewing situation (e.g. picture size, viewing distance, viewing angle, luminance, contrast). Customer perceptions such as e.g. sharpness, depth, eye strain, presence, naturalness form the basis of the quality preference or judgement by the customer. Customers do not make image quality judgements based on technology variables, but they express a preference on the basis of what they see. As a first step within ATTEST we will limit ourselves to determine the relationship between technology variables and their perceived effects that contribute to the overall customer's preference.

In this way we will be able to systematically analyse the perceptual issues originating from technological advances. In addition, usability issues will be a topic of considerable interest to WP5, as the ATTEST project aims to develop a feasible 3-D TV broadcast system that will eventually be accepted by consumers in their homes. For this reason, one of the studies currently planned in WP5, will look at 3-D TV viewing behavior in a home-type setting using a broad range of stereoscopic broadcast content.

Bibliography

- Ahumada, A. J. (1993). Computational image quality metrics: a review. *SID Digest*, 24:305–308.
- Aizawa, K. and Huang, T. S. (1995). Model-based image coding: Advanced video coding techniques for very low bit-rate applications. *Proceedings of the IEEE*, 83:259–271.
- An, P., Zhang, Z., and Shi, L. (2001). Theory and experiment analysis of disparity for stereoscopic image pairs. In *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 68–71.
- Aydinoglu, H. and Hayes, M. H. (1994). Compression of multi-view images. *IEEE International Conference on Image Processing, ICIP-94*, 2:385–389.
- Aydinoglu, H. and Hayes, M. H. (1996). Performance analysis of stereo coding algorithms. *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96*, 4:2191–2195.
- Aydinoglu, H. and Hayes, M. H. (1998). Stereo image coding: A projection approach. *IEEE Transactions on Image Processing*, 7:506–516.
- 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. C. (2001). Difscal: A tool for analyzing difference ratings on an ordinal category scale. *Behavior Research Methods, Instruments, & Computers*, 33:10–20.
- Bradshaw, M. F. and Rogers, B. J. (1999). Sensitivity to horizontal and vertical corrugations defined by binocular disparity. *Vision Research*, 39:3049–3056.
- Bruce, V., Green, P., and Georgeson, M. (1996). *Visual perception: Physiology, psychology and ecology*. Psychology Press, E.Sussex, UK.
- Bruno, N. and Cutting, J. (1988). Minimodularity and the perception of layout. *Journal of Experimental Psychology: General*, 117:161–170.
- Chang, G.-C. and Lie, W.-N. (2000). Multi-view image compression and intermediate view synthesis for stereoscopic applications. *IEEE International Symposium on Circuits and Systems, ISCAS-2000*, 2:277–280.
- Chen, J.-Y., Liwei, Z., and Ding, S.-Q. (1998). The effect of JPEG coding scheme on the perceived quality of 3-D images. *SID symposium*, 29:1211–1214.

- Cruz-Neira, C., Sandin, D., and DeFanti, T. (1993). Surround-screen projection-based virtual reality: The design and implementation of the CAVE. *Computer Graphics: Proceedings of SIGGRAPH*, pages 135–142.
- Cutting, J. and Vishton, P. (1995). Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In Epstein, W. and Rogers, S., editors, *Perception of Space and Motion*, pages 69–117. Academic Press, San Diego.
- 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. The 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 variations in color images of natural scenes. *Journal of Imaging Science and Technology*, 40:487–493.
- de Ridder, H. (2001). Cognitive issues in image quality. *Journal of Electronic Imaging*, 10:47–55.
- 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.
- Dinstein, I., Guy, G., Rabany, J., Tzelgov, J., and Henik, A. (1988). On stereo image coding. *Proceedings of 9th International Conference on Pattern Recognition*, 1:357–359.
- Duwaer, A. and van den Brink, G. (1981). What is the diplopia threshold? *Perception and Psychophysics*, 29:295–309.
- Engeldrum, P. (2000). *Psychometric Scaling*. Imcotek Press, Winchester, Massachusetts, USA.
- Eskicioglu, A. A. and Fisher, P. S. (1995). Image quality measures and their performance. *IEEE Transactions on Communications*, 43:2959–2965.
- Falkus, D. (1996). Digital TV: A testing problem. *International Broadcasting*, January:17–19.
- Fehn, C., Kauff, P., Op de Beeck, M., Ernst, F., IJsselsteijn, W., Pollefeys, M., van Gool, L., Ofek, E., and Sexton, I. (2002). An evolutionary and optimised approach on 3D-TV. *IBC 2002*.
- Franich, R. (1996). Disparity estimation in stereoscopic digital images. Ph.D. dissertation, Delft University of Technology, The Netherlands.
- Freeman, J. and Avons, S. (2000). Focus group exploration of presence through advanced broadcast services. *Proceedings of the SPIE*, 3959:530–539.
- Frisby, J., Buckley, D., and Freeman, J. (1996). Stereo and texture cue integration in the perception of planar and curved large real surfaces. In Inui, T. and McClelland, J., editors, *Attention and Performance XVI*, pages 71–91.
- Gonzalez, R. C. and Woods, R. E. (1992). *Digital image processing*. Addison-Wesley publishing company, Inc.

- Gooding, L., Miller, M., Moore, J., and Kim, S. (1991). The effect of viewing distance and disparity on perceived depth. *Proceedings of the SPIE*, 1457:259–266.
- Grau, O., Minelly, S., and Thomas, G. (2001). Applications of depth metadata. In *IBC2001*.
- Haskell, B. G., Puri, A., and Netravali, A. N. (1997). *Digital Video: An Introduction to MPEG-2*. Chapman & Hall, New York.
- Hayes, R. (1989). *3-D Movies. A History and Filmography of Stereoscopic Cinema*. McFarland & Company, Jefferson, NC.
- Hayslip Jr., B. and Panek, P. (1989). *Adult Development and Aging*. Harper & Row, New York, NY.
- Heynderickx, I., Seuntiens, P., Langendijk, E., and IJsselsteijn, W. (2002). Visibility of a color gradient across the screen for stimuli presented time-sequentially or simultaneously. In *SID Digest 2002*, Boston, Massachusetts.
- Hiruma, N. (1990). Accommodation response to binocular stereoscopic tv images. In Noro, K. and Brown Jr., O., editors, *Human Factors in Organizational Design and Management III*, pages 233–236.
- Hiruma, N. and Fukuda, T. (1993). Accommodation response to binocular stereoscopic tv images and their viewing conditions. *SMPTE Journal*, 102:1137–1144.
- Howard, I. and Rogers, B. (1995). *Binocular Vision and Stereopsis*. Oxford University Press, Oxford.
- IJsselsteijn, W., Bouwhuis, D., Freeman, J., and de Ridder, H. (2002). Presence as an experiential metric for 3-D display evaluation. In *SID Digest 2002*, Boston, Massachusetts.
- 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.
- Inoue, T. and Ohzu, H. (1990). Accommodation and convergence when looking at binocular 3D images. In Noro, K. and Brown Jr., O., editors, *Human Factors in Organizational Design and Management III*, pages 249–252.

- 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.
- Jiang, J. and Edirisinghe, E. A. (2002). A hybrid scheme for low bit-rate coding of stereo images. *IEEE Transactions on Image Processing*, 11:123–134.
- Johnston, E., Cumming, B., and Landy, M. (1994). Integration of stereopsis and motion shape cues. *Vision Research*, 34:2259–2275.
- Johnston, E., Cumming, B., and Parker, A. (1993). Integration of depth modules: Stereopsis and texture. *Vision Research*, 33:813–826.
- Julesz, B. (1971). *Foundations of Cyclopean Perception*. University of Chicago Press, Chicago, IL.
- 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: model. *Journal of the Optical Society of America A*, 13:1178–1188.
- Kayargadde, V. and Martens, J.-B. (1996b). Perceptual characterization of images degraded by blur and noise: experiments. *Journal of the Optical Society of America A*, 13:1166–1177.
- Kooi, F. and Lucassen, M. (2001). Visual comfort of binocular and 3-D displays. *Proceedings of the SPIE*, 4299:586–592.
- Landy, M. and Brenner, E. (2001). Motion-disparity interaction and the scaling of stereoscopic disparity. In Harris, L. and Jenkin, M., editors, *Vision and Attention*, pages 129–151. Springer Verlag, New York.
- Landy, M., Maloney, L., Johnston, E., and Young, M. (1995). Measurement and modelling of depth cue combination: In defense of weak fusion. *Vision research*, 35:389–412.
- 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.
- 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.
- 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. dissertation, Eindhoven University of Technology, The Netherlands.

- Meyer, L. and Fontaine, G. (2000). Development of digital television in the European Union. Final Report to the European Commission DG XIII, June 2000, Institut de l'Audiovisuel et des Télécommunications en Europe (IDATE).
- Mitsuhashi, T. (1996). Evaluation of stereoscopic picture quality with CFF. *Ergonomics*, 39:1344–1356.
- Mon-Williams, M., Wann, J., and Rushton, S. (1993). Binocular vision in a virtual world: Visual deficits following the wearing of a head-mounted display. *Ophthalmic and Physiological Optics*, 13:387–391.
- Motoki, T., Isono, H., and Yuyama, I. (1995). Present status of three-dimensional television research. *Proceedings of the IEEE*, 83:1009–1021.
- Naemura, T., Kaneko, M., and Harashima, H. (1999). Compression and representation of 3-D images. *IEICE Trans. Inf. & Syst.*, E82-D:558–567.
- Norman, J., Dawson, T., and Butler, A. (2000). The effects of age upon the perception of depth and 3-D shape from differential motion and binocular disparity. *Perception*, 29:1335–1359.
- Ohzu, H. and Habara, K. (1996). Behind the scenes of virtual reality: Vision and motion. *Proceedings of the IEEE*, 84:782–798.
- Okoshi, T. (1980). Three-dimensional displays. *Proceedings of the IEEE*, 68:548–564.
- Okuyama, F. (1999). Evaluation of stereoscopic display with visual function and interview. *Proceedings of the SPIE*, 3639:28–35.
- Op de Beeck, M., Fert, E., Fehn, C., and Kauff, P. (2002). Broadcast requirements on 3D video coding. *ISO/IEC/JTC1/SC29/WG11 MPEG02/M8040*.
- Op de Beeck, M. and Redert, A. (2001). Three dimensional video for the home. *Proceedings of EUROIMAGE ICAV3D 2001, International Conference on Augmented, Virtual Environments and Three-Dimensional Imaging*, pages 188–191.
- Palmer, S. (1999). *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge, Massachusetts.
- Pastoor, S. (1993). Human factors of 3D displays in advanced image communications. *Displays*, 14:150–157.
- Pastoor, S. and Wöpking, M. (1997). 3-D displays: A review of current technologies. *Displays*, 17:100–110.
- Patterson, R. (1997). Visual processing of depth information in stereoscopic displays. *Displays*, 17:69–74.
- Patterson, R. and Fox, R. (1984). The effect of testing method on stereoanomaly. *Vision research*, 24:403–408.

- Patterson, R. and Martin, W. L. (1992). Human stereopsis. *Human Factors*, 34:669–692.
- Peli, E. (1998). The visual effects of head-mounted display (HMD) are not distinguishable from those of desk-top computer display. *Vision Research*, 38:2053–2066.
- Perkins, M. G. (1992). Data compression of stereopairs. *IEEE Transactions on Communications*, 40:684–696.
- Reynolds, W. D. and Kenyon, R. V. (1996). The wavelet transform and the suppression theory of binocular vision for stereo image compression. *3rd IEEE International Conference on Image Processing*, 1:557–560.
- Richards, W. (1970). Stereopsis and stereoblindness. *Experimental Brain Research*, 10:380–388.
- Rogers, B. and Graham, M. (1982). Similarities between motion parallax and stereopsis in human depth perception. *Vision Research*, 22:261–270.
- Roufs, J. A. J. (1992). Perceptual image quality: Concept and measurement. *Philips Journal of Research*, 47:35–62.
- Rule, J. (1941). The geometry of stereoscopic pictures. *Journal of the Optical Society of America*, 31:325–334.
- Schertz, A. (1992). Source coding of stereoscopic television pictures. In *IEE Inter. Conf. on image processing and its applications*, pages 462–464, Maastricht, The Netherlands.
- Seferidis, V. and Papadimitriou, D. (1993). Improved disparity estimation in stereoscopic television. *Electronics Letters*, 29:782–783.
- Sethuraman, S., Siegel, M. W., and Jordan, A. G. (1995). A multiresolutional region-based segmentation scheme for stereoscopic image compression. *Proceedings of the SPIE*, 1429:265–274.
- Sexton, I. and Surman, P. (1999). Stereoscopic and autostereoscopic display systems. *IEEE Signal Processing Magazine*, pages 85–99.
- Siegel, M., Sethuraman, S., McVeigh, J. S., and Jordan, A. (1997). Compression and interpolation of 3D-stereoscopic and multi-view video. *Proceedings of the SPIE*, 3012:227–238.
- Smith, C. and Dumbreck, A. (1988). 3-D TV: The practical requirements. *Television: Journal of the Royal Television Society*, pages 9–15.
- Spottiswoode, R., Spottiswoode, N., and Smith, C. (1952). 3-D photography - Basic principles of the three-dimensional film. *Journal of the SMPTE*, 59:249–286.
- Stelmach, L., Tam, W., Meegan, D., Vincent, A., and Corriveau, P. (2000a). Human perception of mismatched stereoscopic 3D inputs. *IEEE international conference on image processing*, 1:5–8.
- Stelmach, L., Tam, W. J., Meegan, D., and Vincent, A. (2000b). 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.
- Strintzis, M. and Malassiotis, S. (1998). Review of methods for object-based coding of stereoscopic and 3D image sequences. *IEEE International Symposium on Circuits and Systems*, 5:510–513.
- Tam, W. and Stelmach, L. (1998a). Display duration and stereoscopic depth perception. *Canadian Journal of Experimental Psychology*, 52:56–61.
- Tam, W., Stelmach, L., and Corriveau, P. (1998). Psychovisual aspects of viewing stereoscopic video sequences. *Proceedings of the SPIE*, 3295:226–235.
- Torgerson, W. S. (1958). *Theory and methods of scaling*. John Wiley & Sons, New York.
- Travis, A. (1990). Autostereoscopic 3-D display. *Applied Optics*, 29:4341–4342.
- Tseng, B. L. and Anastassiou, D. (1995a). Perceptual adaptive quantization of stereoscopic video coding using MPEG-2's temporal scalability structure. In *Proceedings of the International Workshop on Stereoscopic and Three-Dimensional Imaging*, pages 52–57.
- Tseng, B. L. and Anastassiou, D. (1995b). A theoretical study on an accurate reconstruction of multiview images based on the viterbi algorithm. *International Conference on Image Processing, ICIP-95*, 2:378–381.
- Tzovaras, D., Grammalidis, N., and Strintzis, M. G. (1998). Disparity field and depth map coding for multiview 3D image generation. *Signal Processing: Image Communication*, 11:205–230.
- 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.
- van der Meer, H. (1979). Interrelation of the effects of binocular disparity and perspective cues on judgments of depth and height. *Perception and Psychophysics*, 26:481–488.
- van Dijk, A. and Martens, J. (1996). Subjective quality assessment of compressed images. *Signal Processing*, 58:235–252.
- Wade, N. (1998). *A Natural History of Vision*. MIT Press, Cambridge, Massachusetts.
- Wandell, B. A. (1995). *Foundations of vision*. Sinauer Associates, Inc, Sunderland, Massachusetts.
- Watson, A. B. (1987). Efficiency of a model human image code. *Journal of the Optical Society of America A*, 4:2401–2417.
- Watson, A. B. and Kreslake, L. (2001). Measurement of visual impairment scales for digital video. *Proceedings of the SPIE*, 4299:79–89.
- Winkler, S. (1999). Issues in vision modelling for perceptual video quality assessment. *Signal Processing*, 78:231–252.

- Woo, G. and Sillanpaa, V. (1979). Absolute stereoscopic threshold as measured by crossed and uncrossed disparities. *American Journal of Optometry and Physiological Optics*, 56:350–355.
- Woo, W. and Ortege, A. (1999). Optimal blockwise dependent quantization for stereo image coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 9:861–867.
- Woods, A., Docherty, T., and Koch, R. (1993). Image distortions in stereoscopic video systems. *Proceedings of the SPIE*, 1915:36–48.
- Woods, A., Docherty, T., and Koch, R. (1996). 3D video standards conversion. *Proceedings of the SPIE*, 2653:210–218.
- Woodson, W. (1981). *Human factors design handbook*. McGraw-Hill, New York, NY.
- Wu, H. R., Yuen, M., and Qiu, B. (1996). Video coding distortions classification and quantitative impairment metrics. In *Proceedings of ICSP'96*, pages 962–965.
- Xu, J., Xiong, Z., and Li, S. (2002). High performance wavelet-based stereo image coding. *IEEE International Symposium on Circuits and Systems*, 2:273–276.
- Yamaguchi, H., Tatehira, Y., Akiyama, K., and Kobayashi, Y. (1989). Stereoscopic images disparity for predictive coding. *International Conference on Acoustics, Speech, and Signal Processing, 1989. ICASSP-89.*, 3:1976–1979.
- Yamanoue, H. (1997). The relation between size distortion and shooting conditions for stereoscopic images. *Journal of the SMPTE*, pages 225–232.
- Yamanoue, H., M.Nagayama, M.Bitou, and and, J. (1998). Orthostereoscopic conditions for 3D HDTV. *Proceedings of the SPIE*, 3295:111–120.
- Yamanoue, H., Okui, M., and Yuyama, I. (2000). A study on the relationship between shooting conditions and cardboard effect of stereoscopic images. In *IEEE Transactions on Circuits and Systems for Video Technology*, volume 10, pages 411–416.
- Yano, S. and Yuyama, I. (1991). Stereoscopic HDTV: Experimental system and psychological effects. *Journal of the SMPTE*, 100:14–18.
- Yuen, M. and Wu, H. (1998). A survey of hybrid MC/DPCM/DCT video coding distortions. *Signal Processing*, 70:247–278.
- Ziegler, M., Falkenhagen, L., ter Horst, R., and Kalivas, D. (1998). Evolution of stereoscopic and three-dimensional television. *Signal Processing: Image communication*, 14:173–194.
- Zone, R. (1996). The deep image: 3D in art and science. *Proceedings of the SPIE*, 2653:4–8.

Index

- ATTEST, 4, 22, 23, 29, 30, 54, 57
- Binocular rivalry, 10, 26, 29, 33
- Camera
 - ZcamTM, 22
 - configuration, 27
 - converging, 18, 48
 - geometry, 16, 18
 - IR range, 22
 - keystone distortion, 20, 54
 - nearness factor, 16
 - parallel, 16, 51
 - perspective transformation, 16, 18
 - stereoscopic, 15
 - toed-in, 18, 51, 52
- Cardboard effect, 34, 48
- Coding distortions
 - blockiness, 31, 33, 47, 55
 - blurring, 32–34, 47, 48, 55
 - cardboard effect, 34
 - color bleeding, 32
 - double contouring, 32, 34
 - ghosting, 32, 33
 - jerkiness, 32–34
- Compression, *see* Stereoscopic compression
 - 2-D compression, 28, 31
 - lossless, 26
 - perceptually lossless, 26, 31
 - perceptually lossy, 26, 31
- Content generation
 - 2D-to-3D, 23
 - stereoscopic camera, 15
 - stereoscopic films, 2
- Cross-talk, 37, 38, 54
- DCT, *see* Transform coding
- Depth, 39, 48
 - cardboard effect, 34
- cues, 7, 34
 - averaging, 11
 - disambiguation, 11
 - dissociation, 11
 - dominance, 11
 - reinterpretation, 11
 - summation, 11
- map, 22
- perceived, 22, 34, 48
- perception, 7
- Diplopia, 8, 44, 48
- Disparity, 7, 15, 30, 35, 39, 47–49, 54
 - crossed, 8
 - estimation, 27
 - horizontal, 16, 18, 20
 - uncrossed, 8
 - values, 35
 - vertical, 18, 20
- Display
 - 3-D TV, 3, 4, 22, 39, 41
 - 3D-IMAX, 2
 - anaglyph, 38
 - autostereoscopic, 39
 - direction-multiplexed, 39
 - electro-holographic, 40
 - holographic, 39
 - lenticular, 40
 - parallax barrier, 39
 - volumetric, 39, 40
 - cross-talk, 37, 38
 - HDTV, 3
 - history, 1
 - monoscopic, 30
 - multiview, 40, 41, 51
 - Pulfrich effect, 39
 - stereoscope, 1
 - stereoscopic, 30, 37

- circular polarization, 38
 - location multiplexing, 38
 - polarization-multiplexed, 38
 - time-multiplexed, 37
 - time-parallel, 37, 38
 - stereoscopic cinema, 1
 - stereoscopic television, 1, 2
- Eye strain, 10, 53
- Focus group, 52
- Hopter, 8
- Hyperacuity, 10
- Image quality, 49
 - appreciation oriented, 44
 - model, 55
 - objective, 31
 - perceptual, 31, 33
 - performance oriented, 44
- Image redundancy
 - coding redundancy, 25
 - inter-pixel redundancy, 25
 - psycho-visual redundancy, 26
- Inter-pupillary distance (IPD), 20
- JPEG, 28, 33, 34, 50
- Keystone distortion, 20, 54
- Low-pass filtering, 32, 49, 50
- Motion parallax, 30, 49, 51
- MPEG, 28, 33, 34, 49, 50
- Multiview, 30, 40, 41
 - intermediate-view, 30
 - key-view, 30
 - motion parallax, 30
 - multiviewer scenario, 30
- Naturalness, 51
- Occlusion, 28, 30
- Orthostereoscopic view, 52
- Panum's fusional area, 8
- Perceptual attributes, 47
 - depth, 48
 - eye strain, 53
 - naturalness, 51
 - presence, 52
 - sharpness, 47
 - subjective image quality, 49
- Presence, 52
- Pulfrich effect, 39
- Puppet-theater effect, 49, 51
- Quantization, 26, 28, 32, 34
- Range finding techniques
 - active, 28
 - passive, 28
- Sharpness, 33, 47
- Stereoblindness, 12
- Stereopsis, 27
- Stereoscopic compression, 47, 50
 - asymmetric, 30, 32
 - block-based, 27
 - camera configuration, 27
 - DCTDP, 27
 - depth map, 28
 - depth-annotated images, 29
 - disparity, 30, 35
 - disparity estimation, 27
 - feature-based, 27, 28
 - intensity-based, 27
 - mixed resolution, 29, 48
 - multiview, 30
 - symmetric, 32
- Strabismus, 13
- Subjective assessment, 43
 - context effects, 46
 - DSCQS, 44
 - DSIS, 45
 - focus group, 43
 - matching, 47
 - single-stimulus, 44
 - SSCQE, 45
 - stimulus presentation, 46
 - stimulus-comparison, 44
- Transform coding
 - DCT, 26, 33

wavelet, 26, 28, 29

Vieth-Muller circle, 8

Viewing factors, 20

Visual discomfort, 15, 21, 22

Wavelet, *see* Transform coding