DISS. ETH NO. 21730

Computational Models for Stereoscopic Perception

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
Steven Charles Poulakos
Master of Science,
Graduate School of Library and Information Science,
University of Illinois at Urbana-Champaign
born on 13.03.1977
citizen of United States of America

accepted on the recommendation of
Prof. Dr. Markus Gross, examiner
Prof. Dr. Martin Banks, co-examiner
Dr. Aljoscha Smolic, co-examiner

2014

Abstract

Binocular vision enables a precise estimation of depth in visual space through a process called stereopsis. Combined with other visual cues to depth, humans can efficiently interpret a complex three-dimension world. Various forms of technology exist to recreate this visual experience. The most common today is the use of stereoscopic 3D images or video. The fundamental idea is to present two images, one to each eye, which are fused to provide a compelling sensation of depth. Stereoscopic image viewing, however, introduces several perceptual distortions, which impact visual quality, depth interpretation and comfort.

Since the complexity is high, there is a need to develop technology to facilitate understanding of how stereoscopic images are perceived. The technology is based on computational models that encompass theory, representation and implementation of the models. This thesis presents a combination of perceptual models based on existing research and novel experimental methods. The findings support the challenging, yet important endeavor to develop models of stereoscopic image quality.

This thesis makes four primary contributions. First, a collection of modeling topics related to stereoscopic imaging are presented. This helps to frame the space of limitations influencing both the presentation and perception of stereoscopic images. A second contribution is the exploration of how a dominant perceptual conflict between eye vergence and accommodation influences a viewers ability to change visual attention. A third contribution is the exploration of another perceptual distortion when one of the strongest depth cues, occlusion, is in conflict with another strong cue, stereopsis. Finally, because visual attention plays such a critical role in the perception of stereoscopic depth, we develop a framework producing an edge-aware, spatio-temporally smooth stereoscopic saliency representation.

Zusammenfassung

Das binokulare Sehen ermöglicht eine präzise Schätzung der Tiefe im sichtbaren Raum durch einen Prozess namens Stereopsis. In Kombination mit anderen optischen Tiefenmerkmalen kann der Mensch die komplexe dreidimensionale Welt effizient interpretieren. Heutzutage existieren verschiedene Technologien, die das visuelle Erlebnis replizieren können. Am häufigsten werden stereoskopische Bilder oder stereoskopisches Video verwendet. Die grundlegende Idee besteht darin, zwei Bilder, eines für jedes Auge, so zu präsentieren, dass diese fusioniert werden, um dann ein überzeugendes Gefühl von Tiefe zu bieten. Stereoskopische Bildbetrachtung führt jedoch zur Verzerrungen der Wahrnehmung, welche die optische Qualität, Tiefenempfinden und den Komfort des Betrachters beeinflussen.

Da die Komplexität hoch ist, besteht ein Bedarf dafür Technologien zu entwickeln, die zum Verständnis beitragen, wie stereoskopische Bilder wahrgenommen werden. Solche Technologie basieren oft auf computergestützte Modelle, die aus Theorie, Darstellung und Umsetzung der Modelle bestehen. Diese Arbeit stellt eine Kombination von Wahrnehmungsmodellen auf Grundlage der aktuellen Forschung sowie neuen experimentellen Forschungsmethoden dar. Die Ergebnisse unterstützen das schwierige aber wichtige Bestreben Modelle für die Qualität von stereoskopischen Bildern zu entwickeln.

Diese Arbeit beinhaltet vier Hauptbeiträge. Zuerst wird eine Taxonomie über Wahrnehmungs- und technologische Stereoskopiethemen vorgestellt. Diese Taxonomie beschreibt insbesondere die Grenzen, die die Darstellung und Wahrnehmung von stereoskopischen Bildern beeinflussen. Ein zweiter Beitrag ist die Erforschung des dominanten Wahrnehmungskonflikts zwischen der Konvergenz und der Akkommodation des Auges sowie wie dieser Konflikt die Fähigkeiten des Betrachters, seine visuelle Aufmerksamkeit zu ändern, beeinflusst. Ein dritter Beitrag besteht in der Erforschung einer anderen Wahrnehmungsverzerrung, die dann entsteht, wenn die stärksten Tiefenmerkmale, nämlich die Verdeckungen, in einem Konflikt mit dem stereoskopischen Sehen sind. Auf Grund der Tatsache, dass die visuelle Aufmerksamkeit eine so wichtige Rolle in der Wahrnehmung der stereoskopischen Tiefe spielt, entwickeln wir schliessich ein System zur Herstellung einer räumlich und zeitlich glatten aber kantensensitiven Merkmalskarte zur Darstellung von stereskopisch-hervorstechenden Merkmalen.

Acknowledgements

There are many people to thank along the journey of my PhD thesis. First of all, I would like to sincerely thank my advisor, Prof. Markus Gross. His enthusiasm, focus and dedication to a wide variety of research topics is truly inspiring. Many thanks go to Dr. Aljoscha Smolic who leads our Advanced Video Technology group at Disney Research Zurich (DRZ) and also provided significant direction and support during my thesis work. Also special thanks go to Prof. Thomas Gross and Prof. Cary Kornfeld for the initial opportunity to delve into the uncharted territory of stereoscopic 3D technology and visual perception topics. I would also like to thank Prof. Marty Banks for his advice, support and for providing many examples of excellent spatial vision research.

Special thanks go to Gerhard Roethlin and Rafael Monroy for choosing to do master theses with me resulting in significant research contributions. I would also like to thank other collaborators during my thesis including Jeroen van Baar, Manuel Lang, Tunc Aydin, Federico Perazzi, Niko Stefanoski, Paul Johnson, Oliver Wang, Alex Chapiro, Simon Heinzle, Rasmus Tamstorf, Wojciech Jarosz, Alex Sorkine-Hornung, Maurizio Nitti, Derek Nowrouzezahrai, Katharina Quintus, Jisien Yang, Rafael Huber and Prof. Adrian Schwaninger.

All of my friends and colleagues at DRZ as well as the Institute for Visual Computing and Computer Systems Institute at ETH Zurich have provided an energetic and fun environment for doing research. Each group fosters creativity and produces significant results. I would also like to thank those who helped build and maintain the original stereoscopic imaging lab, including Marco Des Santis, Noe Lutz and Nicoletta De Maio. We built a very capable stereoscopic video production environment with basic hardware and a lot of ingenuity. I would like to thank the administrative and support staff at ETH Zurich and Disney Research Zurich for their support. We sometimes had unusual requests to build experimental systems or develop stereoscopic production equipment.

Without funding, research would not be possible. I am grateful for the ETH Research Grant TH-23/04-3, which funded my initial exploration of stereoscopic 3D perception. The founding of Disney Research Zurich and opportunity to further my studies within that environment provided not only continued financial support, but also rich source of industrial experience in stereoscopic media. The transition from doctoral to post-doctoral research and ex-

ploration of immersive stereoscopic 3D applications was made possible by the European Commission under the Contract FP7-ICT-287723 REVERIE.

Finally, I thank all friends and family for their support, encouragement and generally for keeping life interesting. Thank you to my parents, William and Karen, for motivating my studies and brothers, Michael and Nicholas, as well as my parents-in-law, William and Barbara. I would especially like to thank my wife, Anneliese, and children, Elena and Timothy, for their endless support and sharing the journey with me.

Contents

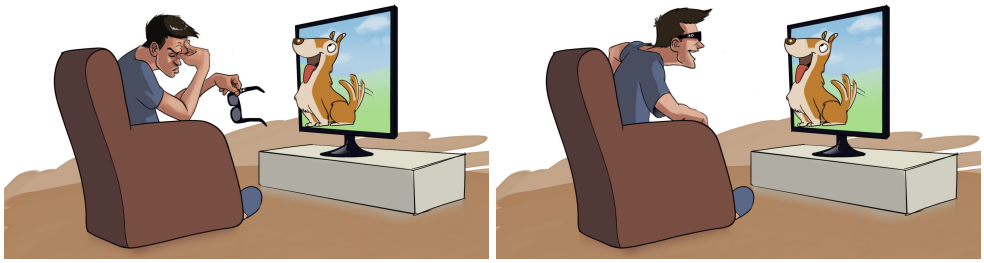
CHAPTER 1

Introduction

Stereoscopic 3D (S3D) has experienced a revival in entertainment applications including cinema, television and interactive games. Digital video technology has enabled much of the recent success of S3D, making it possible to more easily capture, edit, transmit and display stereoscopic content. While the continued success of S3D benefits from digital video technology, human factors, in terms of quality of viewing experience, are gaining importance. The aim is to produce compelling stereoscopic content that provides a more immersive visual experience while not increasing visual discomfort and fatigue.

Evaluating the quality of stereoscopic viewing experience is a challenging task, often requiring an experienced professional stereographer who can predict when a stereoscopic scene composition will be visually problematic. This is a complex process requiring much experience to balance artistic, perceptual and technical aspects of S3D production. In order to assist more stereoscopic content creators in the production of "good" stereoscopic 3D images or video, technological solutions are needed which provide guidance or quality analysis. This requires computational models of how S3D content is perceived.

David Marr [1982] proposed the importance of a information-processing perspective on computational modeling. He advocated that while algorithms



(a) Poor quality of experience (b) Good quality of experience

Figure 1.1: *Simple assessment of quality of experience. (a) Although the visual content may seem good, the viewer experiences visual discomfort and is motivated to stop watching. (b) Subtle changes to the same content that maintain visual comfort enables to the viewer to better engage in the visual experience.*

and mechanisms facilitate understanding a system, it is more important to know the nature of computations underlying perception of a computational problem. Essentially, the nature of the problem is more important than an algorithm that solves the problem. Marr identifies three levels for carrying out information-processing tasks. First, a *computation theory* must be developed to understand what must be computed and why. Second, it is necessary to determine the information *representation and algorithm* that can achieve the computational theory. Finally, a *hardware implementation* must be capable of physically realizing the data representation and algorithm.

1.1 Motivation

This thesis aims to explore computational modeling as it relates to the perception of stereoscopic image viewing. At a high level, the task can seem easy: Simply build a detector for a specific problematic situation. However, building the detector can be quite challenging and can also be confounded by a variety of perceptual factors. The aim of this thesis is to explore the space of existing computational models of stereoscopic image perception. We then explore several aspects of stereoscopic perception with the goal of understanding specific stereoscopic image quality problems and the application of that knowledge in computational systems.

We utilize empirical methods to evaluate quality of experience and to construct a computational perceptual model. The goal is to identify the sometimes subtle factors that influence the viewing experience as represented in

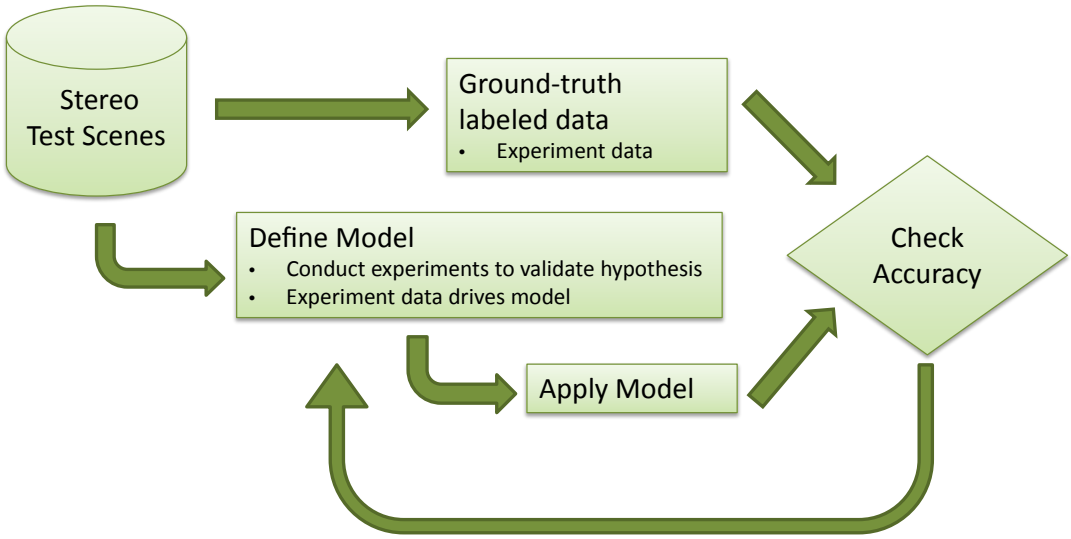


Figure 1.2: *Typical workflow for computational system development. The bottom portion of the flow diagram represents the construction of model by careful creation of visual stimuli and application of that model to real image content. The upper portion of the flow represents the development of ground truth or user labeled data, which is used to evaluate model performance. This is an iterative process requiring refinement.*

Figure 1.1. Quality of stereoscopic viewing experience is a complex, multidimensional problem, which is best explored in parts that end up in a meaningful technical system. The experience of professional artists and content creators is useful for getting first-hand experience with problematic situations encountered while presenting stereoscopic 3D. However, building computational systems of stereoscopic visual perception requires interdisciplinary skills ranging from image processing and computer vision to psychophysics and neurophysiology. The combination of these diverse domains enables a better prediction of visual experience.

In additional to combining knowledge from diverse domains, developing a computational system generally requires a workflow as represented in Figure 1.2. It is necessary to create visual stimuli to facilitate the creation of a perceptual model. The stimuli should enable modification to specific perceptual factors without introducing additional confounding factors. The model can then be applied to real image content that is representative of the problem motivating the computational system development. It is also useful to create ground truth data, used to evaluate the performance of the computational model. An iterative process is often necessary to refine the stimuli and also to create representative ground truth stimuli.

1.2 Thesis Overview

A significant challenge in stereoscopic perceptual research is understanding the limitations imposed by both human perception and stereoscopic image technologies. Many dimensions exist with an ever growing amount of research. Chapter 2, *Modeling Topics in Stereoscopic Imaging*, provides a structured overview of relevant topics and significant research. This knowledge has not only been utilized to support research contributions of this thesis, but also plays a critical role in formulating computational theories about stereoscopic perception.

Chapter 3, *Attention Transitions in Stereoscopic Depth*, presents novel research demonstrating how additional visual information embedded in a stereoscopic scene can facilitate the time to achieve visual attention transitions. The decoupling of eye vergence and accommodation is believed to hinder attention transitions, and we demonstrate the use of a visual cue to help compensate. Both objective performance measures and subjective self-assessments were utilized to observe visual performance and comfort during attention transition tasks. The findings provide a significant example of how stereoscopic 3D content creators may learn scene composition, framing and montage from visual psychophysics.

Another significant stereoscopic distortion is presented in Chapter 4, *Stereoscopic Window Violations*. In this case, the dominant conflict is between the perception of occlusion and stereopsis depth cues near the image border. We demonstrate how experimental psychophysics can be used to both develop perceptual models and validate the results. Our window violation detector assists content creators in identifying problematic window violations so the scene composition may be manually adjusted or to utilize automatic techniques for removing the violation.

Visual attention is a critical component of any visual quality metric. On one hand, visual distortions can influence visual attention. On another hand, visual attention can help guide the image regions where perceptual quality should have greater significance. In Chapter 5, *Multimodal Stereoscopic Saliency*, we demonstrate how multiple models of visual saliency estimation can be combined to provide an edge-aware, spatio-temporally smooth saliency map for stereoscopic video. We present the challenges constructing and analyzing a stereoscopic data set as well as the performance of our saliency model on that data set. Stereoscopic saliency enables many useful forms of quality control and content manipulation, such as the remapping of stereoscopic depth.

The thesis concludes by summarizing the key results, which span the space of utilizing existing perceptual models, validating hypothesis through visual psychophysics, development of new perceptual models and creation of applications for those models. An outlook for future research is then discussed, which provides some future directions for exploration of the challenging task of modeling the quality of visual experience for stereoscopic 3D content.

CHAPTER 2

Modeling Topics in Stereoscopic Imaging

Exploration in stereoscopic vision research benefits from understanding the many aspects influencing the perception of stereoscopic images. This chapter presents a collection of relevant stereoscopic topics and perceptual models that were explored during the thesis, either to support the design of experiments or the creation and analysis of stereoscopic content.

2.1 Introduction

Over the past century, knowledge of visual perception has grown considerably, which has lead to the development of many models of the perception of both 2D and stereoscopic 3D image content. Digital image and video technology has enabled the creation and presentation of specialized visual content, supporting analysis of many perceptual topics. This chapter provides a structured overview of relevant topics, significant research and computational models influencing our understanding of stereoscopic 3D perception.

We have applied this knowledge in many ways throughout the thesis. First, it guides the creation of stimuli, both the carefully controlled stimuli to iso-

late experiment factors as well as stimuli that is representative of real-world stereoscopic content. Second, this knowledge has guided the process of creating and presenting stereoscopic content while exploring the stereoscopic content production pipeline. Several short stereoscopic movies have been produced using knowledge presented in this chapter. Third, it has guided our work in disparity editing (aka. disparity remapping) to recompose stereoscopic content to be more comfortable or pleasing to view.

Chapter Organization. The chapter is organized into the following sections:

- Depth Interpretation
- Stereopsis
- Vergence, Accommodation, and Comfort
- Distortions
- Attention and Saliency

2.2 Depth Interpretation

Human vision utilizes many different types of visual information to perceive depth in visual space. This section presents cues to depth as well as theories about cue integration and examples of the effects of conflicting cues.

2.2.1 Depth Cues

Depth interpretation is influenced not only by binocular disparity, but also a collection of other depth cues. Depth cues include the following classification described by Cutting and Vishton [1995]:

- Occlusion - provides depth ordering, nearer objects occlude farther objects [pictorial]
- Relative size and relative density - can provide scaled information about depth, based on the retinal size of objects (or textures) or retinal density of clusters of objects (or textures). Texture gradients and linear perspective can also be considered a subset. Size information from light and shading can also be grouped here. [pictorial]

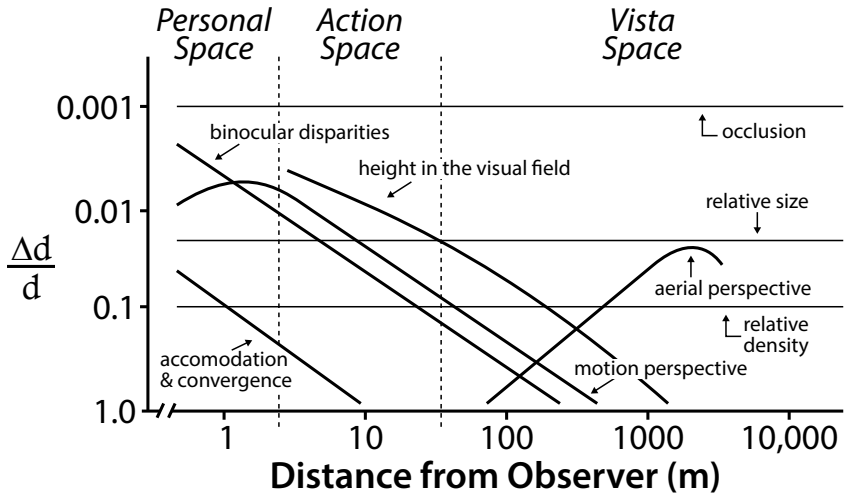


Figure 2.1: *Just-discriminable depth thresholds as a function of the log of distance from the observer [Cutting and Vishton, 1995].*

- Height in Visual Field - when viewed from above a planar surface, objects on the plane will appear higher with greater distance. [pictorial]
- Aerial perspective - moisture and pollutants in the atmosphere represent forms of *participating media* that decrease visual contrast with distance. [pictorial]
- Motion perspective - more distant objects appear to move more slowly. [motion]
- Convergence and accommodation - self-awareness (or proprioception) of eye convergence and lens accommodation can influence perception of depth. [oculomotor] Perception of blur can also provide a pictorial cue to depth. [pictorial]
- Binocular disparity, stereopsis and diplopia - relative difference in projection of the same object on the two eyes. [stereopsis]

2.2.2 Depth Cue Integration

Depth cue integration is the process of combining depth cues to produce a single perceived depth interpretation. There are different approaches to depth cue integration including: cue range in visual space, cue dominance, cue averaging, and cue spatialization.

Cue range (or range extension) is represented well by the work of Cutting

and Vishton [1995]. They explore the depth cue classification listed above and provide an evaluation of their relative strength as a function of viewing distance. Figure 2.1 provides an idealized representation of the just-discriminable depth thresholds as a function of the log distance from the observer. It is based on measurements of depth discrimination threshold functions for a variety of different depth cues. An important observation is that the distance of visual space has a significant influence on the relative importance of different depth cues. Occlusion is the strongest depth cue throughout the range of personal, action and vista space. The strength of the binocular disparity cue is greatest in a viewer's personal space and falls off in the action and vista spaces.

Cue dominance occurs when a specific depth cue has a more significant impact on depth perception [Howard, 2002]. The assumption is that the strongest depth cue overrules the depth interpretation from other cues. Occlusion is a good example because it can overrule depth ordering perceived by other cues. However, other depth cues may provide more precise position information.

Cue averaging is represented by a weighted combination of depth cues. An example of this approach is a weighted linear cue combination strategy combining the linear perspective and texture gradient cues [Oruç et al., 2003].

Cue specialization considers that the depth interpretation task can influence cue prioritization [Bradshaw et al., 2000; Schrater and Kersten, 2000]. For example, the occlusion depth cue can efficiently be used for depth ordering tasks and perspective projection is useful when parallel lines converge. Ware and Mitchell [2008] demonstrated how the depth relationship of complex 3D graphs can be better interpreted with a combination of stereo and motion cues. Such a task would be difficult with linear perspective (or relative size/density) alone.

Depth cues can also play complimentary roles compensating for limitations in our visual system. For example, blur and disparity cues co-vary according to geometric optics. The strength of blur cue increases at distances as the disparity cue becomes weaker [Mather and Smith, 2000]. Held et al. [2012] utilized a volumetric display to independently control disparity and blur depth cues. They found disparity to be a more precise depth cue near fixation and blur to be more precise in depths away from fixation. Additionally, they hypothesize that blur provides a stronger depth cue at visual eccentricities beyond central foveal fixation.

Hybrid approaches exist, exploring the combination of *cue averaging* and *cue specialization*. Cipiloglu et al. [2010] developed a weighted cue summation

model with linear cue prioritization using fuzzy logic. Their system takes as input the scene and task. It then prioritizes depth cues to include for efficient rendering. The focus is on reducing computational cost while preserving the most important visual information supporting depth interpretation tasks. This concept could be applied to help one parameterize depth cues and evaluate their relative strength within a scene (and task). For example, the strength of other depth cues may reduce the importance of exaggerating disparity within a scene.

2.2.3 Depth Cue Distortions

Conflicts between depth cues can influence interpretation in stereoscopic images. For example, it has been observed that decoupling eye vergence and accommodation, which is inherent in stereoscopic image viewing, can result in distorted perception of depth [Watt et al., 2005; Banks et al., 2008; Hoffman et al., 2008]. Occlusion can also play a significant role in depth interpretation, due to the loss of disparity information. Harris and Wilcox provide an overview of different types of occlusion and their influence on depth perception [2009]. Tsirlin et al. [2010] demonstrated how monocular occlusion clues alone can be used to infer location and direction of depth discontinuities and object boundaries in a scene.

When evaluating stereoscopic image content, it is often beneficial to consider which depth cues are visible and to verify that they agree and are appropriate for intended depth discrimination of the content.

2.3 Stereopsis

Binocular vision is the process of seeing the world with our two eyes. At a high level, there are two theories about how visual information is perceived between the two eyes to produce a single perception of the world [Steinman et al., 2000]. On one hand there is *binocular alternation or suppression* theory, in which perception of the world alternates between the eyes based on a process of *binocular rivaly*. The alternate theory is that the world is perceived through a *binocular fusion*, in which visual information from both eyes is fused to produce a single percept of the three-dimensional world. Not surprisingly, binocular vision utilizes both theories, although the fusion is more often used for natural viewing.

Stereopsis represents the perception of three-dimensional depth that comes about from fusing the different projection of the world on our two eyes. The resulting visual information, which can only be perceived when stereoscopically fused by both eyes, is called *cyclopean vision* [Steinman et al., 2000]. The Random dot stereogram (RDS) is an example of encoding visual information that cannot be perceived by monocular luminance or color changes [Julesz, 1960].

2.3.1 Disparity

The difference between corresponding points in the two eyes is termed *binocular disparity* or *retinal disparity*. Figure 2.2 provides an overview of several aspects of *horizontal disparity*, which are presented in this section. Based on the geometry of our eyes, horizontal disparity has a more significant effect than representing vertical disparity, and it is easier to visualize. Figure 2.2 visualizes the scenario where the eyes are converged on a fixation point, F. Point P represents a nearer point that is simultaneously fused while fixated on point F. The disparity, δ, between points F and P is represented by

$$\delta = \alpha_L - \alpha_R = \alpha_{FL} - \alpha_{FR} - (\alpha_{PL} - \alpha_{PR}) = \alpha_F - \alpha_P. \tag{2.1}$$

Disparity is often represented in units of degrees and is expressed as the angular difference in projection of corresponding points on each retina, as visualized in α_L and α_R in Figure 2.2. Visual processing of stereopsis can produce both absolute and relative depth perception.

Absolute and Relative Disparity represent two forms of depth interpretation and disparity processing in the visual system. Absolute disparity represents the interpretation of visual information at absolute depth from the eyes. Relative disparity represents depth relationships between visual information (e.g. the depth between objects). Interestingly, absolute disparity information is produced in early portions of the visual system (neurophysiological components), however, psychophysical reporting is often in terms of relative disparity [Cumming and DeAngelis, 2001]. This is an indication that both low- and high-level forms of visual processing work together.

In the visual processing pipeline, the primary visual cortex (known as V1) represents the first area of the human visual system where single neurons can be activated by stimulation from both eyes [Cumming and DeAngelis, 2001].

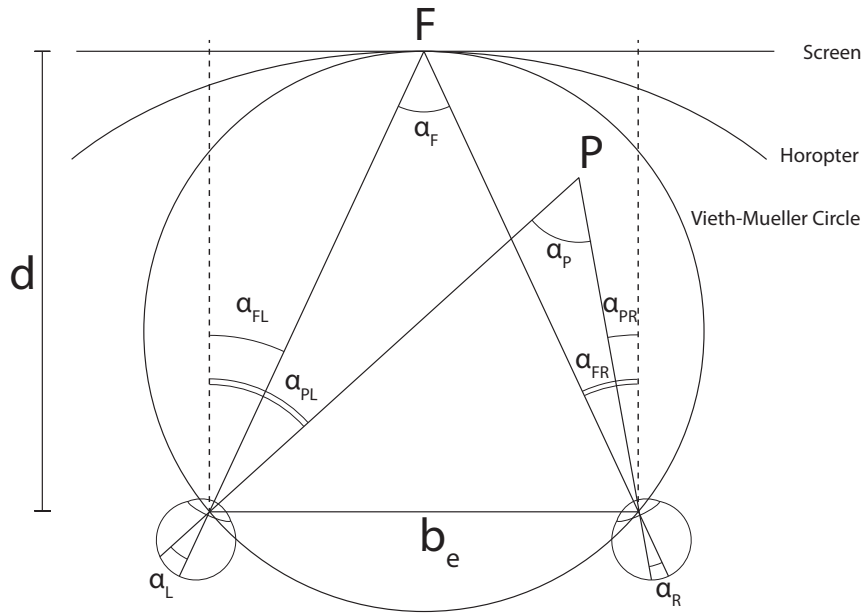


Figure 2.2: *Visualization of angular relationships to compute disparity. The eyes have a baseline, b_e, and are fixated at point F at distance d from the observer. The Vieth-Mueller Circle, Horopter and Screen plane are also represented. Note: angles α_{FL} and α_{PL}, for example, could also be defined relative to the horizontal axis passing through the baseline, b_e.*

Interestingly, neurons in V1 have been found to be tuned to *absolute disparity*, in which the disparity in retinal coordinates of visual information forms a percept of depth information in space. The difference in retinal disparity, determined by displacement from the left and right eye fovea, is influence by foveal fixation (eye vergence).

In Figure 2.2, the fixation point, F, is located at depth, d. The disparity of point P is $\delta_P = \alpha_F - \alpha_P$. The depth of P is then perceived at a corresponding Δd from the fixation depth, d. Production of absolute disparity information in early parts of the visual system (the first stages of visual processing) can be utilized for important visual tasks, such as guiding binocular eye convergence while viewing motion in depth. Absolute disparity can be considered as representing the depth range equation to compute the depth and direction of a point in space.

Relative disparity is the difference between two absolute disparities. Figure 2.4 provides an example. Given a fixation point, F, the disparity between the near and far points is $\delta_{NF} = \alpha_{PN} - \alpha_{PF}$. This disparity corresponds to a depth difference, Δd_{NF}, that is invariant to fixation changes. This representation offers the major advantage that depth relationships remain constant

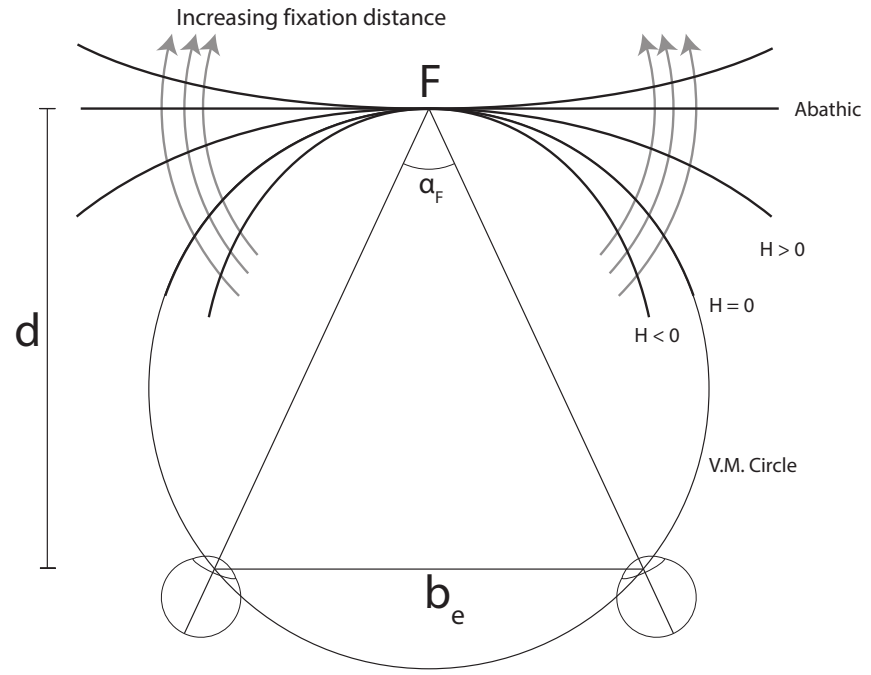


Figure 2.3: *Visualization of a horizontal slice of the horopter. Corresponding points lying on the horopter are perceived to have the same disparity. The Hering-Hillebrand deviation, H, represents how the empirically observed horopter deviates from the theoretical horopter, the Vieth-Mueller (V.M) circle.*

with changes in eye vergence. Visual processing mechanisms utilizing relative disparity can then maintain a stable depth representation of the visual space while viewing motion in depth or changing eye vergence through saccadic motion about a scene.

The Horopter represents all points in visual space whose projection on the retina appear in corresponding points on both eyes. The horopter represents a surface in the visual field [Schreiber et al., 2008]. For simplicity of visualization, Figure 2.3 represents a slice of corresponding points in the horizontal dimension. Given a fixation point (labeled "F" in figure), the curved lines represent example horopter shapes depending on the fixation distance. While fixating, a point on a given horopter is perceived by each retina (monocularly) to come from the same direction (and distance) [Ogle, 1932]. These points are perceived to have zero disparity.

The horopter was originally thought to correspond to the Vieth-Mueller circle, which is the circle passing through both of the eyes as well as the fixation point. Empirically, however, the horopter has been observed to de-

viate from the Vieth-Mueller circle by a fixation dependent value called the Hering-Hillebrand deviation (H) [A. Ames et al., 1932; Ogle, 1932]. H is influenced by the mapping of corresponding points as well as the magnification experienced between the two eyes.

$$H = cot(\alpha_R) - Lcot(\alpha_L) \tag{2.2}$$

The angles α_R and α_L represent the azimuth between the fixation point and another corresponding point for each eye. L is a skew factor related to the magnification of one eye relative to the other.

Curvature of the Vieth-Mueller circle is proportional to fixation distance. The Hering-Hillebrand deviation does not change with fixation distance (see Howard and Rogers for discussion of small variations [2002]). As a result, the difference in curvature between the horopter and Vieth-Mueller circle remains constant. The benefit is that the layout of corresponding retinal points remains constant with changes in eye vergence distance [Steinman et al., 2000].

In Figure 2.3, the horizontal line labeled, Abathic, represents the condition when the horopter actually is flat, forming a frontal parallel plane passing through the fixation point. The abathic distance occurs when

$$H = \frac{b_e}{d} \tag{2.3}$$

where b_e is the interpupillary distance and d is the fixation distance.

Disparity Channels

By studying anomalies in stereoscopic depth perception, Whitman Richards postulated the existence of three forms of disparity processing selective for zero disparity (e.g. corresponding points located on the horopter), crossed (near) disparity or uncrossed (far) disparity [1971]. Richards observed stereoanomolies in which viewers were stereo-blind for one of the three types, for example stereo-blind to only crossed disparities. This is an important observation in that it helps to explain why experimental subjects may have difficulty stereoscopically fusing specific disparity ranges.

Using experiments that induce adaptation effects on the processing of specific types of disparities, it has been possible to measure a tuning curve

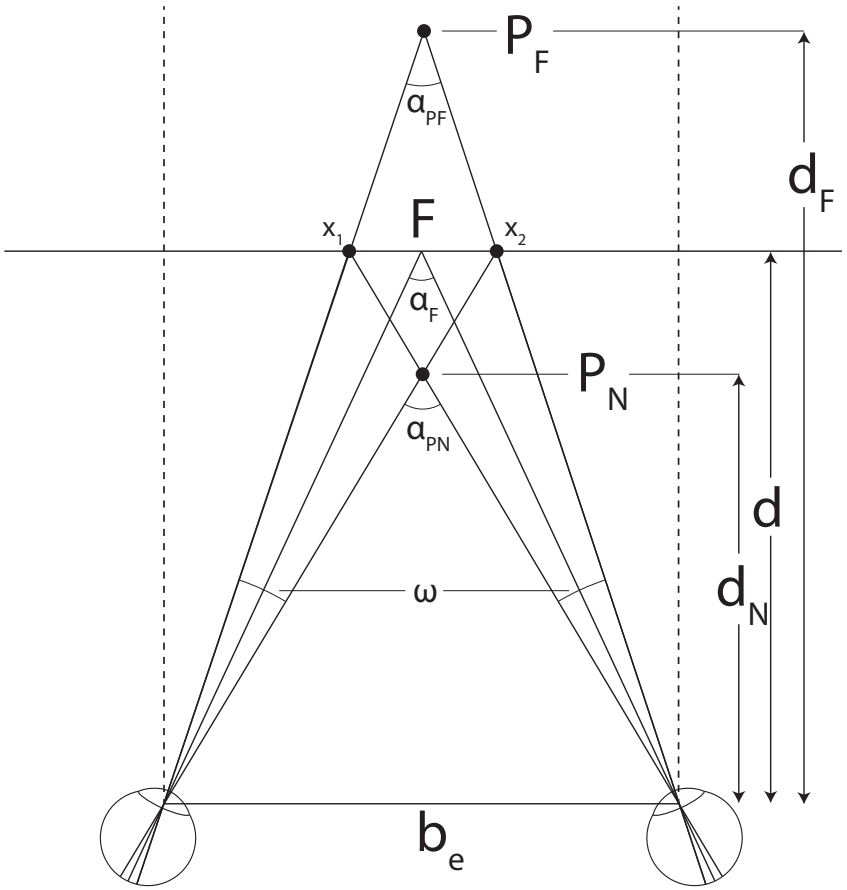


Figure 2.4: *Visualization of geometry to perceive a corresponding feature pair as nearer (P_N) or farther (P_F) than the fixation point (shown at screen plane). Angles, ω, are presented to facilitate presentation of stereoscopic acuity and stereoscopic resolving power in Section 2.3.2.*

for a specific disparity-tuned channel [Steinman et al., 2000]. These experiments have shown that disparity channels are broadly tuned. For example, Steinman [2000] noted that the crossed disparity channel has a maximum of 6 *arcmin* disparity and a half-height bandwidth of 10 *arcmin*, which is broad considering stereo thresholds are on the order of 10 *arcsec*. Recent experiments suggest that more than three channels may exist, but the precise number of disparity-tuned channels is not yet known [Steinman et al., 2000].

Depth Position

In addition to representing disparity as an angular measure, disparity in stereoscopic images are also represented in units of image pixels or millimeters of separation. Figure 2.4 visualizes an example of fixating on an image

plane while fusing points x_1 and x_2. Depth of the disparate point can be computed relative to the viewer:

$$d_v = \frac{db_e}{b_e - s} \tag{2.4}$$

where d is the viewing distance to the stereoscopic screen. s is the on screen separation or pixel disparity of corresponding points. b_e is the baseline eye separation. d_v represents the depth perceived relative to the viewer.

Depth can also be computed relative to screen given a viewing distance:

$$d_s = \frac{ds}{b_e - s} \tag{2.5}$$

A positive pixel disparity corresponds to perceiving points behind the screen. A negative pixel disparity corresponds to points perceived at depths nearer than the screen plane.

2.3.2 Lower Disparity Limit

The lower disparity limit represents the smallest amount of perceived disparity. This limit can also be called the stereoacuity, which is the disparity analog to luminance-based visual acuity.

Visual Acuity

Visual acuity is the resolving limit of spatial vision. The resolving power is the angle subtended by the detail at minimum acuity. Visual acuity is the reciprocal of the resolving power. There are many factors influencing visual acuity including optical properties of the eye, photoreceptor characteristics that collect visual information and receptive field sizes determined by the various processing centers of visual information. To a simple approximation, visual acuity for a bright, high luminance stimulus is approximately 30 *arcsec* in the foveal region and decreases in the periphery.

Vernier acuity measures the ability to align two line segments. It is a form of *hyperacuity*, which means the resolving power to align lines is higher at around 8 *arcsec* than that for visual acuity ($\sim$ 30 *arcsec*).

Stereoacuity

Stereoacuity is the resolving limit of stereopsis. It is the depth discrimination threshold when the only depth cue is binocular disparity. Measurement of stereo acuity is dependent on many factors including the testing method, stimuli characteristics and neurophysiological limitations of the viewer. Howard and Rogers [2002] provide a review. For example, using a Keystone stereo test, Coutant and Westheimer [1993] tested 188 students and observed that 97.3% of them had stereo acuity of 2.3 *arcmin* or smaller. They also observed that at least 80% to have a stereoacuity of 30 *arcsec* disparity. Howard and Rogers note that some observers can achieve a stereo acuity in the range of 2 to 6 *arcsec*. However for for clinical purposes, a stereoacuity better than 40 *arcsec* is an indication of "stereoefficiency" in adults [Howard and Rogers, 2002]. Stereoacuity thresholds are also a form of *hyperacuity*

Figure 2.4 represents example geometry used to compute an approximate stereoscopic resolving power and ultimately the intervals of stereoscopic depth discrimination. The structure of the following formulation is based on derivations presented by Valyus [1966].

Stereoacuity (ω) is the depth-descrimination threshold approximated by

$$\omega = \frac{b_e \Delta d}{d^2} \text{ in radians} \tag{2.6}$$

given the interpupillary distance (b_e), the depth difference (Δd) is the smallest detected depth range at a given distance d. Radians can be converted to seconds by multiplying by $\frac{180}{\pi}$. Stereoacuity is, to a first approximation, proportional to the distance between the eyes and inversely proportional to the square of the viewing distance.

Depth Intervals

The concept of *depth interval* or *depth steps* helps to visualize the range of disparities that are perceived to be the same. Or alternatively, given an object, the distance where a second object will be interpreted to be at a different depth (using only the binocular disparity cue). Although disparity limits are influenced by many factors, such as spatial frequency, it can be helpful to visualize depth intervals as an approximation with fixed stereoscopic acuity.

Equation 2.7 represents the *stereoscopic resolving power* given baseline, b_e, detectable threshold width (aka. *stereoacuity*), ω, and distance, d. For an average observer, let $b_e = 65\ mm$ and 30 *arcsec* threshold width be $w = 0.000145$ radians.

$$W(d) = \frac{b_e}{\omega d^2} \tag{2.7}$$

The *stereoscopic resolving power* units are the inverse of the depth range (e.g. $1/meter$). This reciprocal of the discriminable depth is called the *stereodiopter* or *diopter*.

The *stereoscopic resolving power using a binocular device*, $W'(r)$, proportionally increases the optical magnification, G (decreases the threshold angle), and base magnification, B, given by equation 2.8. Let the product of the optical and base magnifications be represented by π.

$$W'(d) = \frac{B}{(\frac{\omega}{G})d^2} = \frac{\pi b_e}{\omega d^2} \tag{2.8}$$

The quantity, π, is a dimensionless coefficient that represents the depth-sensitivity of a display device. It indicates the magnifying power of an optical system.

The *amount of stereoscopic information*, as shown in equation 2.9, represents the number of discriminable depth planes within a region of space, defined by the nearest, d_1, and farthest, d_2, distances. Remove the d_2 term if the farthest distance is at infinity.

$$|N|_{d_1}^{d_2} = \frac{\pi b_e}{\omega}\left(\frac{1}{d_1} - \frac{1}{d_2}\right) \tag{2.9}$$

The result approximates the number of unique depth planes that can be perceived between two depths.

Stereoacuity at Projection

For stereoscopic projection, the viewing distance and image pixel size influence the minimum stereoacuity threshold. In equation 2.10, Δs represents the threshold detectable width at screen distance d_s.

$$\Delta s = \omega d_s \tag{2.10}$$

If Δs is less than the image pixel image separation, we should modify ω to account for the limit due to screen resolution as opposed to human limit. The *stereoscopic resolving power* and *depth intervals* would be computed with the greater ω.

Stereoacuity and Contrast

In the previous sections, stereoacuity was assumed to be constant. However, there are many additional factors influencing stereoacuity. For example, Cormack et al. [1991] explored the thresholds for interocular correlation. They observed luminance contrast to have a significant influence. At low contrast, stereo acuity was inversely proportional to to the square of contrast. At higher contrast over a range of approximately one log unit, a cube root law contrast dependence was observed.

Stereoscopic Contrast Sensitivity Function

Stereoacuity is also influenced by contrast sensitivity. Frisby and Mayhew's demonstrated a correlation between stereopsis sensitivity and contrast detection sensitivity as a function of spatial frequency [1978]. Their findings show that the shape of contrast detection and stereopsis are similar, although with a shift representing a decreased stereoacuity. These results agree with other findings, for example, Filippini & Banks [2009] and Tyler [1975] observations of the influence of disparity amplitude and spatial frequency on disparity sensitivity. The CSF correlation with stereopsis has lead to models of perceived depth of frequency and magnitude changes in disparity [Didyk et al., 2011].

2.3.3 Upper Disparity Limit

The upper disparity limit represents the maximum range of disparities that can be simultaneously fused. It is best visualized through four concepts: Panum's Fusional Area, Disparity Gradients, Diplopic Depth Perception and the Divergence Limits of Stereopsis.

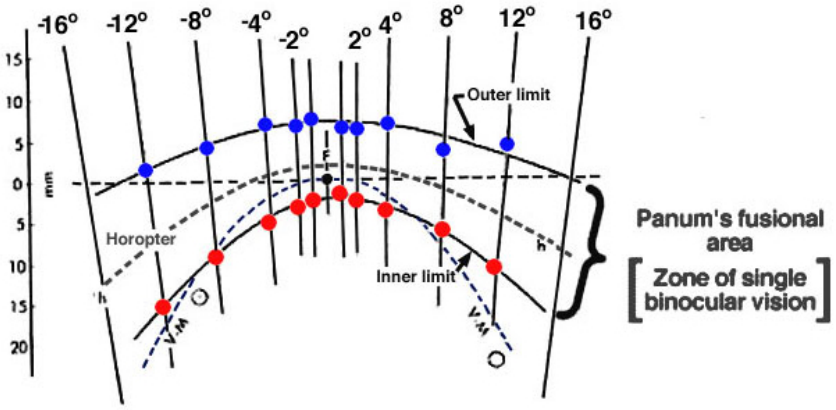


Figure 2.5: *Panum's fusional lies between the inner and outer limits of single binocular vision. This figure represents observations for a stimuli distance of 40 cm. The figure is reproduced with permission of Webvision [Kalloniatis and Luu, 2013].*

Panum's Fusional Area

Panum's Fusional Area (aka Panum's Zone) represents the region of points (see Figure 2.5) that can be simultaneously fused while fixated on a point in space. The area of fusion in the *visual field* is a function of our *receptive field* size and the mapping of disparity-tuned cells in the primary visual cortex. The receptive field size is smallest in *central (foveal) vision* and it increases in the periphery of our *visual field*. Panum's Fusional Area, as represented in Figure 2.5 has an inner limit of maximal crossed (near) disparities and outer limit for maximum uncrossed disparities. See Ogle [1932; 1950] for observational data of Panum's Fusional Area.

The region of simultaneous fusion ranges from roughly ± 10 *arcmin* up to ± 1 degree, depending on retinal eccentricity [Steinman et al., 2000]. It is helpful to differentiate the central region of stereopsis, which is a narrow ± 0.5 degree range at the fovea, from peripheral region of stereopsis, which operates up to ± 7 to ± 10 degrees [Steinman et al., 2000]. The foveal region is selective for small disparities while the periphery is selective for large disparity ranges.

As with the lower disparity limits, it is important to note that Panum's Fusional Area does not represent strict fusion limits. The limits are dependent on other factors such as contrast magnitude and frequency of both luminance and disparity information. As such Panum's Fusional Area can vary when comparing fusion limits observed by fusion of RDS versus line patterns. Furthermore, Panum's Fusional Area has been found to be an estimate of the limits for achieving initial fusion. Fender and Julesz [1967] observed that af-

ter fusion of a 6 *arcmin* is achieved, the disparity can be slowly increased up to 2 degrees horizontal disparity and still maintain fusion. If the disparity increase is too fast or exceeds 2 degrees, fusion is lost.

Disparity Gradient

The disparity limit for fusion described by Panum's Fusional Area can be reduced due to the proximity of corresponding features. Burt and Julesz [1980] found the gradient of the disparity rather than the magnitude of the disparity to be the limiting factor for fusion when objects are near each other in the visual field. They defined the *disparity gradient* between nearby objects to be the difference in their disparities divided by their separation in visual angle. Figure 2.6 provides an example of two nearby points. The binocular disparity difference is the difference between the individual dot disparities, $d_b = d_1 - d_2 = R_r cos\Theta_r - R_l cos\Theta_l$. The disparity gradient is the ratio of the binocular disparity to the binocular dot separation, d_b / R_b. Equation 2.11 represents a critical dot separation, $\hat{R}_b$, marking the boundary between fusion and diplopia.

$$\hat{R}_b = kd_b \tag{2.11}$$

Burt and Julesz experimentally found that fusion is not obtained when the disparity gradient is greater than 1° of disparity per degree of dot separation [1980]. They found that fusion can be lost at less than a third of the value reported by Ogle for the width of Panum's fusional area.

Diplopic Depth Perception

Diplopia occurs beyond Panum's fusional area. Ogle [1950] noted that just beyond Panum's zone there is a range of disparities that, although diplopic, can still produce strong impressions of depth. Ogle called region of strong depth impression from either fusion or diplopia, "*patent stereopsis*" [Howard, 2002]. Beyond this region there is a region of vague depth interpretation, which he called *qualitative stereopsis*. Qualitative stereopsis has also been called *latent stereopsis*. In the foveal area, Ogle observed the fusional area to be ± 5 *arcmin* while the patent stereosis extended to ± 10 *arcmin* and qualitative stereopsis to ± 15 *arcmin*. At 6 degree eccentricity, patent stereopsis increased to 70 *arcmin* and qualitative stereopsis to about 2 degree [Howard, 2002].

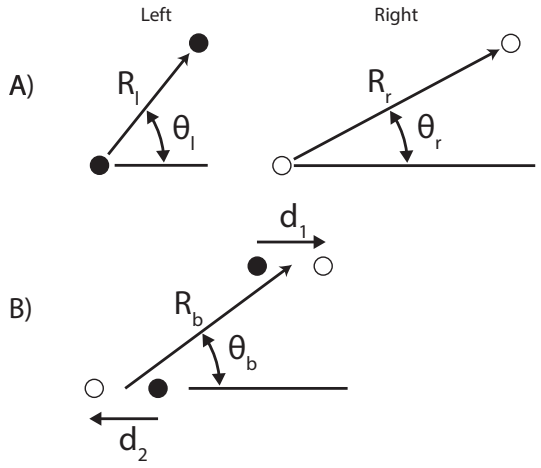


Figure 2.6: *Geometry for computing disparity gradient using a two-dot stereogram. (A) The images shown to each eye and (B) the disparities required for stereoscopic fusion. There is no vertical disparity, so $R_l sin\theta_l = R_r sin\theta_r$. The right eye dots are unfilled to visualize the example. Normally, the left and right eye dots would both be filled. Figure notation from Burt and Julesz [1980].*

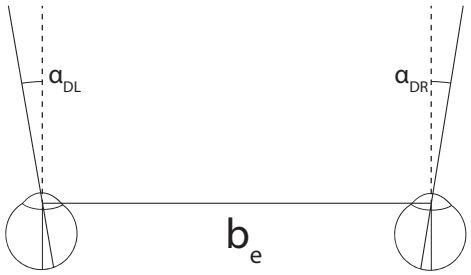


Figure 2.7: *Visualization of divergent disparity. Divergent disparities up to 1.75 degrees can be fused.*

Divergence Limit

Although counter intuitive, it is possible to fuse stereoscopic image disparities that exceed the eye separation. The limits of stereoscopic fusion are a function of receptive field size [Howard, 2002]. Given an eye fixation, there is a range of uncrossed disparities that can be fused. Jin et al. [2005] experimentally found that divergent angular disparities up to 1.75 degrees can be fused. They saw this as a neurophysiological limit across viewing distance. This observation agrees with common production practices as stated by Mendiburu [2009] and Lipton [1982]. Lipton, for example, found 1 degree divergence to be acceptable.

Figure 2.7 provides an example visualization of divergent disparity, where the angle of divergent disparity, δ_{div}, is represented as

$$\delta_{div} = \alpha_{DL} - \alpha_{DR} \tag{2.12}$$

where α_{DL} and α_{DR} are the divergent angles of each eye from parallel viewing. Divergent disparities can be computed as divergent on-screen separation, s_{div},

$$s_{div} = s - b_e \tag{2.13}$$

where s is the on screen separation or pixel disparity. The divergent angular disparity is then

$$\delta_{div} \approx \arctan \frac{s_{div}}{d} \tag{2.14}$$

where d is the viewing distance to the screen. It is important to note that children, for example, will reach the divergence limit earlier due to the smaller eye separation. Divergent viewing ability is especially useful in large cinema viewing environments.

2.3.4 Spatiotemporal Interactions

In the luminance domain, it has been observed that both spatial frequency and temporal frequency influence the detection of visual luminance information. The concept, *window of visibility*, is used to represent the influence motion and time have on the detection of luminance information [Watson et al., 1986]. For example, fine details are best viewed with static images. Sensitivity to fine details decreases with an increase in temporal variation.

The upper and lower disparity limits of stereopsis have also been shown to be influenced by spatio-temporal interactions. Kane et al. [2014] observed that spatial frequency and temporal frequency of disparity modulation influences the detection thresholds of the disparity. They observed that spatial and temporal influence are separable. They fit their observations to a windowed, cross-correlation energy model of disparity, and observed that both the upper and lower disparity limits are influenced by the same spatial and

temporal windowing function. They observed a best mean fit with a spatial windowed Gaussian, $\sigma = 12\ arcmin$, and a temporal windowed Gaussian, $\tau = 24ms$.

Temporal Disparity Gradient

Kane et al [2014] also observed that the disparity gradient concepts extends to the temporal domain. They experimentally observed the expected spatial temporal gradient

$$\nabla d_s = 2af_s \tag{2.15}$$

Where a is the peak-to-trough disparity amplitude and f_s is spatial frequency. They observed a spatial upper disparity limit when the disparity change, a, was greater than 1.5 deg for a 1.0 deg change in spatial position. They observed $\nabla d_s = 1.5$, which is consistent with previous findings [Burt and Julesz, 1980].

They also observed the following temporal disparity-gradient limit:

$$\nabla d_t = 2af_t = 0.7 \tag{2.16}$$

where f_t is temporal frequency and ∇d_t is in units of arcmin/msec.

They then combined the two gradient limits to produce a spatio-temporal gradient, $2af_{st}$, where:

$$f_{st} = \sqrt{f_s^2 + (kf_t)^2} \tag{2.17}$$

k represents a constant such that the two gradients are equivalent. They found $k = 0.09$ minimizes the differences between data points. The resulting spatio-temporal disparity gradient, $\nabla d_{st} \approx 1.5$ for nearly all combinations of spatial and temporal frequency.

Through the experimental design, Kane et al. [2014] were able to demonstrate that the minimum disparity thresholds are separable in terms of disparity variations of spatial and temporal frequencies. Their results demonstrate how spatio-temporal disparity sensitivity is more restricted than spatio-temporal luminance sensitivity.

2.3.5 Modeling Stereopsis

There are many challenges in modeling stereopsis. One challenge is to understand how the functional neural wiring in the visual system responds to binocular stimuli. Another challenge is solving the correspondence problem, which requires strategies to balance both locally and globally optimal solutions to disparity estimation. Finally, computational methods for applying the model must be implemented.

In the 1990s, significant progress was made to experimentally model the functional neural wiring of how V1 complex cells respond to binocular stimuli, such as RDS patterns [Ohzawa, 1998; Fleet et al., 1996]. *Disparity energy models* were developed based on experimental data to model the behavior of binocular energy neurons [Ohzawa et al., 1990]. Ohzawa's model [1998], for example, represents two stages: First, an array of binocular simple cells each produces a half-squaring nonlinearity, achieved by rectification and squaring of the cellular response. The second stage is to sum the output of each binocular simple cell by a binocular complex cell. The result of which produces a disparity tuning curve representing the specific disparity sensitivity of a given binocular complex cell's receptive field. This is a local representation based on the RF size.

The *correspondence problem* represents the challenge of selecting the correct feature matches between the two eyes. Selecting incorrect matches can result in a false binocular interpretation of depth. Although disparity energy models such as Ohzawa's are primarily local, they are able to reduce the complexity of the correspondence problem. Since binocular complex cells can be tuned to different spatial frequencies, the result is that disparity tuning can be achieved at several band-pass frequency scales. Combination of the frequency scales can help to find a more global solution, or at least help to significantly reduce the complexity to find a globally optimal solution. The limitations on finding a global solution are based the receptive field sizes used in the summation process. Fleet et al. [1996] demonstrated how disparity energy models based on position-shift and/or phase-shift can produce highly accurate disparity estimation with the right pooling strategy. They also demonstrate improved performance when linearly pooling over spatial scales in addition to orientations and local spatial neighborhoods.

Disparity energy models are computationally similar to interocular cross-correllation [Ohzawa, 1998; Fleet et al., 1996]. However, there is an important difference. Fleet et al.[1996] point out that disparity energy models modulate a complex cell response about a baseline input, which is the sum of input monocular energies. Cross-correlation does not predict a stimulus dependent

baseline. Initial binocular interaction is multiplicative for cross-correlation and additive for energy models [Fleet et al., 1996].

Cross-correlation is successfully applied to agree with physiological observations of disparity sensitivity. Banks et al. [2004] demonstrated how well a cross-correlation algorithm with adaptive window size can model performance of the human visual system. They observed algorithms to perform well on image content that is frontoparallel surface having a disparity gradient of zero, and with an appropriately sized spatial window for computing the cross correlation [Banks et al., 2004]. Their findings agree with physiological observations that disparity-selective neurons are limited by their receptive field size [Nienborg et al., 2004]. Importantly, Nienborg et al. [2004] observed that V1 receptive fields prefer uniform disparity, which helps to explain why disparity sensitivity is highest for zero disparity gradients. This is also reasoned to be a cause of low spatial stereo resolution [Banks et al., 2004; Filippini and Banks, 2009]. Higher order neurons could be constructed to be selective for a specific magnitude or direction of disparity gradient, however, they would not have a higher stereo resolution than observed in V1 [Filippini and Banks, 2009].

Motivated by the well known contrast sensitivity function, there has also been effort to define a *disparity sensitivity function* [Bradshaw and Rogers, 1999]. This has motivated the application of experimental methodologies used to observe luminance thresholds to detect disparity sensitivity thresholds. Didyk et al. [2011] modeled the influence of disparity amplitude and disparity frequency. They later model the influence of luminance magnitude and luminance frequency on the disparity sensitivity thresholds [Didyk et al., 2012]. These results agee with earlier findings, for example Filippini & Banks [2009] and Tyler [1975] observations of the influence of disparity amplitude and spatial frequency on disparity sensitivity.

This section has detailed several important aspects of stereopsis. The concept of disparity was defined. Upper and lower limits of disparity sensitivity were also presented. The following sections build on stereopsis in ways influencing visual comfort, attention and detection of additional visual distortions.

2.4 Vergence, Accommodation and Comfort

When viewing objects in the real world, eye vergence and accommodation is harmoniously coupled in our visual system. Stereoscopic viewing disturbs this relationship. The visual system must maintain focus on the screen (where

the information is located) and change vergence to fixate on the scene depth. This has been found to influence depth interpretation, fatigue and discomfort [Hoffman et al., 2008; Shibata et al., 2011].

This section introduces eye vergence, accommodation, and the coupling between them. The impact of stereoscopic image viewing is then presented.

2.4.1 Vergence

Eye vergence is the movement of both eyes to converge the central fovea of each eye on a common point in space. This provides the highest visual acuity of the eyes on that point in space. Vergence eye movement serves three basic functions: stabilization of the retinal image as the head moves, fixation and tracking of objects, and convergence of visual axes on a particular object [Howard, 2002]. There are three forms of vergence eye movement: horizontal, vertical and rotational (aka. ocular torsion or cycloversion). We present vergence in the context of horizontal eye movement since it is most important for stereopsis.

The *vergence angle* is expressed using geometry formed by the eyes fixating on a point in space. Figure 2.2 presents the simple geometry representing the horizontal angle of vergence, α_F.

$$\tan(\alpha_F/2) = b_e/2d \tag{2.18}$$

Equation 2.18 represents this basic geometric with b_e as eye separation and d as distance to the fixation point from midpoint between the two eyes.

In Howard's [2002] review of eye vergence topics, he presents the following four types of horizontal vergence:

- Tonic vergence - the default eye vergence state when no visual information is stimulating vergence
- Proximal vergence - image cues can influence perception of where an object is in space
- Accommodative vergence - a change in eye accommodation is normally associated with a change in vergence
- Fusional, or disparity-induced vergence - eye movement driven by absolute disparity and the maximization of correspondence between the two eye images.

As with all neurophysiological processes, eye vergence is not exactly on the intended point in space. For stereopsis, the error need only be within the range of Panum's Fusional Area to achieve proper fusion. Kenneth Ogle [1950] called this occurance *fixation disparity*.

2.4.2 Accommodation

Accommodation is the process of adjusting refractive power of the eye lens to bring the intended image of objects into focus. Refractive power is represented in units of *diopter*. It represents the inverse distance (in units of meters) of the intended visual stimulus in visual space. 1 diopter (aka. $1D$) corresponds to an object that is one meter away. A refractive power of $2D$ is required to accommodate an object that is 0.5 meters away. The refractive power of the lens can vary by approximately 10 diopters. However, this range reduces with age to approximately 1 diopter at age 70 [Howard, 2002].

Similar to vergence, there are four types of accommodation [Howard, 2002]:

- Tonic accommodation - the default, or resting state, of eye accommodation state when no visual information is stimulating accommodation
- Proximal accommodation - image cues can influence perception of where an object is in space
- Blur accommodation - perceived blur of the retinal image
- Convergence accommodation - a change in eye vergence is normally associated with a change in accommodation.

Fortunately, as with vergence, there is also an an acceptable error in the amount of accommodation to perceive an object as clear. This range is called the *Depth of Focus*.

2.4.3 Depth of Focus

Depth of Focus is the range of distances in image space (projected on the retina) that appear in focus. It is a symmetric value, relative to the image sensing plane, which is the *retina* for human vision. The value is represented in diopters. *Depth of Field* is the projection of depth of focus into object space. Depth of Field is the range of distances in object space that appear in focus. This value is not symmetric about the fixation point and often represented in meters.

In a similar way that Panum's Fusional Area provides a neurophysiological tolerance for vergence, depth of focus provides a tolerance for accommodation. As long as the intended object is within the bounds of depth of focus, the object will be perceived to be in focus. It is important to distinguish between *retinal defocus* and *blur* [Wang and Ciuffreda, 2006]. Blur is the perceptual sensation of a decrease in sharpness. Retinal defocus is an optical phenomenon, which results in a smaller retinal image contrast gradient [Wang and Ciuffreda, 2006]. Retinal defocus is acceptable as long as it is within the threshold for detection of blur.

Depth of focus can be measured objectively, however, it is more common to measure it subjectively. Vasudevan et al. [2007] observed a mean objective depth of focus of $\pm 0.59 \pm 0.10D$ with a range of $\pm 0.46D$ to $\pm 0.75D$. Through subjective measures, they observed a larger mean depth of focus of $\pm 0.63 \pm 0.22D$, with a range from $\pm 0.37D$ to $\pm 0.96D$. Wang et al. [2006] reviewed experimental findings of depth of focus and found a large variation in foveal depth of focus from $0.04D$ to $3.50D$. They state that the large variance is due to different experimental stimuli and methodology. However, they stated the typical depth of focus in young experienced observers is approximately $0.8D$ to $1.2D$.

Wang et al. [2006] summarized internal and external factors influencing depth of focus. External factors include attributes of the observed visual information (e.g., luminance, contrast, spatial frequency, wavelength). Generally, for the external factors, a decrease in detectability of the external factor helps to increase the subjective depth of focus. Internal factors refer to optical and neurological attributes of the viewer (e.g., visual acuity, pupil size, age, retinal eccentricity, refractive state). A complete model of depth of focus requires consideration of these factors.

For example, considering internal factors, pupil diameter influences depth of focus much like an aperture of a traditional camera. Ogle and Schwartz [1959] observed a 0.12 diopter reduction in depth of focus per millimeter of increase in pupil size. Ogle and Schwartz also found stimulus size to influence the depth of focus. They observed an increase of $0.3D$ to $0.4D$ per 0.25 *arcmin* increase in stimulus target size.

2.4.4 Coupling of Vergence and Accommodation

In natural image viewing, eye vergence, accommodation and pupil diameter work together to form a clear image. The interrelated change is known as a *near-triad* response [Howard, 2002]. For example, when looking near,

the pupil diameter reduces to increase depth of field and decrease spherical aberration. When looking far, the pupil dilates (increases diameter) to improve retinal illumination and decrease diffraction. The primary benefit of the coupling is an improved visual performance reducing the amount of time to transition visual attention.

Accommodative convergence (AC) occurs when a change in accommodation induces a change in eye vergence. An increase in accommodation will converge the eyes while a decrease in accommodation will diverge the eyes. The *AC/A ratio* represents the amplitude of accommodative convergence (AC) induced by a 1 diopter change of accommodation (A).

Convergence accommodation (CA) occurs when a change in eye vergence induces a change in accommodation. The *CA/C ratio* represents the amplitude of convergence accommodation (CA) induced per unit change in convergence [Howard, 2002].

There is a significant amount of research about the linkage between AC and CA. Ian Howard [2002] provides an informative overview and references to more detailed reviews. The following two important topics arise in the context of accommodation and convergence: *phoria* and *the zone of clear single binocular vision*.

Phoria represents the tendency of the visual system to return to its natural resting state, the tonic points of accommodation and/or vergence. Figure 2.8 represents the effect of this phenomena. The dashed-diagonal line in the left panel represents the ideal correspondence between stimulus induced eye vergence and accommodation (note: both represented in diopter for visualization). The green line represents the influence of Phoria while the other two lines represent the bounds. In typical vision, a viewer under converges for far focal distances and over-converges for near distances. In both case, the deviation of convergence is in the direction of the tonic point. The right panel of Figure 2.8 shows the typical viewing distances for difference stereoscopic displays.

Zone of Clear Single Binocular Vision (ZCSBV) is the range of eye vergence that can be observed clearly for a given fixed focus on a stimuli (see Figure 2.9). The shape of the ZCSBV is roughly parallel to the Phoria.

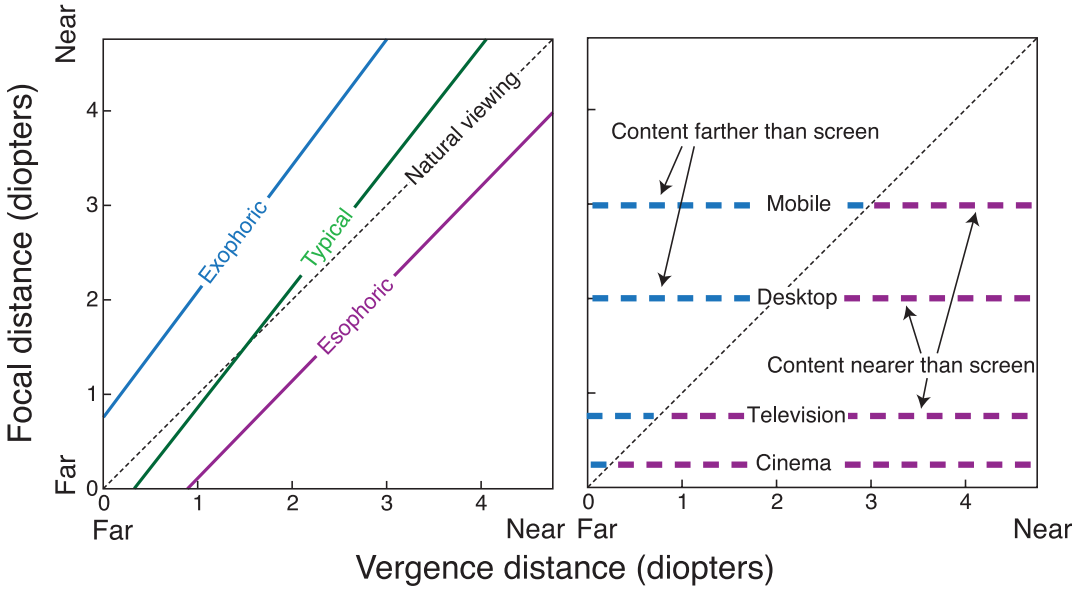


Figure 2.8: *Visualization of natural viewing, phoria, and typical display viewing distances visualized by Shibata et al [2011]. Reproduced with permission from the authors* [1].

2.4.5 Influence of Stereoscopic Imaging

The right panel of Figure 2.8 represents a fundamental problem of stereoscopic image viewing. While each stereo display modality is capable of presenting some range of vergence disparity, each viewing scenario is constrained to a fixed focal distance. This represents the well-known *decoupling of vergence and accommodation*.

Many researchers have stated the visual conflict between vergence and accommodation influences visual comfort, depth interpretation or fatigue [Emoto et al., 2005; IJsselsteijn et al., 2005; Lambooij et al., 2009; Patterson, 2007; Ukai and Howarth, 2008; Yano et al., 2004]. Some have experimentally observed an effect of discomfort or fatigue by comparing stereoscopic image viewing to 2D image viewing [Emoto et al., 2005; Kuze and Ukai, 2008; Yano et al., 2002]. However, those findings do not prove the discomfort is caused specifically by the vergence accommodation conflict. Kooi and Toet [2004], for example, explored a variety of additional perceptual distortions caused by stereoscopic viewing that can cause visual discomfort.

[1] Journal of vision by Association for Research in Vision and Ophthalmology Reproduced with permission of ASSOCIATION FOR RESEARCH IN VISION AND OPHTHALMOLOGY in the format Republish in a thesis/dissertation via Copyright Clearance Center.

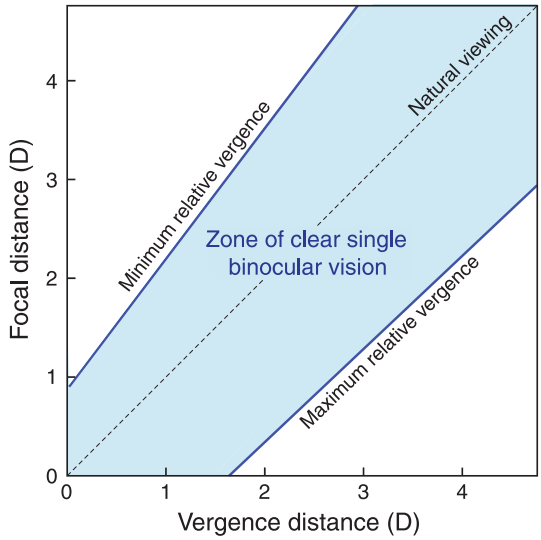


Figure 2.9: *Zone of clear single binocular vision estimated by Shibata et al [2011]. Reproduced with permission from the authors* [1].

Hoffman et al. [2008] were the first to convincingly prove an effect of the vergence-accommodation conflict on visual discomfort, depth interpretation and fatigue. This was achieved through the development of a volumetric stereoscopic display, which enables the control of both vergence and (approximate) accommodation cues [Akeley et al., 2004]. Accommodation was interpolated between several fixed states. This experimental system makes it possible to isolate and control the degree of vergence and accommodation conflict.

The results of Hoffman et al. [2008] were compelling, but experimental observations were limited to a specific viewing distance (39 cm or 2.5D). This research was extended to predict the zone of comfort of a stereoscopic display [Shibata et al., 2011]. They conducted a detailed exploration in three parts: First, demonstrating the influence of viewing distance on discomfort and fatigue. Second, exploring the influence of disparity sign (in front or behind screen) on discomfort and fatigue. Third, measuring the phoria and zone of clear single binocular vision, which are predictors of the discomfort observed in the first two experiments.

Comfortable Depth Ranges

Figure 2.10 represents the comparison of Shibata et al.'s [2011] observed Zone of Comfort to previous, well known predictions. The Zone of Comfort defined by Sheard and Percival are both relative to the zone of clear single

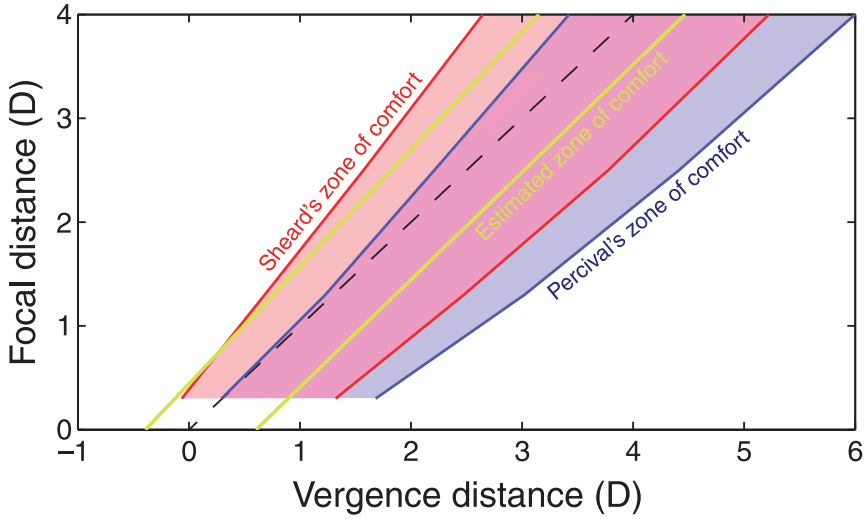


Figure 2.10: *Comparison of the Zone of Comfort as defined by Sheard, Percival, and Shibata et al. [2011]. Reproduced with permission from the authors* [1].

binocular vision (ZCSBV). *Perceival's zone of comfort* is the central 1/3 region of the ZCSBV. *Sheard's zone of comfort* is centered on the phoria and extends 1/3 the distance toward each bound of the ZCSBV. Perceival and Sheard definitions overlap in the purple region. The yellow lines represent the boundaries of the zone of comfort observed by Shibata et al [2011]. The far boundary (left of dashed line representing natural viewing) is represented by Equation 2.19.

$$D_f = m_{near} D_v + T_{near} \tag{2.19}$$

where D_f is focal distance in diopters, D_v is vergence distance in diopters. They found $m_{near} = 1.035$ and $T_{near} = -0.626$.

$$D_f = m_{far} D_v + T_{far} \tag{2.20}$$

Equation 2.20 is defined by $m_{far} = 1.129$ and $T_{far} = 0.442$.

The results of Shibata et al. are also represented in terms of angular disparity in Figure 2.11. The near limit of comfortable stereoscopic viewing increases with viewing distance while the far limit decreases with viewing distance. The discontinuity in the far boundary representation is a correction to prevent eye divergence at larger viewing distances (such as for cinematic viewing).

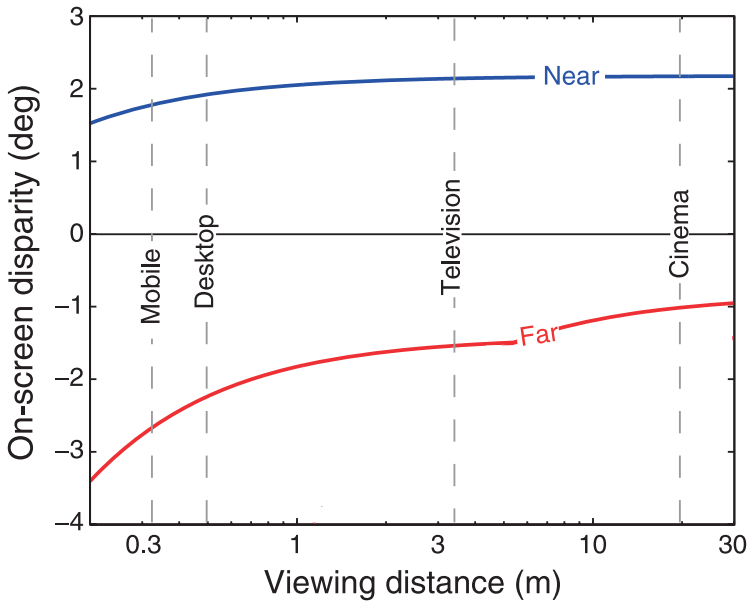


Figure 2.11: *Zone of comfort estimated by Shibata et al. [2011]. Far and near boundaries represented as angular disparity for a given viewing distances. Reproduced with permission from the authors [1].*

Rules of Thumb

Stereoscopic cinematography has evolved with empirical data providing various rules-of-thumb. Film-based stereographers assumed that there should be a maximum on-film deviation. Ferwerda, for example, recommended a maximum deviation of 1.2 mm when using a 35mm film format [2003]. Other cinematographers applied a percentage rule, such as the 3% rule, which corresponds to the maximum allowable pixel or screen disparity relative to the screen width. This rule corresponds to the notion that the baseline separation of a stereoscopic camera system should not exceed 1/30th of the camera distance to the nearest object. For a camera baseline of 6.5 cm, the camera should be approximately 2 meters from the nearest subject. Shooting a subject that is 0.5 meters away would require a baseline of 1.7 cm. This rule is designed to maintain an acceptable amount of depth from the nearest element to infinity.

$$b_c = a_n/30 \tag{2.21}$$

The 3% rule is also sometimes called the 2% or even 1% rule, depending on how conservative the stereographer wants to be. Shibata et al. [2011] demonstrated that the 2% rule is often too conservative compared to their experimental observations of the comfort zone.

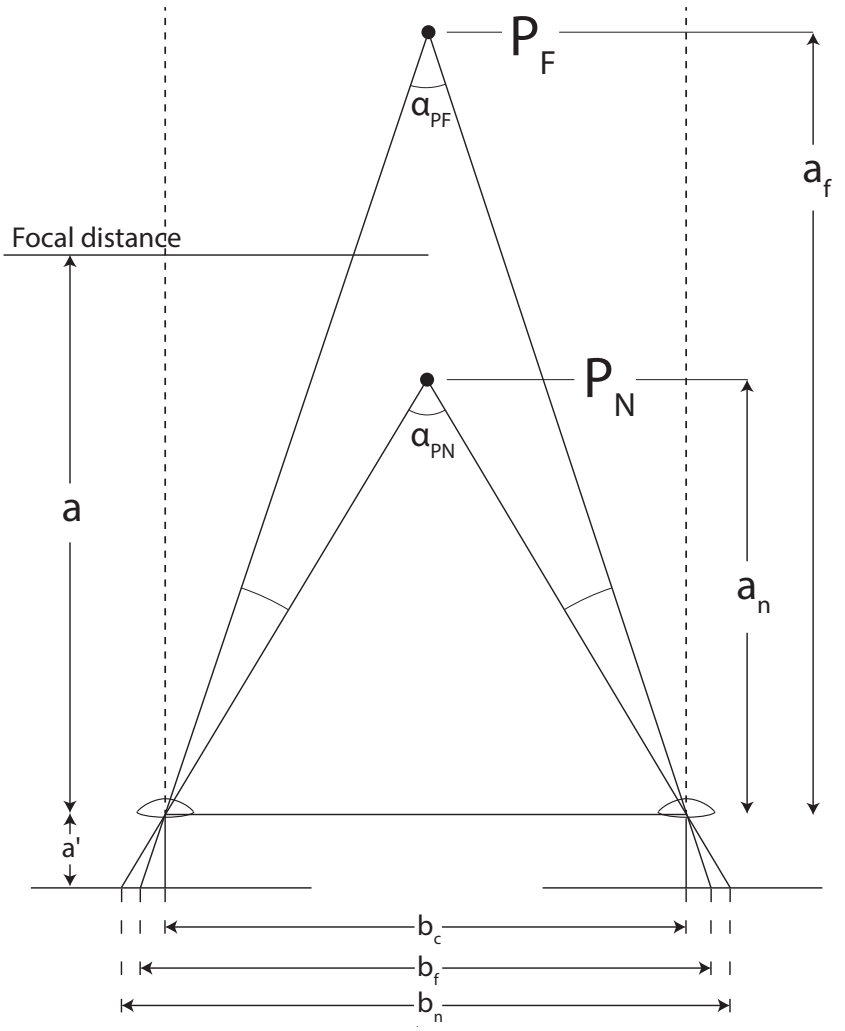


Figure 2.12: *The general geometry of a stereo camera with a focus distance, near object, far object, baseline and projection on an image sensor.*

Depth Range Equations

When capturing (or rendering) stereoscopic content, it can be very useful to quickly compute the appropriate baseline to achieve maximum and minimum comfortable depth ranges. Bercovitz [1998] formulated the stereo camera base equation with a maximum deviation in mind. Figure 2.12 represents the general geometry of the stereo camera. Equation 2.22 computes the baseline as a function of nearest object, a_n, farthest object, a_f, focus distance, a, focal length, f, and maximum deviation, d. The maximum deviation is the difference between the far and near point image separations projected on the sensor, $d = b_f - b_n$.

$$b_0 = d\frac{a_f a_n}{a_f - a_n}(\frac{1}{f} - \frac{1}{a}) \tag{2.22}$$

Selection of the appropriate maximum deviation can then be used to improve the stereoscopic camera parameters.

2.5 Distortions

Stereoscopic imaging has much potential to provide immersive visual experiences. However, it also has much potential to induce perceptual distortions, which cause visual discomfort, depth misinterpretation, and fatigue. An undesirable outcome is for the distortions caused by stereoscopic image viewing to distract or hinder the presentation of story or exploration of the immersive world. There is a fundamental need to characterize the space of visual distortions and provide the technology to resolve them.

The challenge is to understand the perceptual limits as well as cost/benefits of including potential distortions. This requires perceptually-based computational models to predict the quality of visual experience. This section identifies several perceptual distortions. As reference, Kooi et al [2004] conducted a systematic evaluation of 35 different stereoscopic distortions and evaluated them in terms of the discomfort experienced by the viewer.

2.5.1 Depth Distortions

Depth distortions could be considered as the catch-all of stereoscopic distortions. Just as 2D image viewing has the potential to invoke the perception of geometric spatial distortions, the use of stereoscopic 3D imagery only increases the potential for conflict. Geometric distortions come about from a mismatch between stereoscopic capture, display and viewing conditions. These geometric relationships have been well studied [Woods et al., 1993; Jones et al., 2001; Masaoka et al., 2006; Yamanoue et al., 2006; Zilly et al., 2011]. These approaches primarily rely on geometric relationships, ignoring limitations inherent within the human visual system. Depth distortions also require the ability to isolate and evaluate specific factors to make meaningful conclusions. Below, we present several specific stereoscopic distortions, many of which also fall within the category of depth distortion.

2.5.2 Stereoscopic Image Scale

The stereoscopic depth cue can influence the sense of scale and thus influence the perception of space and sense of immersion. In 2D image viewing, *orthoscopic projection* is achieved through alignment of camera perspective to the viewer perspective. *Orthostereoscopic projection* is the 3D equivalent where the baseline also corresponds to the baseline of the viewer. Magnification and the correspondence between captured and observed scene geometries also have an influence.

Magnification and Orthoscopic projection The value of orthoscopy is task dependent. If it were always required, we would need to change our viewing distance depending on the lens used to capture an image. We have learned how to interpret images captured with different projections/focal lengths.

Ortho-stereoscopic Perspective The ortho-stereoscopic perspective is obtained when the field of view, viewing/capture distance, and baseline of the scene capture system and observer match. The stereo camera baseline should be approximately 6.5 cm. The field of view of the projected image should match the camera system.

Hyper- and hypo-stereo A viewer interprets a stereoscopic image as if it were captured with the same baseline as their vision (approximately 6.5 cm). When the baseline of stereo capture is increased (hyperstereo), the convergence angle about the scene objects is higher. A higher convergence angle corresponds a closer object as shown in Figure 2.13. Since the size of the object has not changed, the object is interpreted to be smaller. Hypostereo occurs when the baseline of capture is reduced. Objects are interpreted to be larger.

Interestingly, the phenomenon of scale perception influences viewing the real 3D world. Children and women see the world as larger than men [Ramadan, 2009]. This is due to the average eye separation distance of men (6.6 cm), women (5.7 cm), and children (4.5 cm). For example, women view men 20% larger than men view themselves.

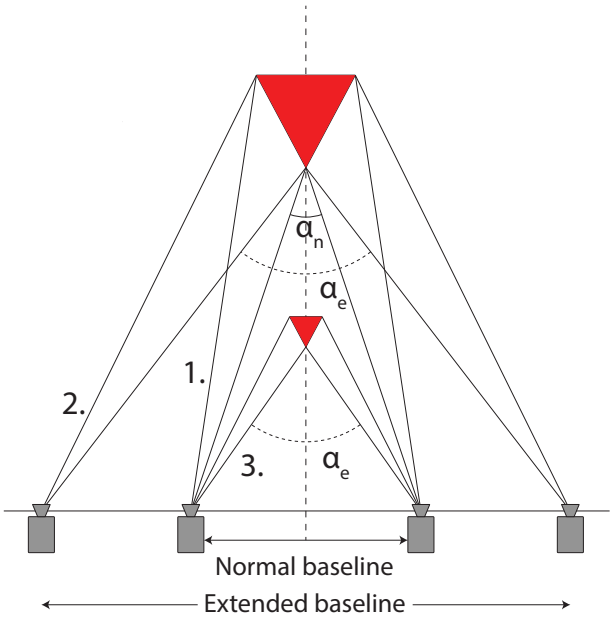


Figure 2.13: *Effect of baseline extension (camera capture separation) on perceived depth and size of an object. (1.) Represents a normal capture at a small convergence angle α_n. (2.) When the baseline is extended, the convergence angle is higher, α_e. (3.) The brain processes the image on the basis of the normal eye baseline. Higher convergence corresponds to a closer position object (magenta triangle). As a result, the object appears to be closer, but it's visible size does not scale as expected. The effect is that the object appears smaller.*

2.5.3 Stereoscopic Cardboarding

Stereoscopic cardboarding is the perception of objects to be flatter than they would be expected to appear. One factor influencing the perception of cardboarding is the mismatch between perception of object size and object disparity with distance. Howard and Rogers [2002] point out that size sensitivity is inversely proportional to distance, while disparity sensitivity is inversely proportional to the squared distance. This results in a conflict between size and depth scaling.

Another significant factor that influences cardboarding is a geometric mismatch between the stereoscopic capture, display and viewing conditions. Masaoka et al. [2006] sought to develop a spatial distortion prediction system to determine the extent of the stereoscopic cardboarding effect. However, they developed geometric relations without taking the subjective perception of the artifact into account.

Yamanoue et al [2000] experimentally evaluated perceived cardboarding by exploring several factors including lighting and variation of spatial thickness. They observed a significant effect of spatial thickness in the subjective rating of perceived cardboarding. Only one object with three spatial thickness values was evaluated, thereby making it difficult to draw more general conclusions regarding finer scale changes in spatial thickness. Yamanoue et al. [2006] later geometrical modeled the cardboarding effect as the ratio of size and depth magnification. They observed a good correlation with their previous experimental observations from one object [2000].

2.5.4 Ghosting

The perception of stereoscopic crosstalk is often called ghosting. Ghosting can be considered a“binocular noise" that further hinders fusion limits and visual comfort. Yeh and Silverstein [1990] demonstrated that crosstalk significantly influences the ability to fuse widely separated images via binocular eye vergence movement. Ghost images may introduce unintended edges and binocular rivalry making visual processing unstable, unpredictable, and impair guiding visual attention [Patterson, 2007]. It has also been found to inhibit the interpretation of depth [Tsirlin et al., 2011a,b].

Use of even minimal crosstalk has been found to strongly affect subjective ratings of display image quality and visual comfort [Yeh and Silverstein, 1990; Kooi and Toet, 2004]. Although acceptable crosstalk may generally be as high as 5-10%, the detection and acceptability thresholds can be significantly reduced with higher image contrast or larger disparity [Wang et al., 2011a]. There is a significant need to remove the detection of crosstalk.

This need has motivated a variety of ghosting removal methods (aka. deghosting). Typically, these methods rely on some form of subtractive compensation [Konrad et al., 2000]. A perceptually motivated extensions to traditional subtractive compensation was presented [Smit et al., 2007], which utilizes a perceptually uniform CIE-Lab colorspace for subtractive compensation. However, these methods fail when negative light is required to achieve sufficient subtractive compensation. We developed a perceptually-based distribution of the ghosting signal to reduced sensitivity regions of the human visual system [van Baar et al., 2011]. We demonstrated that our compensation method produces more comfortable stereoscopic images as compared to traditional subtractive compensation methods.

2.5.5 Vergence-Accommodation Conflict

As presented earlier in Section 2.4, the vergence-accommodation conflict has been experimentally observed to induce visual discomfort, depth misinterpretation and fatigue. This problematic situation has motivated our exploration in Chapter 3 of using scene composition to compensate for visual attention transitions.

2.5.6 Microstereopsis

Motivated by the vergence-accommodation conflict, there are many efforts to reduce the range of disparities to be within the zone of comfortable viewing, as discussed in Section 2.4. One extreme solution is to apply minimal scene disparities. Siegel and Nagata [2000] proposed the concept of *microstereopsis*, in which small interocular separation is combined with alignment of interesting content about the zero parallax plane. Their informal experiments demonstrated sensitivity to small disparities and they hypothesize that minimal detectable disparity is sufficient when combined with other visual cues to stereoscopic depth. Didyk et al. [2011; 2012] formulated perceptually-based depth discrimination thresholds and also demonstrated an application of minimal stereopsis.

2.5.7 Disparity Remapping

Another approach to position stereoscopic content within the comfort zone is to apply disparity remapping operators. A global approach to disparity remapping is to linearly adjust the disparities, effectively by adjusting the camera baseline and, perhaps, reconverging the zero parallax plane (via camera convergence or horizontal image translation). The previously mentioned geometric relationships represent this case [Woods et al., 1993; Jones et al., 2001; Masaoka et al., 2006; Yamanoue et al., 2006; Zilly et al., 2011]. Holliman et al. [2004] proposed piecewise linear remapping to alter the remapping operator based on regions of interest.

Our own disparity remapping work was among the first to implement a framework based on a set of basic disparity remapping operators than can produce local and global disparity remapping [Lang et al., 2010]. We also demonstrated how our algorithms, combined with image warping techniques and sparse disparity information, can be used to benefit a variety of practical stereoscopic disparity editing applications. Later work, such as

Basha et al. [2011], extended seam carving to retargeting of stereoscopic images while maintaining geometric consistency to minimize both image distortion and depth distortion. There have been more recent works combining image retargeting and disparity remapping [Chang et al., 2011; Qi and Ho, 2013]. Perceptually-based remapping operators have been proposed by Didyk et al. [2011; 2012]. More recently we have motivated remapping based on the constraints introduced by autostereoscopic, multi-view displays [Chapiro et al., 2014].

2.5.8 Inconsistent Depth Cues

As discussed in Section 2.2, depth interpretation can be hindered by inconsistent depth cues. The vergence-accommodation conflict is one example, which motivates our attention transitions research presented in the following chapter. Another example is the *Stereoscopic Window Violation*, which is presented in Chapter 4.

2.6 Attention and Saliency

Visual attention is an important survival skill. It enables the focus of limited visual processing resources on interpreting important parts of the real 3D world. The seminal work of Koch and Ullman [1985] proposed a model for pre-attentive visual attention. They identified two stages of visual attention. First, a "preattentive" mode in which simple visual features are processed in parallel over the entire visual field. The second, "attentive" mode represents the process of focusing visual attention. The attentive stage is believed to be a serial process utilizing many visual cues and volitional factors to maintain visual attention.

Koch and Ullman [1985] proposed that the pre-attentive vision model should be composed of parallel processing of simple visual features including color, orientation, movement and disparity in a winner-take-all network. This theoretical framework was later implemented by Itti and Koch [1998] who computed center-surround differences of pre-attentive features. Their system consisting of stimuli-driven, *bottom-up* mechanisms accurately described how attention is deployed within the first few hundreds of milliseconds after the presentation of a new scene. Preattentive temporal dynamics were additionally modeled by allowing the maximal conspicuity feature to decay in the

winner-take-all process. They further acknowledge that modeling the attentive stage requires more sophisticated models with top-down mechanisms accounting for volitional biasing.

Feature contrast is the primary concept supporting saliency models. Itti et al. [1998] utilized center-surround feature structure, which provided a local contrast measure for a given receptive field. Through a global normalization, they were able to reduce the influence of strong contrast that were common while amplifying the strong contrasts that were unique. This process produced conspicuity maps per feature, which were then linearly combined with other feature maps before applying a winner-take-all prediction of preattentive saliency.

There are other saliency approaches, which may be less biologically inspired, however they often use some form of contrast metric. Frequency space methods [Guo et al., 2008; Hou and Zhang, 2007] determine saliency in the frequency domain, evaluating the amplitude or phase spectrum. These methods tend to label object boundaries as salient. Colorspace methods can be global or local. Local colorspace methods can evaluate pixel dissimilarity [Ma and Zhang, 2003] and multi-scale bandpass frequency representations [Itti and Baldi, 2005]. These methods can also emphasize edges and noise. Global colorspace methods include estimation of contrast between image patches [Goferman et al., 2010; Liu et al., 2011; Wang et al., 2011b]. These methods can identify larger image structures as salient, but at the cost of high computational complexity. Dimension reduction can be applied [Duan et al., 2011] with potential loss of important information. Similar to Itti and Koch, another colorspace method applied a simple difference from the mean to provide more global information [Achanta et al., 2009].

Other saliency methods aim to produce binary saliency labels to entire objects or image regions. They utilize image segmentation techniques [Ren et al., 2010] and also clustering [Cheng et al., 2011] to identify individual objects. Perazzi et al. [2012] decomposed their saliency into two color contrast components: uniqueness and distribution of color information. Their approach is able to efficiently label an object as salient. These methods perform well on saliency ground-truth data sets, which have a single labeled salient region. However, they perform poorly when scenes are cluttered. This leads to the importance of identifying the type of content to be analyzed and the preferred saliency representation. For example, the object versus feature region saliency labels.

Motion Saliency

Itti and Dhavale [2003] proposed a complete spatiotemporal framework by extending earlier work on image saliency [Itti et al., 1998] with additional center-surround mechanisms for flicker and motion. More recent methods have been designed to only compute motion saliency [Cui et al., 2009; Belardinelli et al., 2009] or spatiotemporal saliency [Rapantzikos et al., 2009; Mahadevan and Vasconcelos, 2010]. Appropriate motion contrast mechanisms are also a topic of active research.

Stereoscopic Saliency

There has been some research in stereoscopic saliency. Two saliency methods apply depth weighting as a post process after computing monocular spatiotemporal saliency for each eye [Jeong et al., 2008; Fernandez-Caballero et al., 2008]. For example, Jeong et al [2008] developed a biologically inspired saliency framework utilizing similar features to Itti and Koch. They utilize fuzzy logic to support feature combination. They compute saliency for each eye independently and then weight the saliency maps with disparity information, preferring objects that pop out. Both of these saliency methods, however, lack subjective evaluation and the application scenarios enabled by the saliency representation is not clear.

Niu et al. [2012] developed a more recent stereoscopic saliency method for still images. They proposed a disparity contrast metric and combined it with domain knowledge about the preference for specific ranges of disparities. Their disparity metric extends the histogram based color contrast method of Cheng et al [2011]. They segment the image and then identify regions with more abrupt disparity changes as more salient. They evaluate their results relative to other 2D methods on a stereoscopic dataset. However, the dataset also lacks ground truth provided by eye tracking.

There are many open questions related to stereoscopic saliency. First, there is very limited available eye track data on stereoscopic data sets. Second, there has yet to be a comprehensive study specifically exploring the role that stereopsis, comfort or other perceptual distortions play on visual attention in stereoscopic imaging. Finally, there is need for compelling applications to guide the appropriate saliency representation.

2.7 Conclusion

This chapter on modeling topics in stereoscopic imaging has provided a structured overview of many relevant topics, significant research and computational models. Details of these models are provided to describe how they are implemented to support computational analysis of stereoscopic perception. These models have been applied in various forms during the thesis, for example, guiding the capture and display of stereoscopic content for experiments as well as the production of stereoscopic movies. The following two chapters explore influence of different perceptual distortions caused by inconsistent depth cues. The next chapter explores how a conflict between vergence and accommodation can hinder the time to change visual attention within a stereoscopic scene. The subsequent chapter explore how a conflict between the depth cues occlusion and stereopsis influence visual quality and comfort.

CHAPTER 3

Attention Transitions in Stereoscopic Depth

This chapter represents exploration of how changes to stereoscopic scene composition can compensate for a common perceptual distortion in stereoscopic images, the decoupling of vergence and accommodation.

3.1 Introduction

Viewing stereoscopic 3D is inherently an unnatural experience. It has been shown that the decoupling of eye vergence and accommodation experienced during stereoscopic image viewing can lead to depth misinterpretation, discomfort and fatigue [Hoffman et al., 2008]. We explore the impact stereoscopic image viewing can have on the ability to change visual attention in depth. Our hypothesis is that scene composition can help compensate for the eye-vergence and accommodation conflict and facilitate the viewing of stereoscopic content.

To explore this question we attempt to mimic a ubiquitous cinematic scene setting: the basic dialog shot (aka. two-shot) in which viewer attention transitions between two actors [Mascelli, 1965]. Figure 3.1 provides several visu-

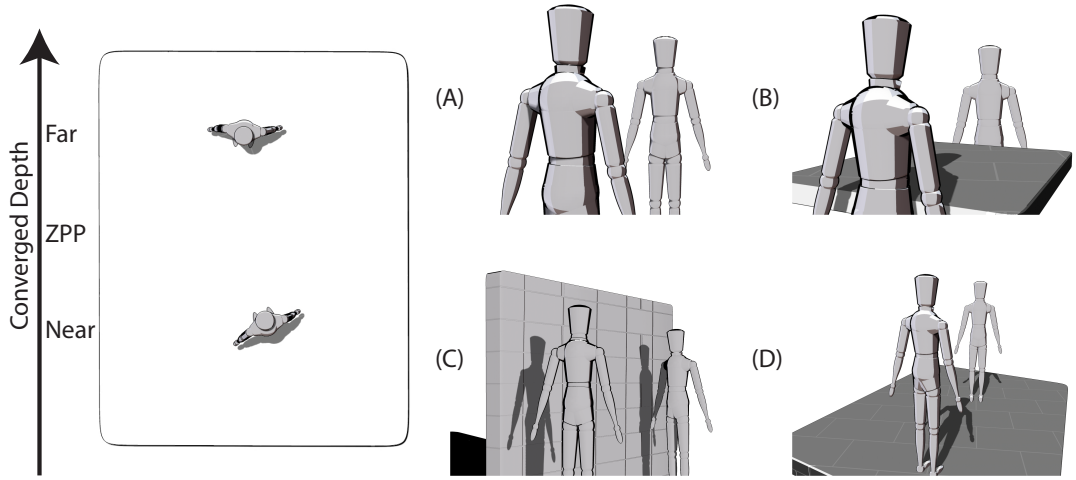


Figure 3.1: *Visualization of different forms of the two-shot. Actor positions are fixed in depth. (A) Represents an over-the-shoulder shot without depth continuity. Insets (B-D) represent potential ways to provide depth continuity.*

alizations of a two-shot (see insets A-D) with stereoscopic scene depth represented on the left. Figure 3.1-(A) represents an over-the-shoulder shot without depth continuity. The remaining three insets (B-D) provide depth continuity through changes to the camera placement and scene composition. We introduce a variable that corresponds to the difference in scene composition between a mid-level shot (inset A) and a down-shot that provides continuity between two actors (insets B-D). In these examples, continuity is provided by a table, wall or floor element that span the depth range between the two actors. We hypothesize that these continuous visual depth cues visually link the two actors and provide an intermediate element that the viewer can use to smoothly saccade from one actor to another.

Our experiment also explores an additional question: can fatigue be measured using a quantitative technique, such as measuring the time required to change visual attention from one actor to another. One would presume that as fatigue grows, the subject will tire and the speed with which they can modify vergence/accommodation will decline, thus providing an indirect measure of fatigue. We administered questionnaires in order to compare self-assessed measure of fatigue against our quantitative approach.

Contributions. Our work makes the following contributions:

- Demonstrate that a continuous depth element can reduce the time required to change visual attention in a stereoscopic scene.

- Present the correlation of data between performance and subjective measures of visual fatigue.
- Motivate that stereoscopic 3D content creators may learn scene composition, framing and montage from visual psychophysics.

Chapter Organization. Related work is presented in Section 3.2 followed by the experimental design in Section 3.3. Results, including our experimental setup and a user evaluation, are presented in Section 3.4. The chapter is concluded in Section 3.5.

3.2 Related Work

Binocular spatial perception is among the most demanding and energy consuming visual tasks viewers perform [Parker, 2007]. In natural image viewing, eye vergence, accommodation and pupil diameter work together to form a clear image. The interrelated change is known as a near-triad response [Howard, 2002]. The primary benefit of the coupling is an improved visual performance reducing the amount of time to transition visual attention.

Stereoscopic image viewing disrupts natural viewing behavior due to the decoupling of eye vergence and accommodation. Many researchers have stated the visual conflict between vergence and accommodation influences visual comfort, depth interpretation or fatigue [Emoto et al., 2005; IJsselsteijn et al., 2005; Lambooij et al., 2009; Patterson, 2007; Ukai and Howarth, 2008; Yano et al., 2004]. Some have experimentally observed an effect of discomfort or fatigue by comparing stereoscopic image viewing to 2D image viewing [Emoto et al., 2005; Kuze and Ukai, 2008; Yano et al., 2002]. However, those findings do not prove the discomfort is caused specifically by the vergence accommodation conflict. Kooi and Toet [2004], for example, explored a variety of additional perceptual distortions caused by stereoscopic viewing that can cause visual discomfort.

Hoffman et al. [2008] were the first to convincingly demonstrate an effect of the vergence-accommodation conflict on visual discomfort, depth interpretation and fatigue. This was achieved through the development of a volumetric stereoscopic display, which enables the control of both vergence and (approximate) accommodation cues [Akeley et al., 2004]. Accommodation was interpolated between several fixed states. This experimental system makes it possible to isolate and control the degree of vergence and accommodation

conflict. The research was continued by Shibata et al. [2011] who demonstrated the influence of viewing distance and disparity sign (e.g. in front or behind the screen) on discomfort and fatigue.

There are several approaches to compensate for the vergence-accommodation conflict. One approach involves presenting microstereopsis, which is a minimally required disparity [Siegel and Nagata, 2000; Didyk et al., 2011, 2012]. Another approach is to apply linear and nonlinear disparity remapping operators to recompose the scene depth and better utilize the limited depth budget [Lang et al., 2010]. Others have sought to ease attention transitions by aligning the depth position of visually salient scene elements between cuts [Koppal et al., 2011]. Our aim is to demonstrate that it is possible to maintain a large depth volume and utilize other visual cues to improve the time to change visual attention.

3.3 Methods

We use a restricted cinematic domain, similar to a dialog (aka. two-shot) between two spatially separated actors, to motivate visual attention transitions. We do so because the two-shot is one of the most widely used cinematic shots. We measure the response time necessary to change attention between the two scene elements. To ensure visual attention at the appropriate depth, we used random dot stereogram targets [Julesz, 1960] to represent the two actors. An example is provided in Figure 3.3. We asked the viewer to determine if the center portion of the target emerges or recedes from the target background. This is a task that requires binocular fusion to discriminate between these two conditions.

Subjects Twelve adult volunteers (8 male, 4 female, ages 21-37), with normal vision (corrected and uncorrected) participated in the study. An evaluation of the subjects spatial perception was assessed before the test was administered requiring consecutively correct evaluation of the RDS test target at increasing disparities until the subject demonstrated their ability to correctly perceive these targets at the disparities used in the test.

Stimuli The stimulus consists of two modified random dot stereograms (RDS) presented at 3 possible stereoscopic depths. Because some participants had trouble viewing binary random dot stereograms, a modified RDS was used. This phenomena was noted by Julesz [1960]. Edge detection and

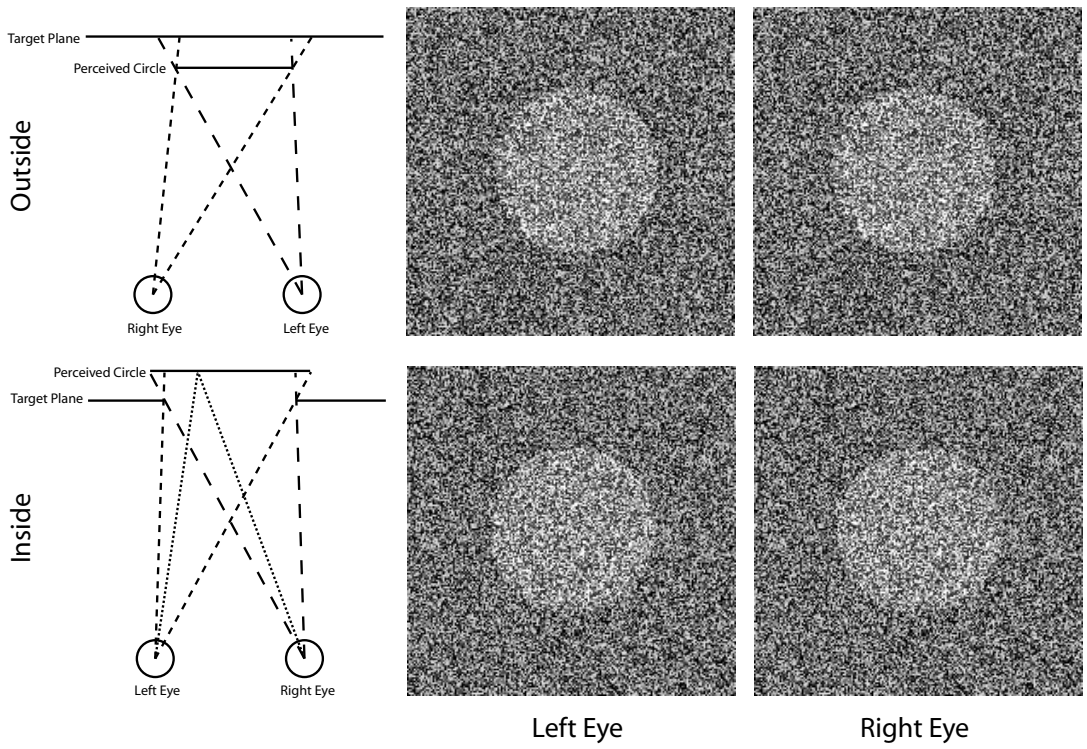


Figure 3.2: *Modified random dot stereogram (RDS) target stimuli. The circle appeared either in front of the square (Outside) or within the square (Inside). A modified RDS was used to make it easier to fuse the circle.*

matching is a critical component of disparity tuned cortical vision to produce stereoscopic fusion. In order to keep our potential subject pool as large as possible, we encoded the stereogram with 8 shades of grey (see Figure 3.2). The darker seven shades are used to encode the dots on the target plane and the lighter seven are used to encode the shape portion of the RDS. Because this shape can be perceived monoscopically, subjects do not identify the encoded shape, but rather interpret the depth location of the shape relative to the target plane (e.g. emerging or receding). The shape is encoded with a 12 pixel disparity and the entire stereogram is 193 pixels square, corresponding to an approximate angular width of 5.6 degrees.

The targets are placed on a floor plane, which is textured with a checkerboard pattern for the continuous depth trials (50%) and not visible (no texture) for the others (see Figure 3.3). When visible, the continuous depth floor plane extended from the nearest target location to the farthest target. Targets are mounted on either the left or right side of the plane at three depth locations. The depth locations were chosen to test three disparities (-1.72, 0, and +1.69 degrees), corresponding to on-screen disparities of approximately -59, 0 and 58 pixels. The relative disparity for the far target was set so as not to ex-

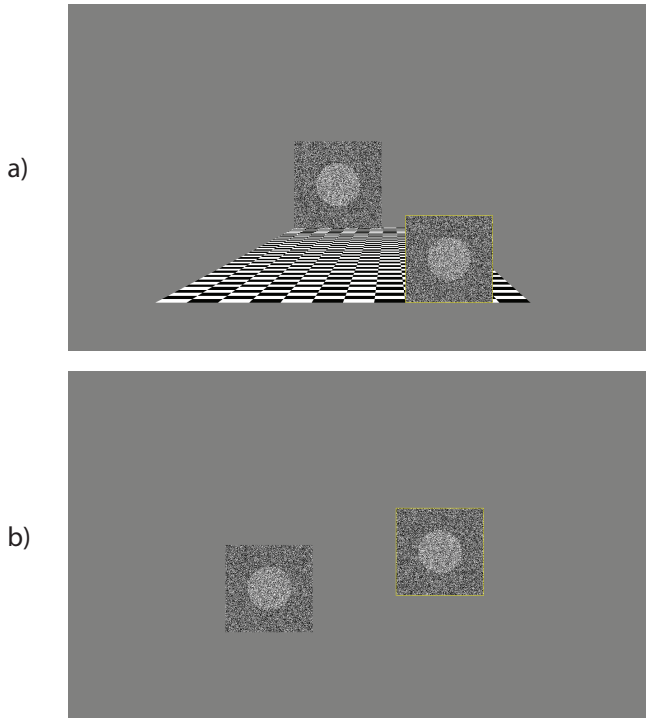


Figure 3.3: *Example trials in monoscopic view. (a) represents a trial with continuous depth cue. Previous target is located in the far depth and the current target (with yellow border) is in the near depth location. (b) represents a trail without continuous depth cue. Previous target is at near depth and the current target (with yellow border) is at the zero parallax depth location.*

ceed average inter-pupillary distance and was reduced to avoid divergent viewing. The near target was set to provide spatial symmetry across the zero parallax plane.

A 50% grey background was used for two purposes. First, a grey background hid some of the crosstalk that is present in our circularly polarized projection system. This crosstalk can be distracting for some viewers, especially when presenting high contrast, black and white images. Second, the use of 50% grey helped to balance the influence of our continuous depth plane. The plane is a black and white checkerboard pattern, which has a local luminance that alternates between black and white and averages globally to 50% grey.

3.3.1 Procedure

Participants were tested individually in the same, darkened experimental room. They were seated at a distance of 2 meters from the projected image.

The image size was 130cm x 78 cm (36 degree FOV) and the resolution was 1280x768 pixels. Circular polarizing filters were used to separate the left and right image channels. Participants wore circularly polarized glasses throughout the experiment to view the stereoscopic images and questionnaires. Image brightness on the projectors was reduced to simulate the illumination levels in a dark theater. Approximately 15 cd/m^2 was measured to be passing through the polarizing glasses at a distance of 2 meters. Less light results in a larger pupil size and decreases the depth of focus. User input was provided by a standard computer keyboard, which was placed on the lap of the experiment participant.

Pre-test Participants were first tested to determine if they could perceive the range of depths presented in the study. They were presented near and far targets (one at a time) and asked to respond to the stimulus in the target. Participants were instructed to press specific arrow keys depending on whether they perceived the shape encoded in the RDS to be in front of or behind the RDS target plane. Correct answers resulted in targets that were presented farther from the zero parallax plane. The test ended when the participant demonstrated perception of the depths used in the study. Participants who could not view the required depths were not permitted to continue with the experiment. A short break was given between the pretest and the experiment.

Questionnaire A 23 question survey was given at the beginning, middle and end of the experiment. All questions are summarized in Table 3.1. The survey first asks a general question about eye fatigue. The next 6 questions originated from a German survey, Kurzfragebogen zur aktuelen Beanspruchen (KAB), which was designed to provide a short scale for assessing stress [Müller and Basler, 1993]. The remaining 16 questions are standardized questions from the Simulator Sickness Questionnaire, which was originally designed to assess motion sickness in virtual reality simulators [Kennedy et al., 1993]. Although we lack motion in our experiment, the oculomotor factors of the SSQ are relevant to the viewing of stereoscopic images. The SSQ enabled us to quickly apply a standard survey that is relevant to stereoscopic eye fatigue. Questions were presented one at a time on the same image screen as the stereoscopic test so that the viewing condition did not change. Using a keyboard, the subject selected the desired response to a question and then confirmed the answer before proceeding.

Experiment Design The experiment is a three factor design. One factor is the *depth change* with 9 levels of change in depth between the two targets

Source	Question	Response (Integer or discrete choice)
Ours	How strong is fatigue of your eyes at the moment?	1 ("not at all") - 6 ("very strong")
KAB	At the moment I feel	1 ("tense") - 6 ("relaxed") 1 ("relaxed") - 6 ("queasy") 1 ("worried") - 6 ("untroubled") 1 ("calm") - 6 ("nervous") 1 ("skeptical") - 6 ("trustful") 1 ("comfortable") - 6 ("miserable")
SSQ	General discomfort Fatigue Headache Eyestrain Difficulty focusing Increased salivation Sweating Nausea Difficulty concentrating Fullness of head Blurred vision Dizzy (eyes open) Dizzy (eyes closed) Vertigo Stomach awareness Burping	"none", "slight", "moderate", "severe"

Table 3.1: *Summary of 23 question survey. The first question is our own. The next six are from the Kurzfragebogen zur aktuelen Beanspruchen (KAB) [Müller and Basler, 1993]. The last 16 are from the Simulator Sickness Questionnaire [Kennedy et al., 1993]. The questionnaire was integrated directly in the experiment using the same display and keyboard.*

(see Figure 3.4-left and also footnote for depth change codes [1]). The second factor is the presence of the *continuous depth* cue (continuous depth or non-continuous depth), as represented by the checkboard plane. The third factor is the experiment *block* depicted in Figure 3.5. Blocks are composed of 2 sub-blocks, one for continuous depth trials and one for non-continuous depth trials. The sub-block order is constant per participant and is balanced between participants.

An RDS target is presented in one of 6 locations (3 depths and 2 positions per depth as presented in Figure 3.4-left). The target to be assessed is outlined with a yellow border. The participant is instructed to decide if the shape encoded on the target is in front of or behind the RDS target plane. Immediately after their response, a new target is presented on the opposite side (left or right) and in one of the 3 depth locations. The new target is outlined with the yellow border as shown in Figure 3.3. The previous target is still visible, but without the yellow border. Two targets are visible at all times

The entire experiment encompasses 2280 trials, composed of 6 blocks made up of balanced sub-blocks. Each sub-block is composed of 10 cycles. Figure 3.5 provides an overview of stimuli presentation. A cycle contains every permutation of depth change. Figure 3.4-right provides an example of a complete cycle. Cycles are used to ensure that the depth change factor is balanced throughout the experiment. Each cycle is composed of 18 trials (9 depth changes and two possible positions: left and right) plus one additional trial (a 19th trial) to transition from one cycle to another. Responses from the additional transition trial are not included in the results because the 19th trial in each cycle would not be balanced. The orientation of the encoded shape (in front or behind the RDS target plane) is randomized. Response time and response are recorded for analysis.

3.4 Results and Discussion

The collected data enabled us to analyze response time, accuracy and self-assessments from the questionnaires. Analysis was performed with a three factor repeated measure ANOVA, using Greenhouse-Geisser adjusted degrees of freedom. Post-hoc pair- wise comparisons with Bonferroni corrections were run for multiple comparisons. All three main effects were significant.

[1] Depth change codes: **F-F** (Far to Far), **N-N** (Near to Near), **Z-Z** (Zero Parallax to Zero Parallax), **F-N** (Far to Near), **F-Z** (Far to Zero Parallax), **N-F** (Near to Far), **N-Z** (Near to Zero Parallax), **Z-F** (Zero Parallax to Far), **Z-N** (Zero Parallax to Near)

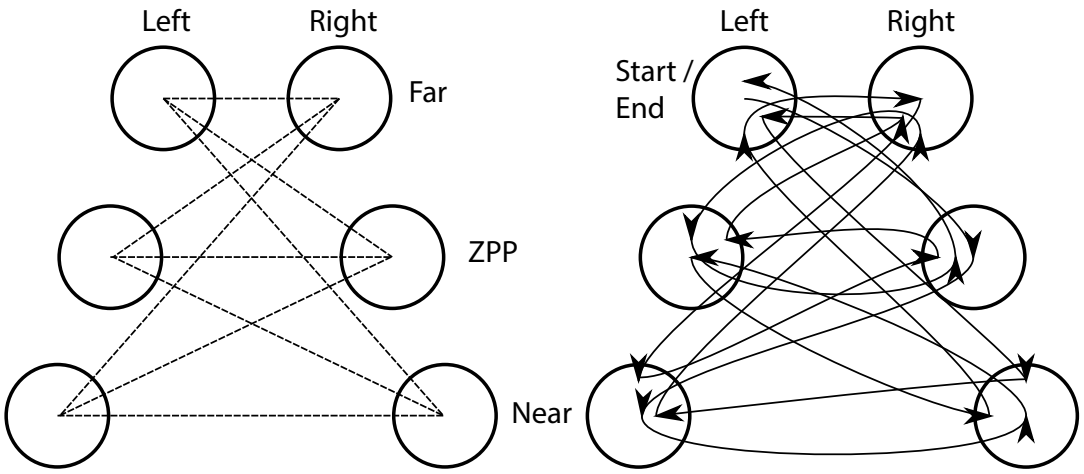


Figure 3.4: *On left: Possible target locations (circles) and depth change levels (lines). Assuming symmetry, the 18 lines reduce to 9. On right: An example cycle, which is balanced to contain 18 (2x9) balanced depth changes.*

Main Effect: Continuous Depth The main effect of Continuous Depth had a significant influence on the response time (F(1,11) = 7.587, MSE = $1.51 * 10^6$, $p < .05$). Participant average response time was 1086 ms with and 1253 ms without continuous depth, an average performance increase of 13.4%.

Main Effect: Depth Change The main effect of Depth Change also significantly influenced the response time (F(8,88) = 10.046, MSE = $6.29 * 10^5$, $p < .001$). A significant interaction was observed between Continuous Depth and Depth Change (F(8,88) = 4.386, MSE = $6.74 * 10^4$, $p < .001$). This implies that the Continuous Depth cue does not always reduce the response time. A detailed discussion of the influence of continuous depth on each of the nine possible depth changes proceeds below.

3.4.1 Depth Change

Response time per depth change are summarized in Figure 3.6. To facilitate analysis, we classify the depth changes into three different types: lateral, inward, and outward changes.

Lateral Change Trials in which both stimuli are located at the same depth are labeled as lateral change. Three depth changes meet this condition: F-F (Far-to-Far), N-N (Near-to-Near), and Z-Z (Zero Parallax-to-Zero Parallax).

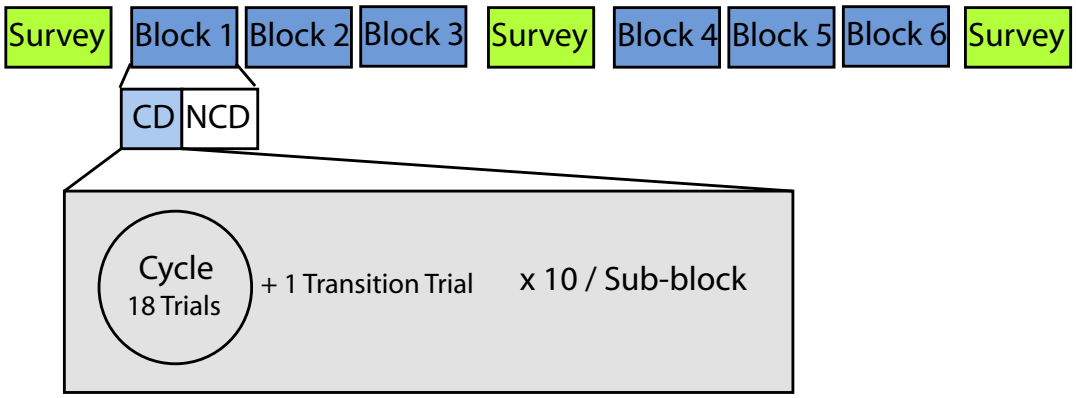


Figure 3.5: *Six blocks total, each containing six sub-blocks of continuous depth and non-continuous depth trials (balanced order across subjects). Each sub-block contained ten randomly selected, balanced cycles of all depth changes, plus transition trials to the next cycle.*

For the Z-Z condition, attention transitions are along the zero parallax plane. We observed an improvement of 4.73% in response time in the presence of the CDP, however the effect was not statistically significant ($p = 0.187$). We hypothesize the small improvement is due to the CDP providing an additional cue to direct attention between the two target stimuli [Egly et al., 1994].

The other two lateral changes exhibited significant improvement in the presence of the CDP ($p < .05$). Mean improvement due to the CDP was 15% for near (N-N) and 8.74% for far (F-F) lateral changes. It appears the CDP again provides a visual cue to direct attention. However, we may observe statistical significance because the cue additionally helps the visual system maintain the decoupling of eye-vergence and accommodation during the attention transition. We hypothesize that during the attention transition, the visual system tries to return to a state of natural correspondence between eye-vergence and accommodation.

It should also be noted that the lateral distance between stimuli differed for each lateral change condition. The lateral distance was greatest for N-N and smallest for F-F. In both cases, the continuous depth plane improved visual performance to change attention, with a greater improvement occurring when the lateral distance was longer. The CDP appears to both provide a directed attention cue [Egly et al., 1994] and also to help maintain the necessary decoupling between eye-vergence and accommodation [Howard, 2002].

Inward Changes Trials in which visual attention changes from a far target to a near target are labeled as inward changes. Three depth changes meet

this condition: Far-to-Zero Parallax (F-Z), Zero Parallax-to-Near (Z-N), and Far-to-Near (F-N). The F-Z condition is interesting in that visual attention is transitioned to the zero-parallax plane, where the eye-vergence and accommodation conflict is minimum. As would be expected, we do not observe a statistically significant improvement ($p = 0.102$).

The F-N condition exhibits a trend of reducing response time with the CDP. However, the mean improvement was near, but not yet statistically significant ($p = .078$). The Z-N condition did exhibit statistically significant effect of CDP.

Outward Changes Trials in which visual attention changes from a near target to a far target are labeled as outward changes. The remaining three depth changes meet this condition: Near-to-Zero Parallax (N-Z), Zero Parallax-to-Far (Z-F) and Near-to-Far (N-F). The transition to zero parallax (N-Z) exhibits a similar behavior as F-Z, in which attention transitions to a location of minimum eye-vergence and accommodation conflict. The response time for N-Z was not significantly reduced by the CDP ($p = 0.134$).

The remaining two outward depth changes, Z-F and N-F, do show statistically significant improvement ($p < .05$). The performance improvement is especially interesting. We observe an 18% reduction in response time for Z-F and 20% reduction for N-F. The outward depth changes show a trend to take longer than all other depth changes. We observed a mean reduction in response time of approximately 300ms for those two conditions.

Observations The inclusion of a continuous depth plane (CDP) linking two targets provided a statistically significant 14.31% reduction in response time required to change attention. If we exclude depth changes to the zero parallax plane (Z-Z, N-Z and F-Z) and the F-N condition (because it was not statistically significant), we observe an 18.10% reduction in time to change visual attention. A simple change in scene composition can have a significant influence on the viewer's ability to attend to elements within the scene. Viewers are able to change their spatial attention faster when a continuous depth plane is present.

Another observation from analyzing the data is the direction dependence on transitions in depth. Attention transitions from either near or far locations to zero parallax take approximately the same time. However, the other two outward changes take longer than the other two inward changes. This observation appears consistent with a simple biological model of the eye accommodation.

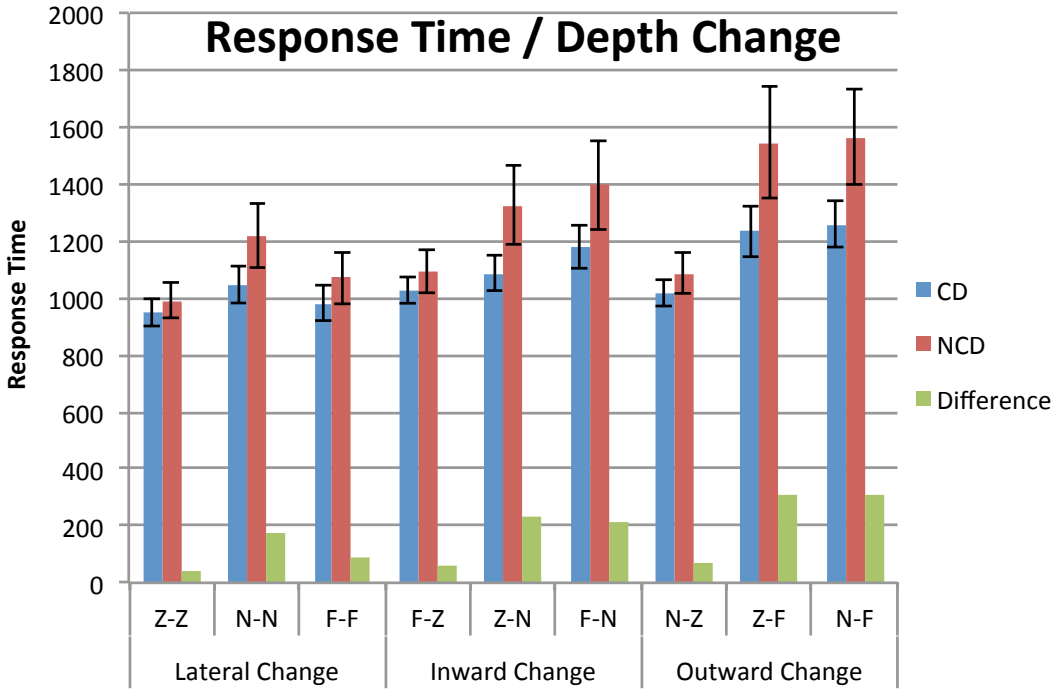


Figure 3.6: *Comparison of the average response time per depth change. Continuous depth improves the outward depth change most. F: Far, N: Near, Z: Zero Parallax.*

The lens of the eye changes shape to accommodate as represented in Figure 3.7. This is accomplished by contracting or relaxing internal ciliary muscle, which adjusts tension on the zonula fibers that radiate from the lens. Since the ciliary is an annulus muscle, contraction decreases the diameter of the muscle resulting in releasing tension and increasing convexity of the lens. When the ciliary muscle tightens, the eye accommodates to a nearer point. Relaxation of the ciliary muscle increases tension on the zonula fibers resulting in far focus. Since contraction of a muscle is always faster than relaxation, we expect to see an asymmetry in the time to change accommodation. This effect agrees with our data: changes of accommodation from far-to-near are faster than near-to-far.

Since vergence can drive accommodation [Nguyen et al., 2008; Howard, 2002], we could reason about the demanding process of changing visual attention in stereoscopic images. When the visual system changes eye-vergence, a natural response causes an initial reaction to adjust accommodation to the new fixation point. However, that adjustment will result in focus deteriorating because all visual information is fixed at the zero paral-

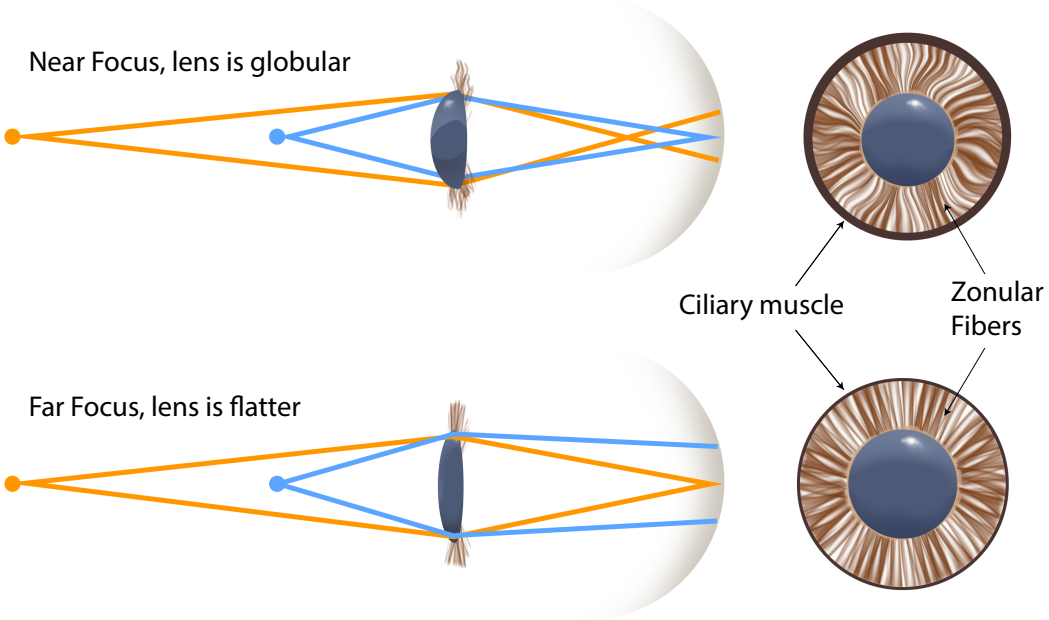


Figure 3.7: *Top: Ciliary muscle contracts, relaxing zonula fibers. The lens thickens to facilitate accommodation of near objects on retina. Bottom: Ciliary muscle relaxes, placing tension on zonula fibers. The lens stretches and becomes more flat to facilitate accommodation of far objects on retina.*

lax plane. The visual system then begins the counter-intuitive response of decoupling eye-vergence and accommodation.

We hypothesize that the continuous depth plane provides additional eye-vergence cues to assist the visual system in compensating for the decoupling with accommodation. In the case of our stimuli, the visual system may saccade via the continuous depth plane to the new target. Without this additional information, the visual system may invoke larger or more time consuming accommodative changes.

3.4.2 Measuring Fatigue

The second question posed in our study was whether fatigue observed by performance measures correlates with self-assessed questionnaire data. This requires an analysis of not only response time, but also response accuracy and the change in questionnaire data throughout the experiment.

There remains one main effect that we did not discuss in the previous analysis of the influence of continuous depth on depth change. That main effect

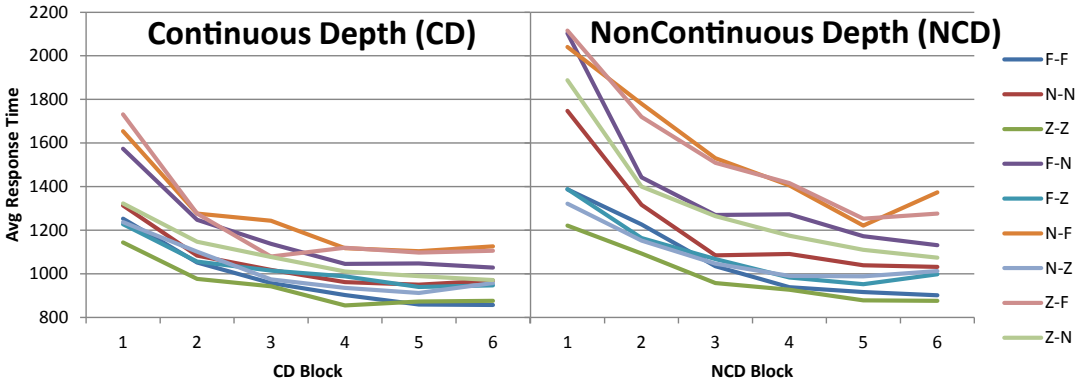


Figure 3.8: *Comparison of response time and of different depth changes across different blocks with and without continuous depth. F: Far, N: Near, Z: Zero Parallax.*

is the *Block*. Our experiment task consisted of six blocks: each block encompassed all permutations of depth changes as well as two conditions: with and without a continuous depth plane (CDP), which are isolated in sub-blocks. Each block consisted of 380 depth discrimination trials. In total, each subject evaluated 2,280 trials. In addition to these blocks, we administered three questionnaires at three points during this task: before, at the mid-point (e.g. Between blocks 3 and 4), and at the end, immediately after block 6, as summarized in Figure 3.5.

Main Effect: Block The main effect of the Block factor has a significant influence on response time ($F(5,55) = 6.509$, $MSE = 8.57 * 10^6$, $p < .001$). Figure 3.8 presents a trend in the data for the response rate to decrease from an average of 1500 ms in Block 1 to minimum at around Block 5 and a small increase in Block 6. Subjects tend to become faster throughout the experiment. We interpret this in two ways: First, there seems to be a learning effect in the first 1-3 blocks. Second, subject accuracy declines and the time to achieve fusion increases in the sixth block in exactly those depth changes that are the most difficult to perform and that often require the longest time to complete (e.g. N-F, N-Z, Z-F, F-Z, N-N).

Analysis of accuracy rate also reveals a significant main effect of Block ($F(5,55) = 10.355$, $MSE = 51.67$, $p < .001$). Closer analysis reveals that more errors are made in Block 6 than Blocks 1, 2, 3, and 4 ($p < .05$).

We observed a trend for slower attention transitions for the difficult depth changes and an increase in the error rate. This behavior is expected when subjects experience performance fatigue. Next, we seek to relate these ob-

servations with subjective self-assessment data provided by the experiment questionnaire.

Questionnaire We break the questionnaire analysis into 3 Survey Blocks. Survey Block 1 is in the beginning. The second is at the midpoint, which occurs 15-36 minutes into the experiment. The final block is at the end, which occurs after 30-70 minutes, depending on the participant's response rate throughout the experiment. Note that the substantial difference in elapsed time is indicative of the performance variance observed among test subjects.

Analysis of the general eye fatigue question as well as the 6 KAB questions resulted in a significant main effect of Survey Block (F(2,22) = 13.221, MSE = 10.111, $p < .001$) The results are as follows: assessment of eye fatigue showed a significant increase between the three questionnaire phases of the experiment ($p < .001$). We observed two other general phenomena. Some assessments increased in the 2nd block, but did not significantly change in the 3rd Survey Block. Those assessments include the subject feeling more queasy ($p < .05$). The following assessments were significant between the first and third Survey Block, which we interpret as a more gradual increase: Subject feels more relaxed, feels more miserable, and more nervous ($p < .05$). Assessment of fatigue and blurry vision had a tendency to increase, but were not statistically significant.

The remaining questions are from the Simulator Sickness Questionnaire (SSQ). We first used the three factor analysis defined by Kennedy, et al. [1993] to determine that factors pertaining to oculomotor were most influenced by our test. The results presented in Figure 3.9 were expected because we did not present moving images that create a conflict between vestibular and visual motion that would influence the other 2 factors: nausea and disorientation.

ANOVA was then conducted on only the 7 questions pertaining to oculomotor factor of the SSQ. The main effect of Survey Block had a significant effect on those 7 questions (F=(2,22) = 14.098, MSE=5.671, $p < .001$). From those questions, General Discomfort gradually increases from Survey Block 1 to 3 ($p < .05$). Eyestrain and Difficulty Concentrating both increased between Survey Blocks 1 and 2 ($p < .052$), but did not change significantly between Survey Blocks 2 and 3. It is likely that these symptoms are perceived as initially worsening before plateauing at a general level of discomfort as stereoscopic viewing continues. General Fatigue and Blurry Vision had a tendency to become stronger, but they were not significant. Headache did not increase during the experiment.

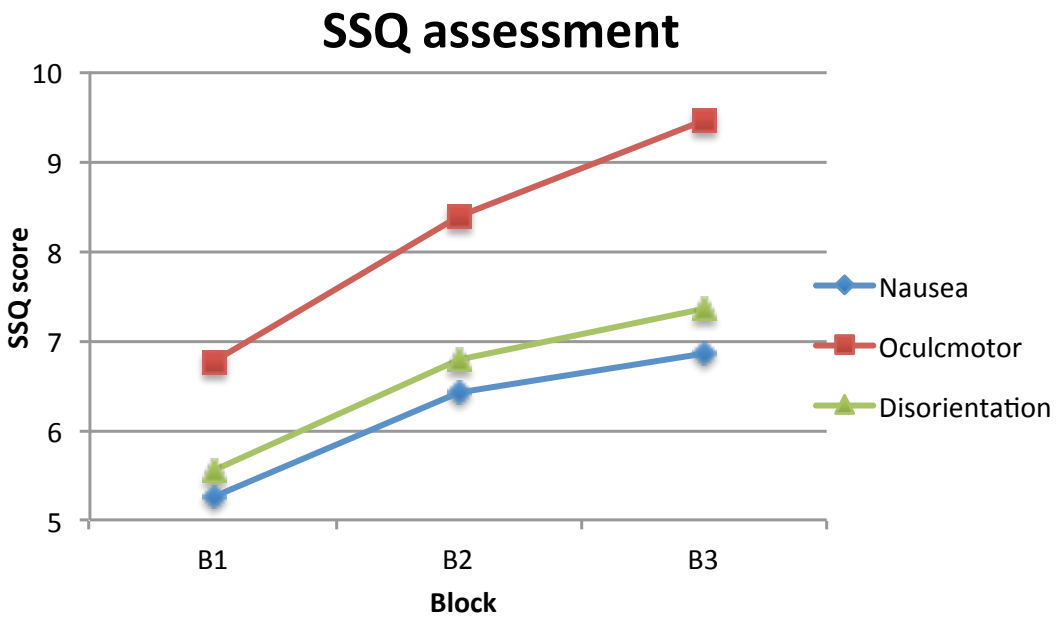


Figure 3.9: *SSQ Assessment: The Oculomotor factor is most influenced during the expeirment.*

The survey data indicates that symptoms such as eye fatigue and general discomfort tend to increase during the stereoscopic activity. Other symptoms may appear when beginning the task, but not worsen through the experiment duration.

Additional Subject Pool Observations We used a screening procedure to verify that all subjects had normal binocular spatial vision. However, among our sample set (psychology and computer science graduate students) we were surprised by an extremely wide variance among the subjects. Some subjects require up to three times longer than others to achieve fusion. Approximately 30% of our potential subjects were unable to achieve fusion for targets with absolute disparities that were within about 10% of parallel viewing. This variance implies that if stereoscopic 3D is to be successful, a conservative approach should be taken in order to establish safe boundaries on the dynamic range of depth within a scene.

3.5 Conclusion

We have shown that changes in scene composition have a significant influence on the viewer's ability to change visual attention among spatially distinct scene elements. We also observe that self-assessment of eye fatigue and general discomfort increase before a decrease in visual performance is observed. For 3D cinema and interactive media to remain as a viable entertainment genre, additional studies of this type may ensure that content will reach the widest possible audience.

The important message is that scene composition, framing and montage can significantly influence visual performance in terms of time to change visual attention. Visually important, or salient, scene elements can be viewed more quickly when connected by visual information that continuously varies in stereoscopic depth.

C H A P T E R 4

Stereoscopic Window Violations

This chapter presents exploration of a visual disturbance that occurs when the depth cue occlusion is in conflict with stereopsis.

4.1 Introduction

This chapter presents a significant problematic stereoscopic artifact influencing S3D quality: the stereoscopic window violation. A full description of the phenomenon is provided in Section 4.2. Stereoscopic window violations occur when a scene element perceived to be in front of the stereoscopic window collides with the window border. The border appears to occlude the object in front of it creating a disturbing visual conflict. Fortunately, window violations are not always problematic. Making that assessment and compensating, if necessary, requires time consuming expert input. We present a computational model that identifies when a window violation is problematic and system to automatically correct it.

Our main contributions are the following:

- Subjective measurement of disturbing window violations as a function

of luminance contrast magnitude, spatial frequency, orientation, and disparity,

- A perceptual model based on these measurements and metric to predict detection of problematic window violations,
- An experimental procedure to calibrate and validate the perceptual model,
- Applications to assist stereo content creators in the detection of problematic window violations and in the automatic application of floating windows.

4.2 Background

This section provides background information about the stereoscopic window as well as related work in visual processing methods.

4.2.1 Stereoscopic Window

The stereoscopic window represents the virtual window through which stereoscopic depth is perceived. Its importance has been recognized since stereoscopic pictures were first made [Spottiswoode and Spottiswoode, 1953]. The stereo window is defined by the parallax of the lateral edges of the two images projected on the screen as represented in line segment p_1p_2 of Figure 4.1-a. At zero parallax, the window is perceived to be at the plane of the screen.

Stereoscopic window violations occur when a scene element meets two conditions: (1) it must be presented with disparity nearer than the stereo window, and (2) it must collide with the lateral image border as shown in Figure 4.1-a. In the real-world, this problem would never exist. An object in front of the window is visible to both eyes (Figure 4.1-b). Stereoscopic images constrain the visible space for presenting objects in front of the screen (Figure 4.1-a). When objects in front of the stereo window are only visible to one eye, the viewer perceives the image border to be occluding the missing information. Two depth cues are in conflict: disparity provided by the visible stereo features and the depth ordering from occlusion. Humans are most sensitive to occlusion in identifying proper depth order [Cutting and Vishton, 1995], therefore violating it has the potential to produce a disturbing visual conflict.

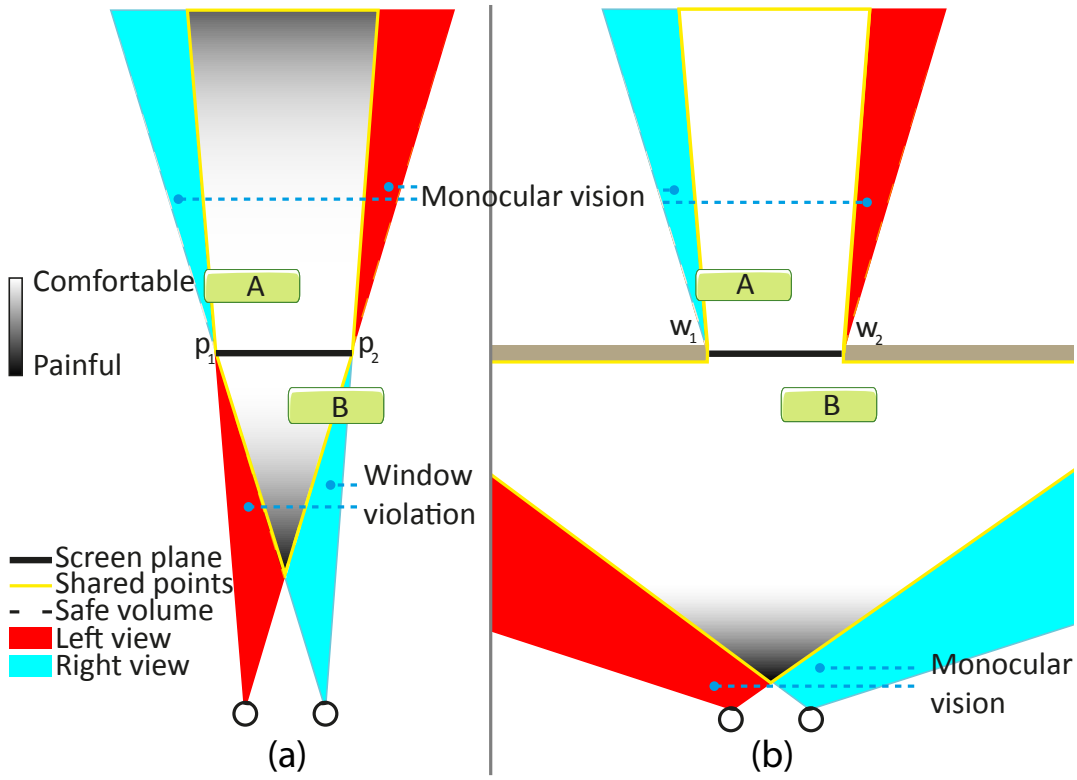
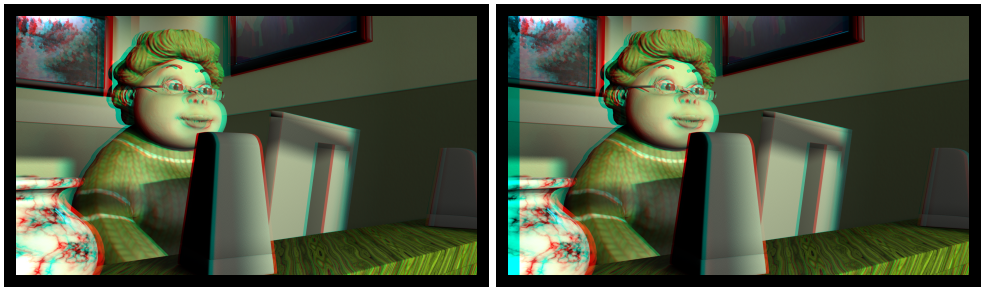


Figure 4.1: *Stereoscopic viewing scenario (a) results in window violation. Features from Object B are occluded by the screen edge p_2 behind it. In a real world viewing scenario (b), features from Object B are visible to both eyes. The window edge, w_2, is occluded by Object B. Note: Object occlusions are omitted.*

Window violations are difficult to avoid. The range of stereoscopic depth that is comfortable to view is limited to a region both in front of and behind the screen [Shibata et al., 2011]. Maximizing this zone of comfort requires placing scene elements in front of the screen (negative parallax), increasing the likelihood that a window violation will occur.

Floating Windows are a common solution to remove stereoscopic window violations. It is produced by applying an asymmetric mask to the left and right eye images. Figure 4.2 provides an example scene with and without floating windows applied. Floating windows remove features that should be visible in the two eyes. The black border to the left of the image appears to float in front of the vase. This preserves the expected depth ordering of occluding elements.

Floating windows have been applied in feature cinematic films as a tool to utilize a larger depth volume, regain artistic control and remove the violation artifacts that distract attention from the story [Neuman, 2009]. Digital



(a) Window Violation (b) Floating Windows

Figure 4.2: *(a) Stereoscopic image with window violation. (b) Window violation removed with floating windows. Note the asymmetric mask on the left.*

cinematography makes it easier to vary floating windows between shots, animate them within shots, and place them to correct only the problematic violations. However, these are all time consuming tasks requiring expert input that can be especially difficult to apply in real-time applications. Our research provides a computational model to assist and automate these operations.

There are other solutions for removing window violations. One strategy is to simply translate the images to globally push the depth back behind the image plane. This method risks exceeding the comfort zone behind the screen plane. Local disparity warping methods could be applied to move only a conflicting object behind the image plane, as demonstrated by Lang et al. [2010]. Less frequently used solutions include blur near the image border reducing visibility of the violation [Lipton, 1982]. Many of these solutions, including floating windows, are currently applied in an empirical, ad-hoc manner. We aim to formalize this through a computational model of perception.

In the next section we will first discuss some relevant topics in 2D perception, and next we introduce additional considerations for modelling binocular 3D perception.

4.2.2 Perceptual Modeling

Contrast Sensitivity Human vision is better at distinguishing two objects when their relative difference in color or luminance is large. This difference can be expressed in terms of *contrast* [Barten, 1999]. The inverse value of the minimum contrast required for detection is called *contrast sensitivity*. The change in contrast sensitivity is a function of spatial frequency and is modeled by the well-known *contrast sensitivity function*. Through our exper-

iments, we found that spatial frequency is a key component that also affects our perception of window violations.

Image Quality Metrics Our method is similar to image quality assessment metrics with the difference being our interest in assessing window violations rather than detecting visible differences. Perceived contrast distortion metrics, such as the *visible difference predictor* (VDP) [Daly, 1992], are designed for detecting near threshold differences. The update of the metric HDR-VDP-2 [Mantiuk et al., 2011] extends the VDP for high dynamic range applications, and it is capable of expressing suprathreshold difference in JND (just noticeable difference) units. These types of models are based on low-level representations of the human visual system (HVS) including contrast sensitivity and visual masking [Watson and Solomon, 1997].

Another approach to modeling perceived image quality is the use of *structural similarity index metric* (SSIM) [Wang et al., 2004], which exploits the structural information from a scene. Higher level visual equivalence models provide metrics to determine when perceived image changes do not result in a perceived change in image quality [Ramanarayanan et al., 2007; Krivánek et al., 2010]. Similar to our method, these can be considered suprathreshold models because they do not predict the probability of detection, but rather if the change is significantly visibile. These considerations are very important and influence the performance of an image quality metric, as shown by Cadik, et al [2012].

4.2.3 Stereoscopic Visual Processing

Stereopsis also has a contrast sensitivity correlate. Frisby and Mayhew's demonstrated a correlation between stereopsis sensitivity and contrast detection sensitivity as a function of spatial frequency [1978]. Their findings show that the shape of contrast detection and stereopsis are similar, although with a shift representing a decrease in stereopsis sensitivity. It is possible to perceive the disparate features, but not achieve stereopsis. The CSF correlation with stereopsis has lead to models of perceived depth of frequency and magnitude changes in disparity [Didyk et al., 2011]. In contrast, our work is focused on detecting disturbing window violations.

To evaluate stereoscopic window violations, we are less concerned about computing the JNDs of disparity, but are more interested in assessing if the visual system can find a viable depth interpretation when point correspondences do not exist. Marr and Poggio [1976] developed a cooperative algo-

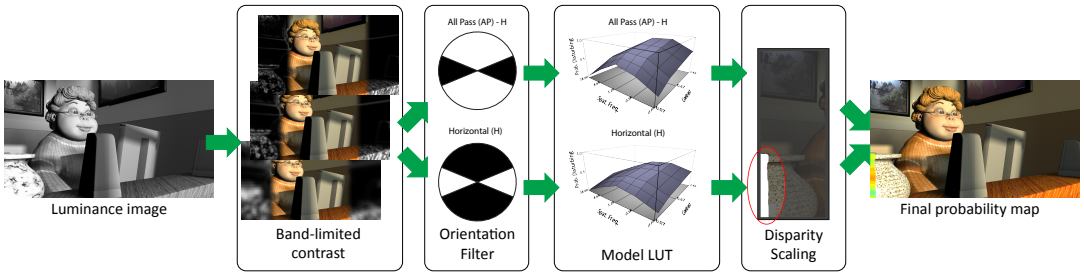


Figure 4.3: *Pipeline of our computational model. A luminance image is decomposed into band-limited contrast and contrast orientation channels. We then apply our predictive model and additional disparity scaling. Results are combined using winner-take-all producing the final probability map.*

rithm for extracting disparity information. They suggest that perceiving the disparity of an object involves finding a continuous and smooth matching of point correspondences.

Mitchison and McKee [1985; 1988] later observed that strong edges can cause a stable, yet incorrect correspondence match of stereoscopic stimuli. The visual system favors strong edges at the expense of misinterpreting the fine texture detail. This is a significant finding. For stereoscopic window violations, the image border can provide a strong edge biasing the interpretation of scene elements to be perceived as if they are behind the stereoscopic window, and free of violation.

4.3 Problem Statement

Our goal is to create a computational model for the perception of stereoscopic window violations to produce a binary classification of a window violation being disturbing or not. We are guided by two key concepts: (1) Stereopsis sensitivity has a CSF-like behavior, and (2) strong edges of the image borders can influence the depth interpretation of a window violation. We hypothesize that a scene element in window violation will be problematic when it is represented by strongly visible contrast.

We isolate four dominant variables to develop our perceptual model: contrast magnitude, spatial frequency, orientation, and disparity. To build our model as shown in Figure 4.3, we take an experimental approach as described in the following sections. The approach is visualized by Figure 4.4. Section 4.4 reports our subjective experiments to evaluate the influence of the different

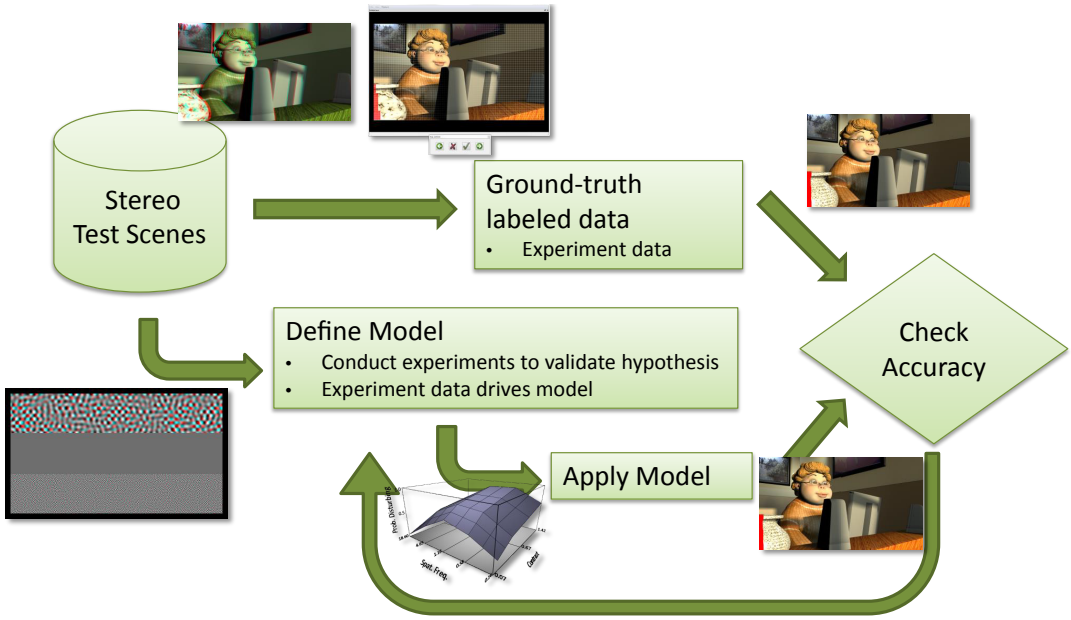


Figure 4.4: *Workflow applied to develop computational system to analyze stereoscopic window violations. The bottom portion of the flow diagram represents the construction of model by careful creation of window violation stimuli and application of that model to real image content. The upper portion of the flow represents the development of ground truth user labeled data, which provides labels of where window violations occur. This is used to evaluate model performance and involves an iterative process of refining model experiments and the computational system.*

variables on perception of window violations. We then describe the computational model in Section 4.5, which represents the bottom portion of the flow diagram of Figure 4.4. Section 4.6 presents how we calibrate and validate the model with complex stereoscopic images. This requires the creation of ground truth labeled data as as represented in the upper portion of the flow diagram in Figure 4.4. We discuss our results in Section 4.7 and finally describe applications in Section 4.8.

4.4 Model Experiments

We now present the details regarding perceptual experiments required to create the computational model in Figure 4.3.

4.4.1 Stimuli

Our experiments require the ability to change the contrast magnitude, spatial frequency, orientation, and disparity of the stimuli. We conducted our initial experiments with stimuli consisting of sinusoidal gratings with Perlin noise [Perlin and Hoffert, 1989] that resulted in textures similar to the ones used by Ferwerda et al. [1997]. However, we found that the sinusoid gratings resulted in periodic ambiguity and the Perlin noise added additional spatial frequencies outside of the considered range. We overcame this problem through the use of random dot stereograms. The random dot stereograms used in our experiments were filtered in order to confine them within specific ranges of contrast magnitude, orientation and frequency.

We presented 15 combinations of spatial frequency and contrast. Five spatial frequency levels were investigated spanning the range from 0.14 cpd to 18.6 cpd. The levels of contrast were 0.21, 0.63 and 1.35. An example trial of our experiment is shown in Figure 4.5. Trials were balanced for presentation at the top and bottom of the stimulus. When orientation-specific stimuli was required, we utilized the fan filter as implemented by Watson [1987]. Disparity was controlled by translating the stimuli.

In order to create the stimuli used at each trial, the generated textures were uniformly applied to two planes. The max pixel disparity in our stimuli was -50px, corresponding to an angular disparity of -0.8°. Under these conditions, window violations were created for both planes. The height of each plane was 350px (angular height of 5.9°) and their width was 1920px (angular width of 30.75°). Further details about the stimuli are provided below when describing specific experiments.

4.4.2 Procedure

The experiments were implemented as a two-alternative forced choice (2AFC) procedures. Human subjects were paid and naive to the experiment. Subjects were first introduced to the concept of window violations. Then they were asked to compare the two planes in the stimuli image (e.g. Figure 4.5) and choose which of the two looked less disturbing or annoying. They were instructed to base their assessment on the regions of the display close to the image borders, where window violations are expected to occur. The stimuli was shown on a 50″ Panasonic 3D plasma TV (TX-P50VT20E) in a darkened room. Subjects were seated two meters away from the display wearing active shutter stereoscopic glasses. All subjects had normal or corrected-to-normal

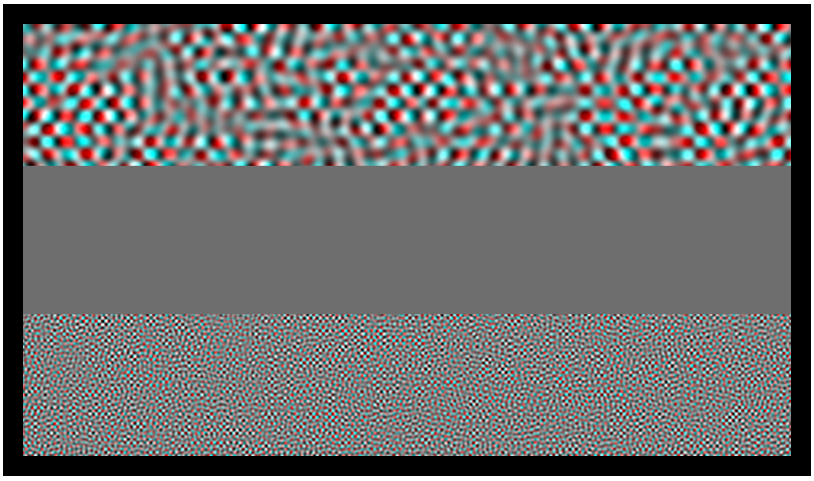


Figure 4.5: *Example stimuli presented in the main model experiments. Both stimuli represent the All Pass condition with no orientation filtering. The top stimulus is spatial frequency condition level three (SF3) at 1.16 cpd, and the bottom stimulus is SF4 at 4.65 cpd. Example provided in anaglyph. The black border is provided by the black border of the HDTV display device in a darkened room.*

visual acuity and stereo acuity. The duration of each session was approximately 40 minutes.

Our **main experiment** produced the primary model look-up table (LUT). It was performed with six subjects each of which evaluated the balanced combinations of five spatial frequency levels and three contrast levels resulting in 225 trials per experiment. Each subject performed three experiments, one for each orientation condition. The orientation condition for this experiment was produced using a broadly tuned, four segment fan filter (45° per segment) to achieve the following orientation conditions: all-pass (no orientation filtering), horizontal and diagonal orientation. The horizontal and diagonal orientation conditions were produced by centering one segment of the fan filter on the respective axis of interest. Only max, -50 pixel disparity condition was used.

Orientation mixing experiment was additionally conducted by mixing stimuli orientations as shown in Figure 4.6. This experiment was conducted with six new subjects who had not participated in the main experiments. The max contrast and disparity condition levels were held constant. Spatial frequency and orientation was balanced resulting in 400 trials per experiment. A narrow orientation filter was used (one 22.5° segment from an eight segment fan filter) to produce the following orientation conditions: all-pass (no orientation filtering), horizontal, diagonal and vertical orientations.

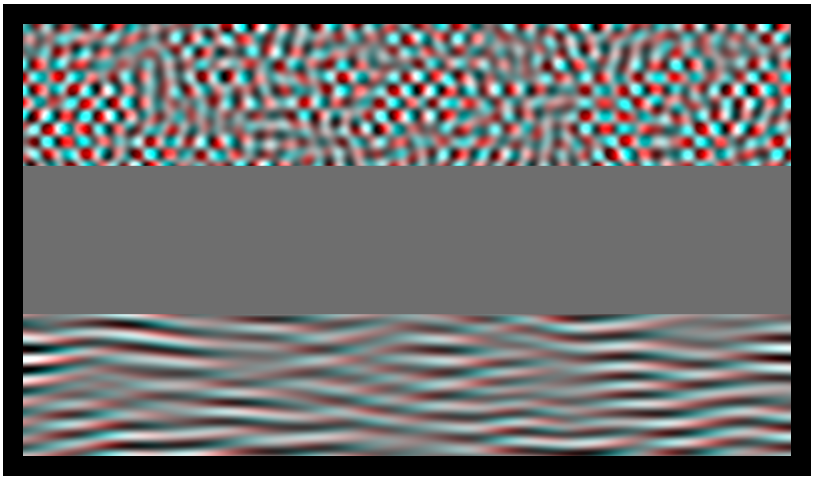


Figure 4.6: *Example stimuli presented in the orientation mixing experiment. The top stimuli is the All Pass (AP) condition. The bottom stimuli is the Horizontal (H) condition. Both stimuli are spatial frequency condition SF3 (1.16 cod). Example provided in anaglyph. The black border is provided by the black border of the HDTV display device in a darkened room.*

Disparity mixing experiment was conducted through a balanced presentation varying the disparity and spatial frequency conditions while holding contrast and orientation constant. The max constant level was used. There was no orientation filtering (only All Pass condition). Four levels of negative disparity were compared within the experiment: 6, 12, 25 and 50 pixels (ranging from -0.1° to -0.8° angular disparity). The experiment involved 400 trials per experiment.

4.4.3 Results

The responses collected from each session of *our main experiment* are stored in a single matrix, with 15 entries in total: one for each combination of spatial frequency and contrast condition level. The values were normalized resulting in a matrix describing the probability for each combination to be preferred.

We performed Two-factor Analysis of Variance (ANOVA) with repeated measures to analyze the influence of orientation, spatial frequency and contrast on viewer preference. The between-subjects main effect of orientation did not have a significant influence on stimuli preference ($p = .391$). We believe this is due to the broad orientation filter width. There was, however, strong statistical significance for within-subjects main effects of contrast (F(2, 30) =

367.32, $p < .001$) and spatial frequency (F(1.32, 19.89) = 100.43, $p < .001$) using Greenhouse-Geisser adjusted degrees of freedom.

The lack of a broad orientation effect motivated the use of only the All Pass orientation condition as data for our primary model LUT. We also analyzed the Bonferroni-adjusted pairwise comparison between the levels of each condition [Sheskin, 2007]. For the All Pass orientation, we observed a statistically significant difference in preference between all contrast levels ($p < .01$) and nearly all spatial frequency levels.

Figure 4.7 provides a visualization of data used to produce the model LUT. The plot describes the inverse of the users' response, meaning that the highest combinations were least preferred. It is based on the assumption that the least preferred combinations are also the most disturbing or annoying. The plot exhibits a monotonic preference response for each individual spatial frequency in terms of contrast. Window violations with higher contrast are less preferred. The plot also shows similar behavior to the familiar contrast sensitivity function for luminance. This result supports the assumption that a disturbing window violation is significantly influenced by contrast magnitude and frequency.

The first experiment motivated further exploration of orientations in narrower bands. We suspected that the insignificance of orientations in the main experiment was due to the broad orientation tuning. The narrower tuning used in the **orientation mixing experiment** enabled us to directly observe a significant orientation effect (F(3, 15) = 14.29, $p < .001$) in addition to reproducing a significant spatial frequency effect (F(4, 20) = 4.06, $p < .05$). The horizontal condition was the only condition to show an insignificant influence ($p = .125$) on spatial frequency. The spatial frequency preference curve was much flatter, nearly not resembling the CSF curve. Pairwise comparison reveal no significant preference between spatial frequency levels for the horizontal condition.

All Pass, Diagonal and Vertical stimuli orientation conditions had similar mean preference scores. However, the horizontal condition was significantly preferred more than the All Pass and Diagonal orientations ($p<.05$) and a trend for preference over vertical ($p = .066$). The mean preference for the horizontal condition was 30% higher than all conditions and 32% higher than All Pass. This experiment motivated us to construct two model LUTs: one representing All Pass except horizontal (AP - H) and the other horizontal only (H). The horizontal LUT is modulated by a coefficient representing the mean preference ratio of horizontal to the other conditions.

The **disparity mixing experiment** produced expected results showing a sig-

nificant effect from varying the disparity condition ($F(3, 15) = 58.35$, $p < .001$). Pairwise comparison revealed a significant preference between disparity levels for all except the the two smallest ($p = .058$). There was also a significant interaction between disparity and spatial frequency ($F(2.77, 13.83) = 8.52$, $p < .001$). Smaller disparities showed a flatter CSF-like preference. However, we also observed that changes in texture alignment with the window violation caused some differences in per disparity spatial frequency preference curves. For this reason, we omitted the shift in spatial frequency preference from our model. We include the significant preference for smaller disparities, which fit a log-linear scaling of preference as a function of disparity. The linear fit of the data is $p_d = 0.1603 \log d + 0.0992$ with a goodness of fit, $R^2 = 0.9959$. p_d is the mean probability that the disparity condition, d, is disturbing.

For application purposes, a clear threshold between disturbing and non-disturbing window violations is beneficial. We present an experiment and method to find such a threshold in Section 4.6 after discussing the computational model.

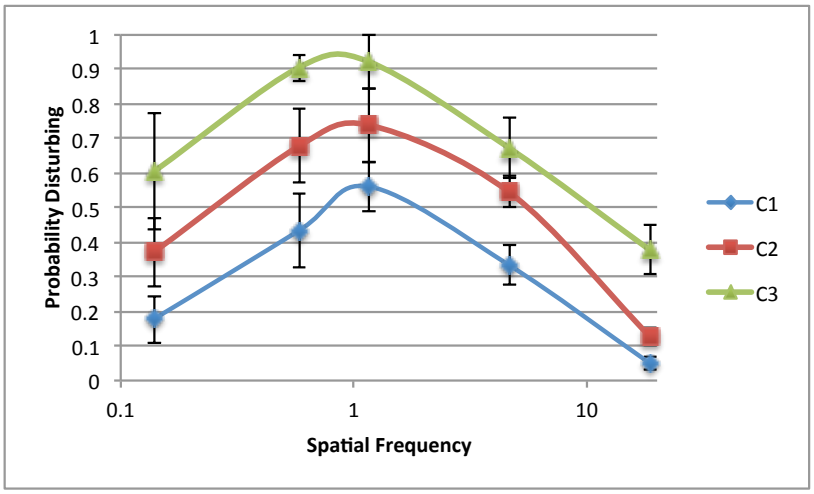


Figure 4.7: *The mean probability for each combination of contrast and spatial frequency to be perceived as disturbing. Contrasts C1, C2 and C3 correspond to .21, 0.63 and 1.35 respectively. Error bars are standard deviation.*

4.5 Computational Model

We now utilize our experimental data to produce the computational model, depicted in Figure 4.3.

Based on the assumption that luminance is the main factor guiding stereopsis, we convert the stereoscopic RGB pair to individual luminance images. Viewing conditions in terms of luminance depend on the spectral emission properties of the display being used. To create a luminance image the display was characterized using a Photo Research PR-730 spectroradiometer. The obtained spectral emission curves for the three channels were used together with the color matching function of the XYZ colorspace [Ohno, 2000] to obtain the luminance image described by the Y component. The resulting image will describe the amount of $\frac{cd}{m^2}$ emitted by the display in a per-pixel basis.

In order to extract contrast and spatial frequency information out of a complex luminance image we use a similar approach as the one presented by Peli [1990]. We decompose the image in band-limited versions by applying cosine-log filters as in the original paper. A cosine-log filter centered at 2^i cycles/picture is defined as:

$$G_i(u,v) = G_i(r) = \frac{1}{2}(1 + cos(\pi log_2 r - \pi i)), \tag{4.1}$$

where u and v are the horizontal and vertical spatial frequency coordinates respectively and r is one of the polar spatial frequency coordinates defined as $r = \sqrt{u^2 + v^2}$.

The contrast of the ith band is computed as:

$$c_i(x,y) = \frac{|a_i(x,y)|}{L'}, \tag{4.2}$$

where c_i is the ith contrast image, a_i is the band-limited luminance image and L' is the mean luminance of the luminance image. L' is motivated by its use in VDP-related metrics [Daly, 1992].

We apply our model to each contrast image using the LUT constructed with data from Figure 4.7. Bilinear interpolation is used between sample points. This results in one probability map for each spatial frequency band. Bands are combined together using a winner-take-all approach, such that for each pixel we use the maximum probability value across all probability maps. This approach is motivated by the independent-channel hypothesis in which disparity discrimination is influenced by the largest active spatial frequency channel [Marr and Poggio, 1979].

The LUT is further modulated by scaling coefficients to reflect the effect of orientation and disparity. The orientation scaling factor is only applied to the LUT for the horizontal component of our pipeline. In this case, a scaling factor, $s_o = 0.7$, is applied to the LUT.

The disparity scaling, s_d, is computed by shifting the linear fit from the disparity experiment such that our max disparity tested, 50 pixels, is represented by $s_d = 1$. Our primary LUT already represents the probability disturbing for 50 pixel disparities. The result is the following disparity scaling function: $s_d = 0.1603 \log d_{max} + 0.3598$. Since disparity is undefined in window violation regions, we set d_{max} equal to the max window violation size for all pixels of the given row.

Since we are only interested in the regions that actually are in window violations, we prune the probability map using a disparity map of the stereoscopic pair removing regions that are not in contact with the borders or do not have negative disparities. This gives a window violation detection mask as shown in the Disparity Scaling component in Figure 4.3 as well as in results Figure 4.10.

The probability map per orientation channel, k, is expressed as follows:

$$P^k(x,y) = \max(P_0^k(x,y), ..., P_N^k(x,y))\ s_o^k\ s_d, \tag{4.3}$$

where the max per frequency band, P_N, is modulated by orientation and disparity scaling coefficients. We then apply the max between orientation channels, H and AP-H, to produce a final per pixel probability map:

$$P(x,y) = \max(P^k(x,y)). \tag{4.4}$$

4.6 Validation Experiments

We conducted a subjective study using real-world and computer-generated images to validate our model and obtain a measure regarding its performance. Since our goal is to predict whether a window violation will be perceived as disturbing, we asked subjects to look at stereoscopic images and indicate where a window violation was disturbing.

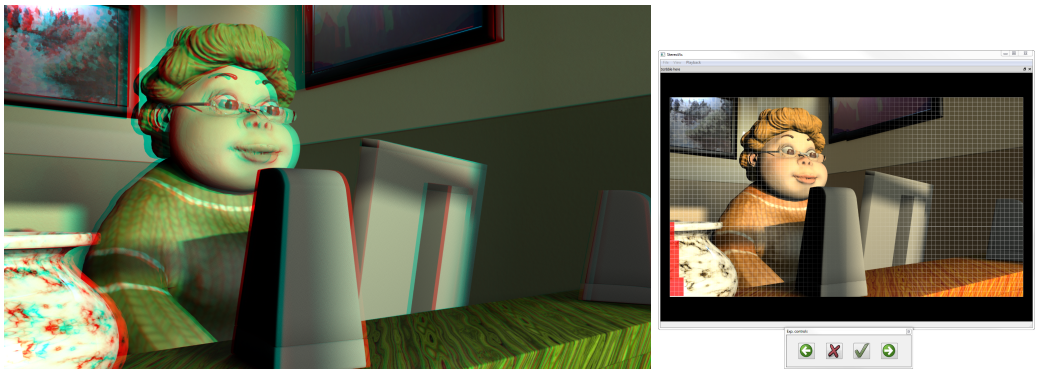
4.6.1 Stimuli

Stimuli consisted of 95 stereoscopic images: screen captures taken from stereoscopic movies (including live action and computer-generated imagery) and in-house produced computer-generated imagery. The stereoscopic images presented window violations as large as 50 pixels in width. The images created in-house described a similar scene, but with varying object texture, position properties, and camera configurations.

4.6.2 Procedure

The experiments used the same setup as the modelling experiments (see section 4.4) with an additional computer monitor for user input (Figure 4.8-b). There were 11 subjects, each had normal or corrected-to-normal vision.

Our subjective methodology was similar to the one presented by Aydin et al. [2010] to evaluate HDR video tone mapping. Subjects were instructed to look at the stereoscopic image in the 3D TV and localize the regions on the lateral borders perceived as disturbing or annoying. An additional 2D computer monitor showed the left half of the left image together with the right half of the right image overlayed with a grid. Each cell had a size of 25 pixels and users were asked to label cells containing disturbing window violations. Figure 4.8 shows an example trial.



(a) Stereoscopic image (b) 2D image with grid overlay

Figure 4.8: *Validation and Calibration Experiment. (a) Stereoscopic Image (presented here as anaglyph). (b) The user interface for grid-based selection of problematic window violations.*

4.6.3 Evaluation

The collected data was stored as binary images representing selected cells. These images were averaged per stereoscopic image across all 11 subjects. The resulting grayscale image contained the mean subjective response per cell, and it provided our ground truth labeled data.

To measure the performance of our model we computed 7-fold cross-validation between the subjects' response and our prediction. The cross-validation finds a threshold value for our model that minimizes the absolute difference between the true positive ratio (TPR) and true negative ratio (TNR).

We compared our model with the ground truth labeled data at a cell resolution of 50 pixels by quantizing both the ground truth and the model. Since the ground truth is subject to human labeling error and inconsistencies, we defined sensitivity for the ground truth. We labeled a cell as disturbing (positive) if 60% of subjects selected the same cell. This gave us in average a TPR: 71% and TNR: 72%.

As mentioned before, the cross-validation was set to find a balance between the TPR and TNR. If true positives or true negatives are not equally important, we could set the cross-validation to find a threshold value which favors true positives or true negatives instead. Figure 4.9 illustrates the performance when matching 60% percent of the ground truth labels while varying the importance of true positives over true negatives. For instance, if we set the true positives to be twice as important as the true negatives we could get a TPR: 89% and TNR: 47%.

Sensitivity of the system can be tuned to different applications (Section 4.8). An automatic floating window generator might be optimized for high TPR, in order not to miss to many disturbing violations, at the cost of correcting some of those that are not disturbing. A quality assistance system might be optimized more towards TNR, predicting for the user when it is not necessary to intervene with high reliability.

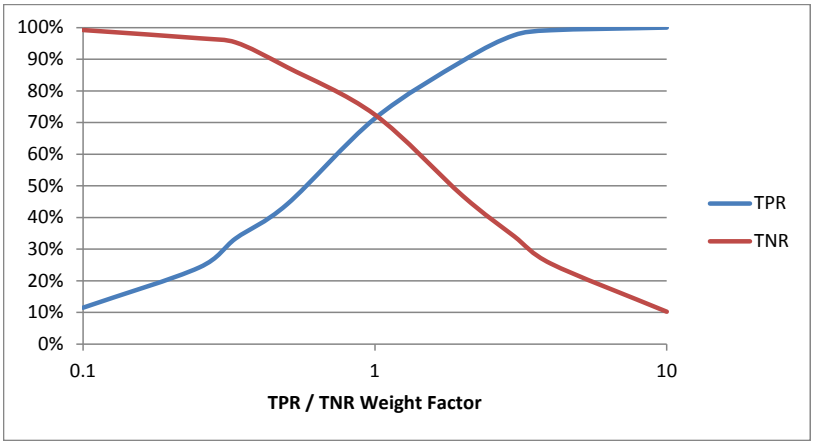


Figure 4.9: *Resulting TPR and TNR after varying their respective importance.*

4.7 Results

Figure 4.10 illustrates our results by providing anaglyph image as well as visualization of the ground truth and model prediction. The middle column

(Raw data) uses the JET colormap to represent the level of subject agreement in labelling window violations. Dark red shows greater agreement. The same colormap is used to represent the model prediction of a disturbing window violation. Dark red shows a higher probability disturbing. The right column shows the comparison of the thresholded ground truth data and thresholded model prediction. The threshold for the ground truth is set to 60% agreement, while the threshold used for the model is set to the value obtained from the cross-validation: 0.4394. The regions in red represent positives and blue negatives.

Our model captures most of the problematic areas without labelling those that are not. Figure 4.10-a shows how our model correctly labeled the hanging lampshade to be disturbing, however, it incorrectly labeled a portion of the desk as problematic. Figure 4.10-b is correctly predicted to not be problematic. This was accomplished by reducing texture detail in both the lampshade and desk. Similarly, Figure 4.10-c and 4.10-d show how disturbing window violations can be removed using depth-of-field. Contrast is reduced in the regions farther from the focal point. Both the thresholded ground truth and model prediction data agree the disturbing window violation is removed.

4.7.1 Limitations

Perceptual models by their nature are limited to certain number of factors. One of the factors that we didn't include was "feature proximity". This happens when strong stereoscopic features (visible to both eyes) meet two conditions: (1) they are spatially near the violation and (2) they have negative disparity similar to the window violation region. These strong features can dominate the stereo matching process increasing the likelihood a disturbing window violation is perceived. Luminance of the image border is also not handled by our model. Our experimental data does not explore the effect of matching the luminance of both the window violation and image border. Our model also does not consider the influence of disparity frequency.

4.8 Applications

We present two applications of our model: visualization of where disturbing window violations occur and the automatic removal of disturbing window

violations using floating windows. The stereo-matching algorithm developed by Werlberger [2010] is used to identify regions of undefined disparity near the lateral borders.

4.8.1 Visualization

Our visualization application is intended to support stereoscopic content producers in the detection of disturbing window violations. We adopt a technique commonly used in camera systems called zebra patterning. It is used to represent regions of an image that are overexposed. A threshold value is set by the user to choose how close to overexposure a region is before the zebra pattern is made visible.

Visualization is provided in two modes. First, it identifies images containing disturbing window violations. Second, it provides a visual representation of the disturbing region by displaying an animated zebra pattern as shown in Figure 4.11. A user parameter is available to adjust sensitivity to disturbing window violations.

Figure 4.11: *Disturbing window violations are visualized by a zebra pattern.*

This visualization could be included in a computational stereoscopic camera system [Heinzle et al., 2011] or stereoscopic analyzers [Zilly et al., 2010; CelSoft, 2012] to provide real-time detection of disturbing window violations. It could also be applied in software rendering tools or as a quality assurance step before releasing stereoscopic content.

4.8.2 Automatic Floating Window Generation

Floating windows (see section 4.2) provide a good solution to resolve window violations. However, it can be difficult to determine how to apply them, especially during live-capture. Our prediction can help reduce the uncertainty. Figure 4.12 shows two results of our automatic floating window generator: one requires the use of floating windows and the other does not. Because our system can localize regions that are disturbing, we can also automate more elaborate crops, such as slanted floating windows, to preserve pixels not in window violation.

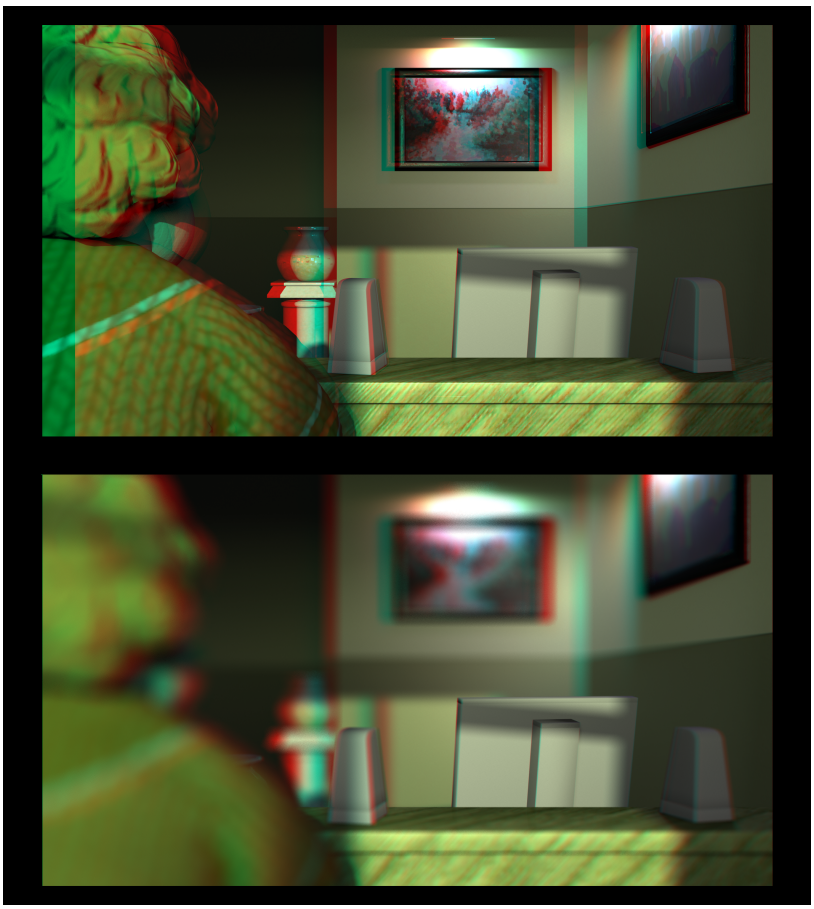


Figure 4.12: *Automatic floating windows. (a) The woman's hair and sweater are predicted to be disturbing and corrected with a floating window. (b) Depth of focus blur reduces the conflict and it not predicted to be problematic. No floating windows are applied.*

4.9 Conclusion

We have demonstrated the development and application of a computational model for the perception of stereoscopic window violations. We presented a method of measuring window violation preference as a function of contrast and spatial frequency. Our data fits the expectation of a CSF-like sensitivity function for stereopsis. The model was calibrated and validated using viewer input from real stereoscopic images. It can successfully detect user-labeled disturbing window violations. We present two important applications of the model: visualization and automatic floating window correction of disturbing window violations.

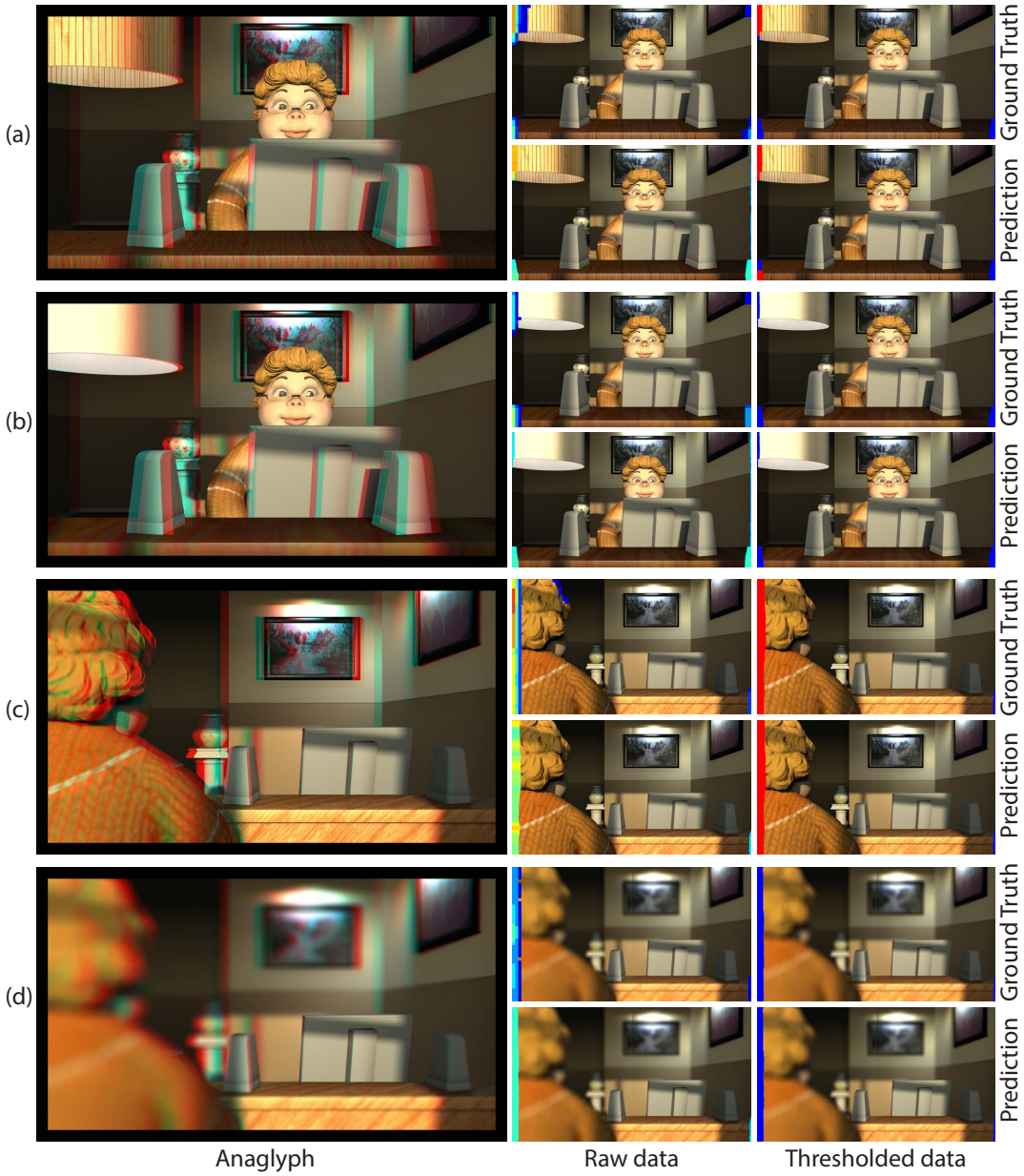


Figure 4.10: *Illustration of results: anaglyph image, ground truth and model prediction. Raw data column uses JET colormap to represent level of agreement (Ground Truth) or probability disturbing (Prediction). Thresholded data represent result of binary threshold: Red is positive (disturbing) and blue is negative (not disturbing).*

CHAPTER 5

Multimodal Stereoscopic Saliency

Since reasoning about depth interpretation is dependent on eye fixation, it is important to have a means to predict eye fixation locations. This chapter presents research combining multiple saliency modalities to produce a prediction of stereoscopically significant salient objects.

5.1 Introduction

The visual saliency estimation problem has been extensively studied by a multitude of disciplines including neurosciences, vision science and computer vision. The seminal work of Koch and Ullman [Koch and Ullman, 1985] asserted that a saliency map can be generated by combining a number of elementary, pre-attentive visual features (such as color, orientation, movement and disparity) in a winner-take-all network. This purely theoretical framework was later implemented by Itti and Koch [Itti and Koch, 2001] who proposed computing center-surround differences of pre-attentive features. Their system consisting of stimuli-driven, *bottom-up* mechanisms accurately described how attention is deployed within the first few hundreds of milliseconds after the presentation of a new scene. However, for longer spans of attention they admit that more sophisticated models with *top-down* mecha-

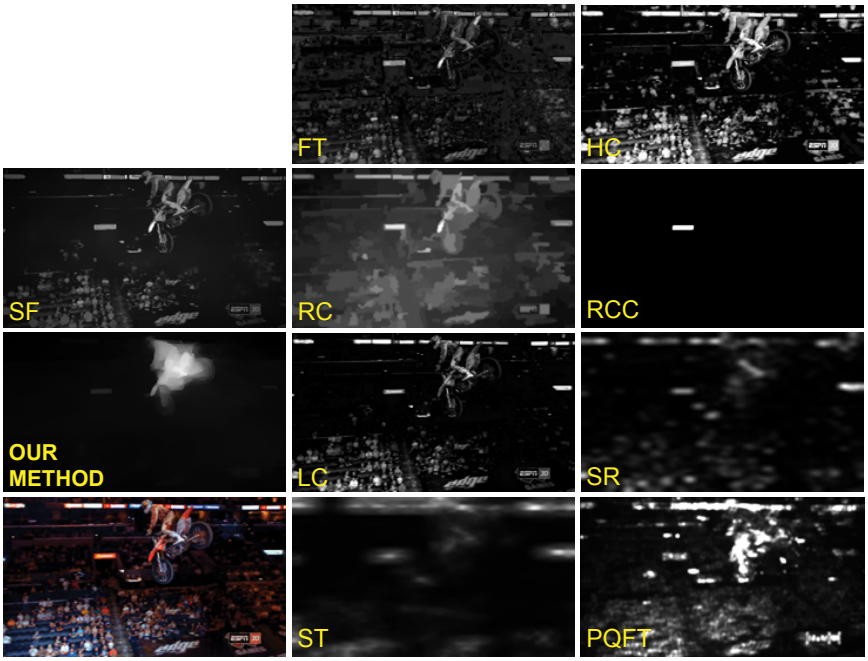


Figure 5.1: *Our method accurately estimates salient objects even in visually cluttered scenes by utilizing motion and disparity, as well as high-level features and low level spatial distribution cues. A comparison to SF [Perazzi et al., 2012], FT [Adams et al., 2010], RC [Cheng et al., 2011], LC [Zhai and Shah, 2006], ST [Seo and Milanfar, 2009], HC [Cheng et al., 2011], RCC [Cheng et al., 2011], SR [Hou and Zhang, 2007] and PQFT [Guo et al., 2008] shows that our method successfully singles out the motorbike as the salient object where other methods fail. Note also the high edge-accuracy of our result near the salient object.*

nisms accounting for volitional biasing are required. Since then, their formulation of complementary bottom-up and top-down mechanisms as well as the center-surround differences have been widely adopted by other researchers.

Depending on the application, saliency estimation techniques have different goals. Some approaches try to predict the most probable fixation points of a human observer in the scene. This line of research helps us to understand how short term visual attention is deployed, and the fixation point predictions can be used for globally applied image operators [Ancuti et al., 2011], or to reduce the search space for computer vision algorithms. Alternately, many image processing applications such as retargeting, collages [Goferman et al., 2010], classification, and object of interest localization [Gao et al., 2009],

require saliency masks that mark specific *objects* as important, rather than fixation points. We present a multi-modal fusion technique that combines together different information sources into a final, edge aligned and temporally consistent stereoscopic saliency estimation.

Working on stereoscopic video presents a number of challenges and opportunities. Additional modalities such as motion and stereo disparity contain valuable saliency cues that can be exploited for more accurate prediction. However, each new modality significantly increases the problem complexity, requiring extensions to computational saliency models. For example, when considering videos rather than images, temporal stability becomes important. Applications such as video retargeting [Krähenbühl et al., 2009], video summarization [Lee et al., 2012], activity recognition [Vig et al., 2012] and perceptual video coding [Lee and Ebrahimi, 2012] all stand to directly benefit from a temporally consistent spatiotemporal saliency estimation. Similarly, accurate saliency estimation for stereoscopic content is a crucial part of automatic content creation for stereoscopic displays [Stefanoski et al., 2013], video retargeting [Basha et al., 2011], and disparity editing [Koppal et al., 2011; Lang et al., 2010]. We extend recent image-based state-of-the art practices to consider spatiotemporal information and stereoscopic disparity, at the same time adding top-down saliency cues as well. As a component of our saliency framework we also discuss a novel approach to computing motion saliency. Our approach fuses together these modalities and leverages a recent edge-aware spatiotemporal video volume filtering approach to generate temporally stable, edge aligned results.

Quantitative evaluation of stereoscopic video is a significant challenge due to the difficulty of obtaining ground-truth data sets. As a second main contribution, we have created and made available for public use, a diverse ground-truth data set of eye track data of stereoscopic video. We use this data set to perform a quantitative evaluation of our method, and compare the results to other existing image saliency approaches. Finally, we demonstrate applications of our method to automatic view synthesis.

5.2 Related Work

Numerous saliency estimation methods have been proposed for monoscopic images. Among them, some bottom-up models are partially influenced by the mechanisms of the human visual system [Itti et al., 1998; Itti and Koch, 2001; Murray et al., 2011]. Other methods use statistics about color or patch distributions in the image to determine significant regions [Goferman et al.,

2010]. Similarly, Perazzi et al. [2012] utilizes two global measures, uniqueness and distribution for estimating saliency. Other researchers have shown that fixation points can also be predicted by top-down models utilizing high-level features such as distance and dissimilarity between image patches and a center bias [Duan et al., 2011]. Our method considers additional modalities of information, such as spatiotemporal cues and stereoscopic disparity.

Including temporal information for video saliency has been proposed by prior work as well. Lang et al. [2012b] computes saliency for each frame of the input video and later applies a spatiotemporal filter to achieve temporal consistency. However, this approach does not take into account the spatiotemporal and disparity related aspects of saliency. Itti and Dhavale [2003] proposed a complete spatiotemporal framework by extending earlier work on image saliency [1998] with additional center-surround mechanisms for flicker and motion. Since then, more methods have been proposed that either only compute motion saliency [Cui et al., 2009; Belardinelli et al., 2009] or spatiotemporal saliency [Rapantzikos et al., 2009; Mahadevan and Vasconcelos, 2010]. We present a novel method for motion saliency estimation. In addition to the aforementioned methods, we utilize stereoscopic information and perform an evaluation in comparison to ground truth eye track video data.

Recent findings on stereoscopic saliency suggest accounting for disparity is crucial for saliency estimation in stereoscopic images [Niu et al., 2012]. Similarly, stereoscopic information has been included to spatial saliency maps by several works [Lang et al., 2010, 2012a]. In Section 5.4.2 we show that accounting for motion *and* disparity modalities significantly improves the saliency estimation for stereoscopic videos. Furthermore, while some previous work describing human visual system motivated saliency models has mentioned the significance of spatiotemporal stereoscopic cues for saliency estimation [Jeong et al., 2008; Fernandez-Caballero et al., 2008], they lack quantitative evaluation and the edge-accurate results required by many applications. Such quantitative evaluation is a challenging task due to a lack of available ground-truth data sets. While multiple fixation data sets have been used for validation on images [Borji et al., 2012], there are fewer sets for monoscopic videos [Dorr et al., 2010], and to our knowledge none for stereoscopic video. We create such a ground-truth fixation data set for stereoscopic video sequences, and use it to validate our method (Section 5.4).

5.3 Saliency Estimation

In this section we discuss the spatial (Section 5.3.1), motion (Section 5.3.2), stereoscopic (Section 2.3.1) and high-level components (Section 5.3.4) of our method. The final saliency estimation is a spatiotemporally filtered weighted average of each component's outcome (Section 5.3.5). The computation of each of these components presents a significant challenge. To that end we take advantage of prior art and utilize concepts that have been shown to work well, such as distribution and uniqueness [Perazzi et al., 2012], and the "comfort zone" and "popping out" rules for disparity saliency [Niu et al., 2012].

Our method takes the left and right views of a stereoscopic video as input, although we utilize only the right view V in all computational steps except disparity estimation. As such, all presented saliency estimations are aligned with the right view. The data flow of our method is illustrated in Figure 5.2. For ease of reference we also provide a list of symbols in Table 5.1. In the next section we discuss the spatial saliency computation.

5.3.1 Spatial Saliency

Our spatial saliency component is a weighted combination of the distribution and uniqueness measures proposed by Perazzi et al. [2012]. We start by computing $\hat{a} = 500$ superpixels for each frame V^t of the input video using the SLIC algorithm [Achanta et al., 2010] with the modifications by Perazzi et al. [2012]. Next, a three dimensional vector $S^{t,i} = [L^{t,i}\ a^{t,i}\ b^{t,i}]$ is extracted from each superpixel i at frame t, that contains the mean values of the luminance and chroma channels. The uniqueness of superpixel $S^{t,i}$ is defined as follows:

$$U^{t,i} = \sum_{j=1}^{\hat{a}} \|S^{t,i} - S^{t,j}\|^2 \cdot w(p^i, p^j), \tag{5.1}$$

where w is a local Gaussian weighting function, p_i and p_j are positions of the superpixels i and j. The distribution measure is expressed in a similar form:

$$D^{t,i} = \sum_{j=1}^{\hat{a}} \|p_j - \mu_i\|^2 \cdot w(S^{t,i}, S^{t,j}), \tag{5.2}$$

where $w(S^{t,i}, S^{t,j})$ represents the similarity of the vectors $S^{t,i}$ and $S^{t,j}$, and $\mu_i = \sum |S^t|_{j=1}\ w(S^{t,i}, S^{t,j})\ p_j$ is the weighted mean position of $S^{t,i}$ in the color space. Similar to the original method, both the element uniqueness

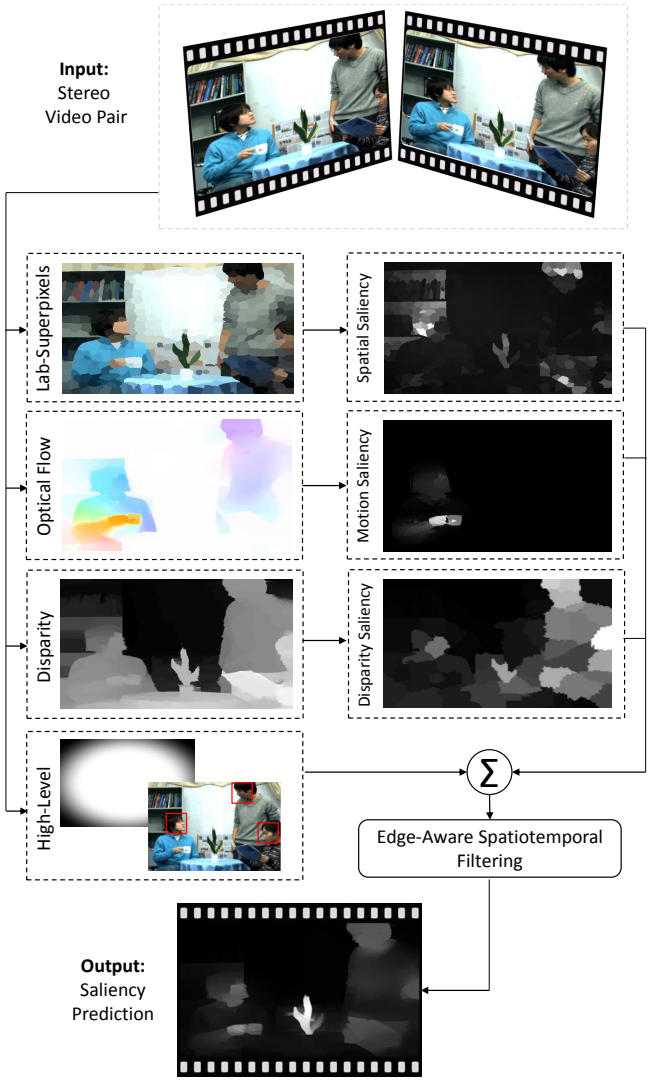


Figure 5.2: *The data flow of our method. See text for details.*

and distribution measures are efficiently computed using permutohedral lattices [Adams et al., 2010]. The spatiotemporal saliency of each superpixel $S^{t,i}$ is obtained by combining the two measures:

$$\Psi_{st}^{t,i} = U^{t,i} \cdot exp\Big(-\hat{b} \cdot D^{t,i}\Big), \tag{5.3}$$

where the model parameter $\hat{b} = 6$ adjusts the significance of the distribution measure with respect to the uniqueness measure. Unlike the original method that applies a bilateral filter to the outcome of equation 5.3, we do not perform any additional processing after this point. Instead, a spatiotemporal filtering is performed as the final step after we combine all saliency components (Section 5.3.5).

5.3.2 Motion Saliency

Motion has a strong influence on saliency in videos. We start by computing the optical flow estimate of V computed using the method by Lang et al. [2012b]. Our novel motion saliency estimation relies on the application of the uniqueness and distribution concepts to optical flow vectors. To that end we recompute equations 5.1 and 5.2 with the two dimensional optical flow vectors at each pixel rather than the average color coordinates at superpixels. We also similarly combine the two measures using equation 5.3 with the same $\hat{b}$ parameter as in the spatial case. Figure 5.3 shows an example of the steps of our motion saliency computation. Note how our final motion saliency estimation (d) estimates the motorbike's motion as salient and isolates it from the camera motion in the background (a).

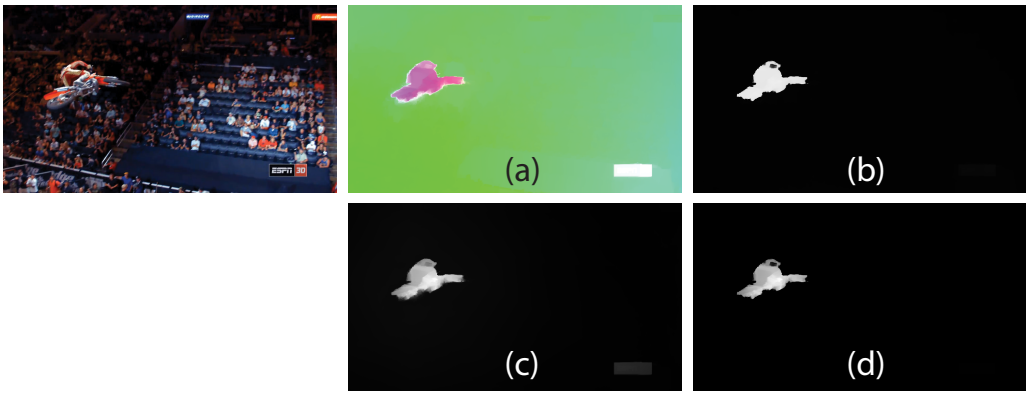


Figure 5.3: *An example optical flow estimation visualized using Middlebury color coding (a), and the corresponding distribution (b) and uniqueness (c) that leads to our final motion saliency estimation (d).*

5.3.3 Disparity Saliency

Stereoscopic disparity is another source of visual information that can help estimating saliency. Our method utilizes the disparity contrast measure from Niu et al. [2012] that takes the abruptness of disparity change over image regions into account. To that end we consider the previously computed superpixels (Section 5.3.1) as image regions instead of utilizing a graph-based segmentation proposed in the original method. Given the disparity $\phi^{t,x}$ at pixel x of frame t, the disparity contrast between any pair of superpixels is

defined as follows:

$$\begin{aligned} \delta(S^{t,i}, S^{t,j}) &= \frac{1}{n^{t,i}\, n^{t,j}} \sum_{p=1}^{n^{t,i}} \sum_{q=1}^{n^{t,j}} f(p,q), \quad \text{and} \\ f\,(p,q) &= w\,(p,q) \cdot \left| \phi^{t,p} - \phi^{t,q} \right|. \end{aligned} \tag{5.4}$$

The Gaussian weight function $w(p,q)$ is defined as $exp(-\|p-q\|^2/\hat{d})$ on normalized image coordinates p and q. The variance of the weighting function $\hat{c}$ is set to the default value 0.4.

Niu et al. [2012] proposes the two saliency rules obtained from domain knowledge on stereoscopic perception, both of which can be implemented given the disparity contrast δ. The first rule states that *objects with small disparity magnitudes tend to be salient*. In practice this rule assigns higher weight to objects well within the stereoscopic comfort zone, and is expressed with the following formula:

$$R_1^{t,i} = \begin{cases} \left(\delta_{max}^t - \delta^{t,i}\right) / \delta_{max}^t & \text{if } \delta^{t,i} \geq 0, \\ \left(\delta_{min}^t - \delta^{t,i}\right) / \delta_{min}^t & \text{if } \delta^{t,i} < 0, \end{cases} \tag{5.5}$$

where the δ_{max}^t and δ_{min}^t denote the maximum and minimum disparity contrast values at frame t, and $\delta_{t,i}$ denotes the mean disparity contrast of superpixel i at frame t. The second rule states that *objects that pop out of the screen tend to be salient*, and is computed as follows:

$$R_2^{t,i} = \frac{\delta_{max}^t - \delta^{t,i}}{\delta_{max}^t - \delta_{min}^t}. \tag{5.6}$$

Our disparity saliency estimation is a weighted combination of these two rules:

$$\Psi_d^{t,i} = (1-\lambda) \cdot R_1^{t,i} + \lambda \cdot R_2^{t,i}, \tag{5.7}$$

where $\lambda = \gamma + (1-\gamma) \cdot n_{neg}^t / n^t$, such that n_{neg}^t denotes the number of pixels with negative disparity, and n^t is the total number of pixels of frame t.

5.3.4 High-Level Features

Previous research on top-down visual saliency has shown the importance of high-level visual features in saliency estimation. To that end we utilize a saliency measure for faces (Ψ_f^t), that consists of an outcome of a face detector for each frame. Each face region of frame t is marked with 1 in Ψ_f^t, whereas all remaining regions are marked with 0. Similarly, subjective studies have

Intermediate features			
V^t	Frame t of the input video's left view	$S^{t,i}$	3D vector representing superpixel t, i
$\phi^{t,i}$	Stereoscopic disparity at pixel t, i	$\delta^{t,i}$	Mean disparity contrast of the superpixel t, i
n^t	Size of frame t in pixels	$n^{t,i}$	Size of superpixel t, i
$U^{t,i}$	Uniqueness of $S^{t,i}$	$D^{t,i}$	Distribution of $S^{t,i}$
R_1^t	"Comfort zone" rule factor for superpixel t, i	$R_2^{t,i}$	"Popping out" rule factor for superpixel t, i
Parameters			
$\hat{a}$	Superpixel number per frame	$\hat{b}$	Significance of D w.r.t. U
$\hat{c}$	Disparity contrast parameter	$w_{s\|m\|d\|f}$	Saliency weights
Saliency maps			
$\Psi_{s\|m\|d\|f\|c}$	Spatial, motion, disparity and face saliency maps and the center bias		

Table 5.1: *Summary of symbols used.*

shown that people are more likely to perceive scene elements as being salient if they are located near the center of the video frame rather than the periphery [Judd et al., 2009].

We model this center-bias by applying a weighting function that resembles the mesa filter [Watson, 1987]. The mesa filter consists of a flat pass-band starting form the center until the 2/3 of the video frame, followed by a transition region characterized by the Hanning window, and a flat stop-band region near the corners:

$$\Psi_c^t = \begin{cases} 1 & \text{if} \quad \rho < \frac{2}{3}, \\ 0 & \text{if} \quad \rho > \frac{4}{3}, \\ \frac{1}{2}\left(1 + cos(\pi\left(\frac{3}{2}\rho - 1\right)\right) & \text{otherwise}, \end{cases} \tag{5.8}$$

where $\rho = \sqrt{x^2 + y^2}$, and $x, y \in \{-1, 1\}$ denote the normalized image coordinates such that the origin lies at the center of the video frame (Figure 5.2 - high-level features).

5.3.5 Multimodal Saliency Fusion

While intuitively it is clear that all the spatial, motion, disparity and high-level cues we discussed in this section have *some* effect on the final saliency estimation, it is challenging to formulate the exact relationship among the individual saliency components. To that end we make use of a diverse ground-truth data set consisting of eye tracking data for stereoscopic videos. Details on the data set and our experimental procedure are discussed later in Section 5.4.

We assume that the final saliency estimation is a weighted linear combination of all saliency components. Given the spatial, motion, disparity and face saliency maps for a video sequence, along with a binary eye-tracking map where recorded gaze points are marked at each frame, we build a linear system $Ax = b$. The $n \times 4$ matrix A comprises spatial, motion, disparity and face saliencies for all the n pixels in all frames of the input video. The $n \times 1$ vector b contains binary ground-truth eye-tracking information. The value of b is equal to 1 at the recorded gaze coordinates, and is 0 otherwise. We obtain the final weights w_s, w_m, w_d, w_f for the saliency components by solving the linear system for x for each video sequence and then averaging the intermediate weights. The fused multimodal saliency estimation is generated by combining saliency components, multiplying each frame of the outcome with the center-bias Ψ_c^t (Equation 5.8), normalizing the fused saliency estimation, and finally applying an edge-aware spatiotemporal diffusion process proposed by Lang et al. [2012b].

5.3.6 Results

The results presented in Figure 5.4 demonstrate the interplay between different components of our saliency estimation[1]. A significant feature of our method its capability of generating edge-precise and temporally coherent results even in visually cluttered scenes. We first will perform a qualitative comparison given in Figure 5.4 (as well as in supplemental video). In the next section, we describe a quantitative comparison for our data set. The **Street** scene starts with the gray car moving forward (1), while the dark red car is waiting to make a left turn (2). Consequently the gray car is detected as salient in the first frame due to its motion. The dark red car becomes salient as soon as it starts making the left turn (3). The final frame shows another car passing through and being detected as salient (4). Note that in all the frames, the no-parking sign is also detected as a weakly salient object due to the strong color contrast.

Similarly, the first two frames of the **Dino** scene also show the spatial and motion components working together. However, in the last frame where the ball rolls towards the camera (4) the disparity component significantly increases the saliency of the ball, making it the most salient object in the scene.

Saliency estimation for the **Balloons** scene is challenging due to the many salient objects competing for the viewer's attention. The last two frames in this set where the actor's face becomes visible (3, 4) show the effect of the

[1] The entire set of results are presented in supplemental video.

face saliency component. The results of these frames show that even in such cluttered scenes the face saliency dominates the final saliency estimation.

In addition to the results we presented, we also generated a ground-truth data set and evaluated the performance of our method, which is discussed in the following section.

5.4 Subjective Evaluation

A diverse ground-truth data set is essential for making any predictions on a highly complex task such as stereoscopic video saliency. In the absence of a ground-truth data set for stereoscopic video saliency, we performed an eye tracking experiment and collected fixation data for a diverse set of stereoscopic videos (Section 5.4.1). We used this data set to evaluate the performance of our method (Section 5.4.2), and make it available for future research and validation.

5.4.1 Experiment Setup and Execution

Stimuli: we prepared a diverse test set of 15 video clips that comprises both real-world and computer generated scenes. The resolution of the video clips in our test set varies from 960×768 to full HD, and their frame rate varies between 24 and 30. The total number of frames in our test set is more than 3000[2]. The video sequences have both human and non-human salient scene elements at various disparity ranges, with various types of movements, located both in dark and bright scene regions as well as near or far away from the center of the video frame. We also made an effort to include videos with various levels of saliency prediction difficulties: ranging from simpler scenes with a single, clear-cut salient object to more difficult scenes with numerous salient objects.

[2]See supplemental material for detailed statistics.

Figure 5.5: *A picture of our eye tracking experiment setup.*

Setup: we performed an experiment where we obtained eye tracking data for all the video sequences in our test set. Our experiment setup consisted of an EyeLink II head mounted eye tracker (SR Research) and a 55″ row interleaved stereo display. The subjects were placed at 2.05 meters away from the display. Figure 5.5 shows a picture of our setup during one of the trials.

Procedure: all subjects were asked to do nothing but watch the video sequences that were presented in random order. Each subject performed 30 trials where each video sequence was viewed twice. There were no time limitations, and the subjects were free to wait as long as they wanted before proceeding to the next trial. On average, each subject required an hour to finish the study. The experiment was performed on 10 paid subjects, 3 males and 7 females within the age range of 22 to 29. Subjects were confirmed to have good stereo acuity prior to taking part in the study.

As the result of this experiment we obtained horizontal and vertical coordinates of each viewer's gaze points for our entire data set. The gaze points were used as the ground-truth for the performance evaluation of our method as discussed in the next section.

5.4.2 Performance Evaluation

For performance evaluation, we computed the ROC curves of our method for the entire ground-truth set discussed in Section 5.4.1. We treated the gaze

points and their 8 immediate neighbors as the positive data, and computed true positives and false positives accordingly using 100 threshold values uniformly sampled from the range of saliency values. In order to separate the test and training data we utilized leave-one-out cross-validation. The resulting ROC curves are shown in Figure 5.6-top, along with the results of prior art for comparison. We generated the results of other authors' methods using publicly available code.

As our performance measure we compute the commonly used AUC score [Borji et al., 2012] for all the frames in our data set and present the results in Figure 5.6-bottom. Our method achieves the median AUC score of 0.84 using linear fusion, which is significantly better than the closest performers RC and FT at 0.78. The spatial component of our method (S) alone achieves 0.74, which suggests that the additional performance of our method is due to the combined use of additional modalities with the spatial component. In addition to the quantitative evaluation, Figure 5.4 demonstrates the visual quality of our saliency maps. As Figure 5.1 shows, our method is significantly better in producing edge-accurate maps of salient objects compared to prior art.

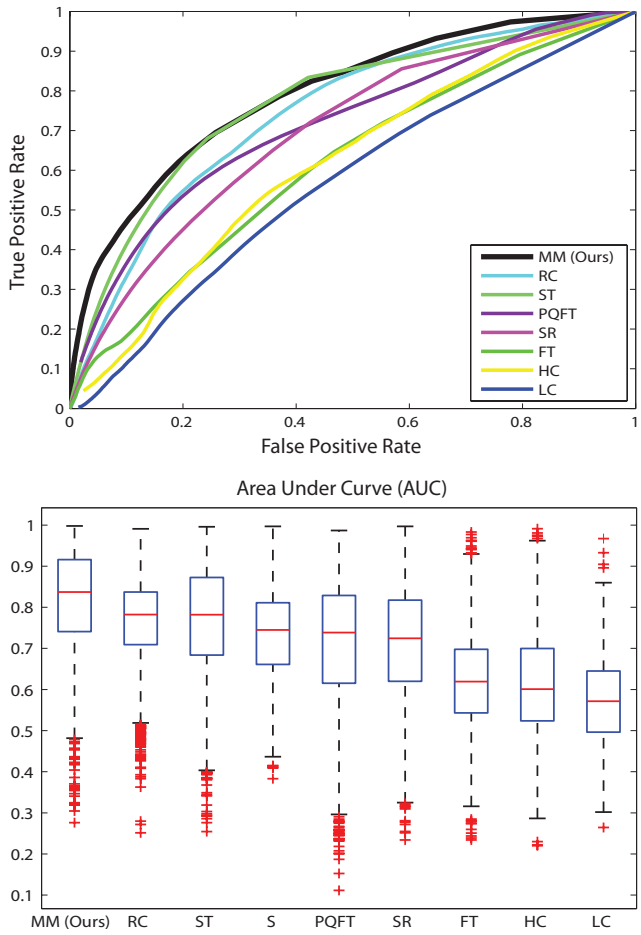


Figure 5.6: *The ROC curves (top) and AUC measures (bottom) of our method with linear (Lin) and SVM fusion compared to RC [Cheng et al., 2011], ST [Seo and Milanfar, 2009], our method's spatial component (S), PQFT [Guo et al., 2008], SR [Hou and Zhang, 2007], FT [Adams et al., 2010], HC [Cheng et al., 2011], LC [Zhai and Shah, 2006]. The plot shows results for the entire evaluation set consisting of 15 videos and nearly 3000 frames. In the box plot, red lines are median values, blue boxes show the variance and the red crosses denote outlier AUC scores.*

5.5 Applications

Our multi-modal stereoscopic saliency method labels specific objects or regions as important. This information can support applications that reason about the location of important objects. Our results are spatial-temporally smooth following objects as they are salient during a sequence.

Figure 5.7: *Improving automatic view synthesis. The visualization (c) shows the differences (in red color) between the same view generated with our saliency result (b) and the saliency from the original method (a). The actual views are presented in the supplemental video.*

Saliency is a fundamental problem of visual computing and as such has many applications. As we discussed in the introduction a better saliency estimation directly improves results of applications such as video retargeting, video summarization, activity recognition, and video coding. Furthermore, a precise saliency estimation for stereoscopic content enables several practical applications. Disparity editing [Koppal et al., 2011; Lang et al., 2010], including reconvergence and remapping of stereoscopic depth, can be guided by knowledge of where important objects are composed within a scene. It has been stated that a stereoscopic saliency would be highly beneficial for spatial retargeting of stereoscopic content [Basha et al., 2011]. Also for stereoscopic comfort analysis, an important consideration is where the viewer will fixate [Shibata et al., 2011]. Finally, automatic content creation for autostereoscopic displays is another application. In the following paragraph, we show an example application for automatic view synthesis.

Automatic View Synethesis: We provide an example application that improves the generation of synthetic views from stereoscopic video. This approach is used to process traditional stereoscopic content for presentation on multi-view autostereoscopic displays. Since our method is automatic, we extended an automatic view synthesis method of Stefanoski, et al. [Stefanoski et al., 2013] to improve view synthesis in regions lacking sufficient image feature matches. Their method utilizes saliency as a stiffness constraint in computing the optimization of generating interpolated or extrapolated stereoscopic views. Their spatio-temporal saliency was computed using phase spectrum of quaternion Fourier transform [Guo et al., 2008], referred to as PQFT, which they combined with a traditional edge map. We replace the original saliency, PQFT, with our own, but keep the edge map to preserve the

constraint on image edges. As can be seen in Figure 5.7, our saliency method labels the entire elephant as salient and reduces the warping-based image distortion in the elephant's face.

5.6 Limitations

While our method accounts for a number of visual saliency components, our linear model of how they affect the final saliency estimate is the main limitation of our method. An example of where such a model fails is shown in Figure 5.8. Given the input frame (a), the output of our face saliency component (b) shows the faces of the audience. Even though the saliency of these faces at the positive disparity region are somewhat reduced by our disparity saliency component, they are still estimated as salient regions in the final result (c). While the assumption that human faces are highly salient is true in most cases, in this example where the fighters are clearly the center of attention, it leads to erroneous saliency estimations (eye tracking data shown as red dots). Such complex interactions are beyond the capabilities of our method, but we hope to stimulate further research on the topic by making the ground-truth data set public.

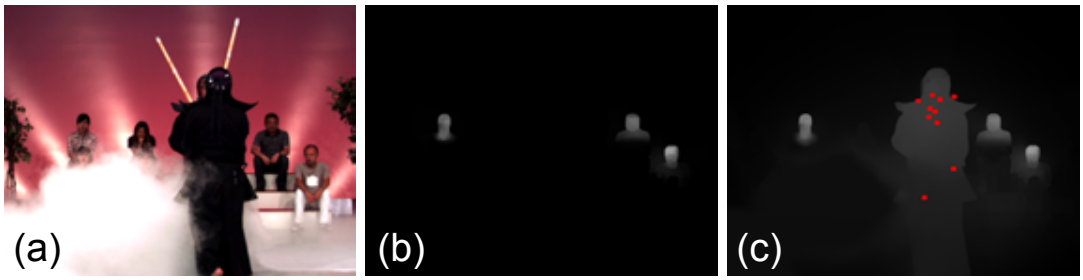


Figure 5.8: *Despite the common-sense notion, human faces are not always the most salient objects. See text for the discussion.*

5.7 Conclusion

We presented a novel multimodal saliency estimation method for stereoscopic video that utilizes spatial, motion, and disparity cues as well as face detection and a center prior. The major contributions of this work are (i) a diverse ground-truth data set of gaze points for stereoscopic video, and (ii) a method for computing temporally coherent and edge-precise saliency maps,

that is trained using this data and comprises state-of-the-art methods along with a novel motion saliency estimation. We presented both qualitative and quantitative comparisons to prior art and showed that our method achieves better performance compared to other saliency methods. As an example application, we demonstrated that automatic multi-view synthesis can notably benefit from using our saliency method.

The exploration of more sophisticated models of the complex interactions between the saliency components is an interesting future direction of this work. We also believe that the performance of our method can be further improved by more accurate optical flow and disparity estimations, as well as better face detection.

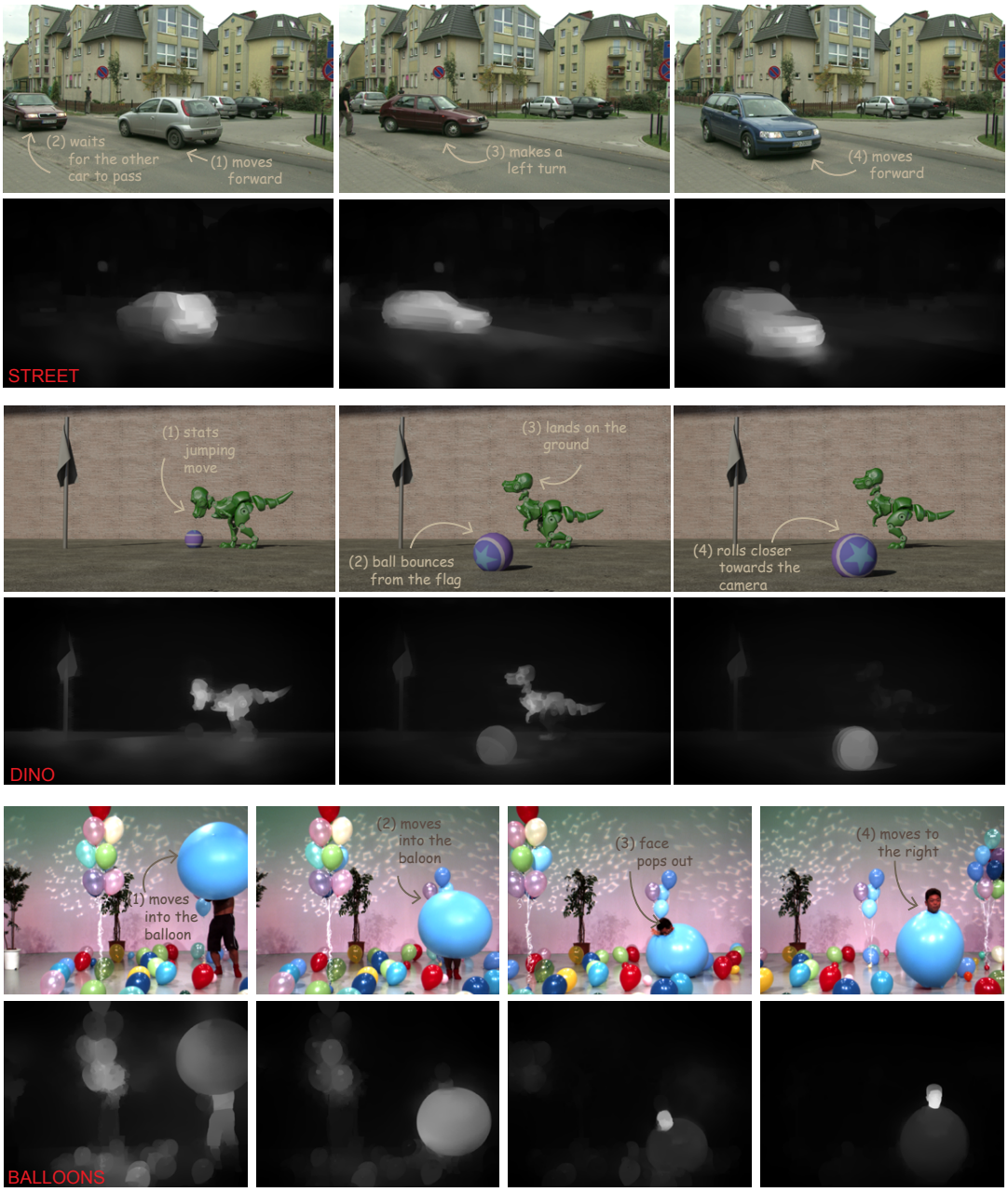


Figure 5.4: *Saliency predictions (bottom rows) for selected representative frames of the **Street**, **Dino** and **Balloons** scenes (top rows) computed using the linear fusion. Higher gray values of the saliency map indicate higher saliency. Each video frame is annotated with the significant events that affect saliency computation. See text for further discussion of the results.*

CHAPTER 6

Conclusion

The evaluation and visualization of stereoscopic content is an important and challenging goal. This thesis presents an exploration of computational modeling as it relates to the perception of stereoscopic image viewing. The findings represent an integration of existing computational models and creation of new theories, representations and implementations of stereoscopic perception. Several contributions are summarized below.

6.1 Key Results

The high level goal of the thesis is to explore the complexity involved in modeling topics related to stereoscopic perception. There is a need to develop technologies to facilitate understanding of how stereoscopic images are perceived. This thesis presents a combination of perceptual models based on existing research and novel experimental methods. The key results support this challenge to develop models of stereoscopic image quality.

Existing perceptual models have been utilized to understand the space of limitations influencing both the presentation and perception of stereoscopic images. This is represented in the structured overview of relevant topics and significant research in the modeling of perception of stereoscopic images.

We have experimentally observed that changes to scene composition can improve visual performance viewing stereoscopic images. Specifically, the application of continuous stereoscopic depth can reduce the time to change visual attention.

We have demonstrated how a stereoscopic window violation, caused by a conflict between the occlusion and stereopsis depth cues, is not always problematic. And we have demonstrated that is possible to predict such situations, providing useful input to stereoscopic content creators.

Finally, we combine the fusion of multiple saliency modalities with edge-aware, spatio-temporally smooth saliency representations to produce better results than other state of the art methods. This is useful for a variety of stereoscopic editing applications.

6.2 Summary of Technical Results

Modeling Topics in Stereoscopic Imaging provides a structured overview of relevant topics, significant research and computational models. Details of these models are provided to describe how they are implemented to support computational analysis of stereoscopic perception. The overview is grouped to represent topics of depth interpretation and stereopsis. Vergence, accommodation and visual comfort is then discussed as it is a dominant factor influencing depth interpretation and visual comfort. Several important stereoscopic distortions are presented followed by a discussion of visual attention. These models have been applied in various forms during the thesis, for example, guiding the capture and display of stereoscopic content for experiments as well as the production of stereoscopic movies.

Attention Transitions in Stereoscopic Depth represents research in developing a computational theory about the influence of stereoscopic scene composition on the speed of visual attention transitions. We have demonstrated how a change in scene composition to provide depth continuity significantly influences a viewer's ability to change visual attention among spatially distinct scene elements. We made several significant observations. First, visual attention transitions to the zero parallax plane are not significantly improved by depth continuity. This is explained by the assumption that the transition target is not presenting the vergence-accommodation conflict. Second, lateral attention transitions (N-N and F-F) excluding the zero parallax plane are significantly improved by continuous depth. Lastly, the

inward (Z-N) and outward (Z-F and N-F) attention transitions were significantly improved by the continuous depth plane. The outward depth changes, Z-F and N-F, resulted in 18% and 20% reductions in response time and provided an approximately 300ms improvement. Finally, we also observed that self-assessment of eye fatigue and general discomfort increase before a decrease before a decrease in visual performance is observed. The findings provide a significant example of how stereoscopic 3D content creators may learn scene composition, framing and montage from visual psychophysics.

Stereoscopic Window Violations induce visual discomfort due to the conflict between two cues to depth: stereopsis and occlusion. We have isolated the following four dominant factors that influence the detection of disturbing window violations: contrast magnitude, spatial frequency, orientation and disparity. We conducted subjective experiments to evaluate the influence of different variables on the perception of disturbing window violations. Window violations with increasing contrast magnitude cause a monotonically increasing perception of a disturbing window violation. The spatial frequency factor resembles the well-known contrast sensitivity function. Spatial frequencies closer to the peak sensitivity of human vision are more disturbing. Horizontally oriented information was observed to be less disturbing than other orientations. Increasing disparity magnitude of a window violation was also observed to increase the likelihood of being disturbing.

A perceptual model based on the subjective measurements was developed to predict the detection of problematic window violations. The model is developed by constructing a look-up table, representing the preference observed through pairwise comparison of the contrast magnitude and spatial frequency conditions. Applying the independent channel hypothesis [Marr and Poggio, 1979], we selected the contrast and frequency combination that is rated as most problematic. The value is scaled by the orientation and maximum window violation size (disparity) for each row of pixels.

We demonstrate a method for defining a threshold between disturbing and non-disturbing window violations. A validation experiment was conducted using 95 stereoscopic images, many containing window violations. Human subjects labeled regions of disturbing window violations for each image. We apply our model and vary the model threshold predicting a disturbing window violation. Using cross correlation between the model and labeled data, we find that threshold that provides best agreement.

Our model demonstrates how specific types of window violations are not problematic. We have constructed detectors to automate this assessment.

This enables stereoscopic content creators to maximize the range of comfortable stereoscopic depth while avoiding or resolving only problematic window violations. We describe two applications of the model. One is to visualize the thresholded prediction of disturbing window violations. Another is to automatically remove problematic window violations using floating windows.

Multimodal Stereoscopic Saliency contributions demonstrate how the fusion of multiple modalities generates better saliency results compared to other state of the art methods. To make this claim, we constructed a stereoscopic data set including both animated and live-action content. We then collected eye track data from observations of the data set, which is used to perform a quantitative evaluation of our multimodal saliency method, specifically to define how best to combine the saliency modalities.

We combine four saliency modalities: spatial, motion, disparity and face detection. Our motion saliency is produced by extending the existing spatial modality to compute the uniqueness and distribution of motion features. The disparity modality is based on an existing method combining disparity contrast with domain specific knowledge about depth preferences. Our face saliency modality combines an off-the-shelf face detector with a spatio-temporal diffusion strategy to produce saliency labels covering the faces within the scene. The final combined result is also spatio-temporally smoothed providing an edge-aware and temporally smooth saliency. We demonstrate the better performance of our method using the area under the ROC curve metric.

Our saliencies are edge-aware and temporally smooth to enable many useful forms of stereoscopic image analysis and manipulation, such as the remapping of stereoscopic depth. In such cases, it is beneficial to isolate specific salient objects to manipulate those salient objects. Our saliency metric can also be used to guide the automatic reconstruction of synthetic views, for example to convert stereoscopic images to multi-view image content. Saliency can be used as a stiffness constraint to reduce distortions in salient image contents.

6.3 Future Work

The results and limitations of this thesis lead to two general paths of future work. First, each of the contributions can be expanded. Second, each of the

contributions can be combined in the context of visual saliency. Below, we outline future work grouped by the main contributions of the thesis and then discuss their combination.

Future work of *Attention Transitions in Stereoscopic Depth* can proceed in several ways. One could vary the magnitude of the vergence-accommodation decoupling by evaluating different target depth magnitudes (e.g. reducing the positive disparity) and also by varying the viewing distance while maintaining the same perceived target size. One could explore alternative forms of disparity connectivity. And one could also utilize our methodology of target discrimination and perhaps utilize eye track data to model the time to change visual attention. It would be beneficial to represent such a model as a function of vergence-accommodation conflict as well as depth and screen space separation. Such a contribution would require additional modeling and validation on real image content.

Stereoscopic Window Violations can be improved through future work exploring the influence of additional factors. We currently prepare to conduct new experiments to evaluate the influence of luminance magnitude in addition to contrast magnitude near the border. This may further account for interactions between image content and the perceived stereoscopic window. After enhancing this model, we intend to validate the benefits of using our model to automatically remove window violations using stereoscopic floating windows.

Future work of *Multimodal Stereoscopic Saliency* can proceed in several ways. More accurate estimation of optical flow and disparity will help improve the results. Exploration of additional modalities as well as more sophisticated models of the complex interactions between saliency components is an interesting future direction.

The second general path of future work is the combination of each contribution in the context of visual saliency. An additional saliency modality could be provided by the spatial and temporal connectivity of image contents. Two salient objects that have connectivity may each be more salient than two objects lacking connectivity. This could provide additional stereo saliency metrics to be applied within a given sequence as well as between sequence transitions. Connectivity could be defined in terms of on screen location, disparity and strength of vergence-accommodation conflict. An additional saliency modality could be the influence of disturbing stereoscopic artifacts, such as window violations or stereoscopic ghosting. Do these artifacts attract or repel visual attention and potentially distract the viewer from the intended story?

In summary, this thesis has incorporated many concepts inspired by neurophysiological and psychological observations of binocular perception. The contributions of this thesis represent the integration of additional computational theories to support a better understanding of stereoscopic image perception.

List of Figures

List of Tables

Bibliography

A. Ames, J., Ogle, K. N., and Gliddon, G. H. (1932). Corresponding retinal points, the horopter and size and shape of ocular images. *J. Opt. Soc. Am.*, 22(10):538–572.

Achanta, R., Hemami, S., Estrada, F., and Susstrunk, S. (2009). Frequency-tuned Salient Region Detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 1597 – 1604.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Susstrunk, S. (2010). SLIC Superpixels. *EPFL Technical Report no. 149300*.

Adams, A., Baek, J., and Davis, M. A. (2010). Fast high-dimensional filtering using the permutohedral lattice. *Computer Graphics Forum*, 29(2):753–762.

Akeley, K., Watt, S. J., Girshick, A. R., and Banks, M. S. (2004). A stereo display prototype with multiple focal distances. *ACM Trans. Graph.*, 23(3):804–813.

Ancuti, C. O., Ancuti, C., and Bekaert, P. (2011). Enhancing by saliency-guided decolorization. *IEEE CVPR*, pages 257–264.

Aydin, T. O., Cadík, M., Myszkowski, K., and Seidel, H.-P. (2010). Video quality assessment for computer graphics applications. *ACM Trans. Graph.*, 29:161:1–161:12.

Banks, M. S., Akeley, K., Hoffman, D. M., and Girshick, A. R. (2008). Consequences of incorrect focus cues in stereo displays. *J Soc Inf Disp.*, 24(7).

Banks, M. S., Gepshtein, S., and Landy, M. S. (2004). Why is spatial stereoresolution so low? *The Journal of Neuroscience*, 24(9):2077–2089.

Barten, P. G. J. (1999). *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*. SPIE - The International Society for Optical Engineering.

Basha, T., Moses, Y., and Avidan, S. (2011). Geometrically consistent stereo seam carving. In *ICCV*, pages 1816–1823. IEEE.

Belardinelli, A., Pirri, F., and Carbone, A. (2009). Motion saliency maps from spatiotemporal filtering. *Attention in Cognitive Systems*, pages 112–123.

Bercovitz, J. (1998). Image-side perspective and stereoscopy. In *Stereoscopic Displays and Virtual Reality Systems V*, volume 3295.

Borji, A., Sihite, D. N., and Itti, L. (2012). Salient object detection: a benchmark. *Proc. of ECCV*.

Bradshaw, M. F., Parton, A. D., and Glennerster, A. (2000). The task-dependent use of binocular disparity and motion parallax information. *Vision Research*, 40(27):3725 – 3734.

Bradshaw, M. F. and Rogers, B. J. (1999). Sensitivity to horizontal and vertical corrugations defined by binocular disparity. *Vision Research*, 39(18):3049 – 3056.

Burt, P. and Julesz, B. (1980). A disparity gradient limit for binocular fusion. *Science*, 208(4444):615–617.

Cadik, M., Herzog, R., Mantiuk, R., Myszkowski, K., and Seidel, H.-P. (2012). New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts. In *SIGGRAPH Asia 2012*.

Cel-Soft (2012). Cel-scope3d stereoscopic analyser. Available from http://www.cel-soft.com/celscope3d/.

Chang, C.-H., Liang, C.-K., and Chuang, Y.-Y. (2011). Content-aware display adaptation and interactive editing for stereoscopic images. *Multimedia, IEEE Transactions on*, 13(4):589–601.

Chapiro, A., Heinzle, S., Aydin, T., Poulakos, S., Zwicker, M., Smolic, A., and Gross, M. (2014). Optimizing stereo-to-multiview conversion for autostereoscopic displays. *Computer Graphics Forum, proc. of Eurographics 2014*, 33(2).

Cheng, M. M., Zhang, G. X., Mitra, N. J., Huang, X., and Hu, S. M. (2011). Global contrast based salient region detection. *IEEE CVPR*, pages 409–416.

Cipiloglu, Z., Bulbul, A., and Capin, T. (2010). A framework for enhancing depth perception in computer graphics. In *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, APGV '10, pages 141–148, New York, NY, USA. ACM.

Cormack, L. K., Stevenson, S. B., and Schor, C. M. (1991). Interocular correlation, luminance contrast and cyclopean processing. *Vision Research*, 31(12):2195 – 2207.

Coutant, B. E. and Westheimer, G. (1993). Population distribution of stereoscopic ability. *Ophthalmic and Physiological Optics*, 13(1):3–7.

Cui, X., Liu, Q., and Metaxas, D. (2009). Temporal spectral residual: fast motion saliency detection. *Proceedings of the ACM international Conference on Multimedia*.

Cumming, B. G. and DeAngelis, G. C. (2001). The physiology of stereopsis. *Annual Review of Neuroscience*, 24(1):203.

Cutting, J. E. and Vishton, P. M. (1995). Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Handbook of perception and cognition*, volume 5, chapter Perception, pages 69–117. Academic Press.

Daly, S. (1992). Visible differences predictor: an algorithm for the assessment of image fidelity. *Proc SPIE*, 1666.

Didyk, P., Ritschel, T., Eisemann, E., Myszkowski, K., and Seidel, H.-P. (2011). A perceptual model for disparity. *ACM Trans. on Computer Graphics*, 30(4):96:1–96:10.

Didyk, P., Ritschel, T., Eisemann, E., Myszkowski, K., Seidel, H.-P., and Matusik, W. (2012). A luminance-contrast-aware disparity model and applications. *ACM Trans. on Computer Graphics*, 31(6):184:1–184:10.

Dorr, M., Martinetz, T., Gegenfurtner, K. R., and Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of vision*, 10(10).

Duan, L., Wu, C., Miao, J., Qing, L., and Fu, Y. (2011). Visual saliency detection by spatially weighted dissimilarity. *IEEE CVPR*, pages 473–480.

Egly, R., Driver, J., and Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology*, 123(2):161–177.

Emoto, M., Niida, T., and Okano, F. (2005). Repeated vergence adaptation causes the decline of visual functions in watching stereoscopic television. *Journal of Display Technology*, 1(2):328–340.

Fender, D. and Julesz, B. (1967). Extension of panum's fusional area in binocularly stabilized vision. *J. Opt. Soc. Am.*, 57(6):819–826.

Fernandez-Caballero, A., Lopez, M. T., and Saiz-Valverde, S. (2008). Dynamic stereoscopic selective visual attention (dssva): Integrating motion and shape with depth in video segmentation. *Expert Systems with Applications*, 34(2):1394 – 1402.

Ferwerda, J. A., Shirley, P., Pattanaik, S. N., and Greenberg, D. P. (1997). A model of visual masking for computer graphics. *International Conference on Computer Graphics and Interactive Techniques*.

Ferwerda, J. G. (2003). *The world of 3-D: A practical guide to stereo photography*. 3-D Book Productions, Borger, The Netherlands.

Filippini, H. R. and Banks, M. S. (2009). Limits of stereopsis explained by local cross-correlation. *Journal of Vision*, 9(1).

Fleet, D. J., Wagner, H., and Heeger, D. J. (1996). Neural encoding of binocular disparity: Energy models, position shifts and phase shifts. *Vision Research*, 36(12):1839 – 1857.

Frisby, J. P. and Mayhew, J. E. W. (1978). Contrast sensitivity function for stereopsis. *Perception*, 7:423–429.

Gao, D., Han, S., and Vasconcelos, N. (2009). Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE PAMI*, 31(6):989–1005.

Goferman, G., Zelnik-Manor, L., and Tal, A. (2010). Context-aware saliency detection. *IEEE PAMI*, 34(10):1915–1926.

Guo, C., Ma, Q., and Zhang, L. (2008). Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *IEEE CVPR 2008*.

Harris, J. M. and Wilcox, L. M. (2009). The role of monocularly visible regions in depth and surface perception. *Vision Research*, 49(22):2666 – 2685.

Heinzle, S., Greisen, P., Gallup, D., Chen, C., Saner, D., Smolic, A., Burg, A., Matusik, W., and Gross, M. (2011). Computational stereo camera system with programmable control loop. *ACM Trans. Graph.*, 30:94:1–94:10.

Held, R., Cooper, E., and Banks, M. (2012). Blur and disparity are complementary cues to depth. *Current Biology*, 22(5):426 – 431.

Hoffman, D. M., Girshick, A. R., Akeley, K., and Banks, M. S. (2008). Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *J. Vis.*, 8(3):1–30.

Holliman, N. (2004). Mapping perceived depth to regions of interest in stereoscopic images. In *Proceedings of the SPIE*, Stereoscopic Displays and Applications XV.

Hou, X. and Zhang, L. (2007). Saliency detection: A spectral residual approach. *IEEE CVPR*, pages 1–8.

Howard, I. P. (2002). *Seeing in Depth – Basic Mechanisms*, volume 1. I Porteous, Thornhill, Ontario, Canada.

Howard, I. P. and Rogers, B. J. (2002). *Seeing in Depth: Volume 2: Depth perception*. Oxford University Press, NY, USA.

IJsselsteijn, W. A., Seuntiãns, P. J. H., and Meesters, L. M. J. (2005). Human factors of 3d. In *3D Videocommunication: Algorithms, Concepts and Real-Time Systems in Human Centred*, chapter 12, pages 219–233. John Wiley & Sons, Ltd.

Itti, L. and Baldi, P. (2005). A principled approach to detecting surprising events in video. *IEEE CVPR*, 1:631–637.

Itti, L. and Dhavale, N. (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. *Proc. SPIE*, pages 64–78.

Itti, L. and Koch, C. (2001). Computational modeling of visual attention. *Nature reviews neuroscience*, 2(3):194–203.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259.

Jeong, S., Ban, S., and Lee, M. (2008). Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment. *Neural Networks*, 21(10):1420–1430.

Jin, E. W., Miller, M. E., Endrikhovski, S., and Cerosaletti, C. D. (2005). Creating a comfortable stereoscopic viewing experience: Effects of viewing distance and field of view on fusional range. In *Stereoscopic Displays and Virtual Reality Systems XII*, volume 5664 of *Proc. of SPIE-IS&T Electronic Imaging*, pages 10–21.

Jones, G., Lee, D., Holliman, N., and Ezra, D. (2001). Controlling perceived depth in stereoscopic images. In *Stereoscopic Displays and Virtual Reality Systems VIII*, volume 4297 of *Proc. SPIE*.

Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). Learning to predict where humans look. *IEEE ICCV*.

Julesz, B. (1960). Binocular depth perception of computer-generated patterns. *Bell System Tech.*, 39(5):1125–1161.

Kalloniatis, M. and Luu, C. (2013). Webvision: Perception of depth. Retrieved from http://webvision.med.utah.edu/ on December 2013.

Kane, D., Guan, P., and Banks, M. S. (2014). The limits of human stereopsis in space and time. *The Journal of Neuroscience*, 34(4):1397–1408.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., and Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *International Journal of Aviation Psychology*, 3:203–220.

Koch, C. and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 4(4):219–27.

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

Kooi, F. L. and Toet, A. (2004). Visual comfort of binocular and 3d displays. In *Displays*, volume 25, pages 99–108.

Koppal, S. J., Zitnick, C. L., Cohen, M. F., Kang, S. B., Ressler, B., and Colburn, A. (2011). A viewer-centric editor for 3d movies. *IEEE Computer Graphics and Applications*, 31(1):20–35.

Krähenbühl, P., Lang, M., Hornung, A., and Gross, M. (2009). A system for retargeting of streaming video. *ACM Trans. Graph.*, 28(5).

Kriv\'anek, J., Ferwerda, J. A., and Bala, K. (2010). Effects of global illumination approximations on material appearance. In *ACM SIGGRAPH 2010 papers*, SIGGRAPH '10, pages 112:1–112:10. ACM.

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

Lambooij, M., IJsselsteijn, W., Fortuin, M., and Heynderickx, I. (2009). Visual discomfort and visual fatigue of stereoscopic displays: A review. *Journal of Imaging Science and Tech.*, 53(3):030201.

Lang, C., Nguyen, T., Katti, H., Yadati, K., Kankanhalli, M., and Yan, S. (2012a). Depth matters: Influence of depth cues on visual saliency. In *Proc. of ECCV*, pages 101–115.

Lang, M., Hornung, A., Wang, O., Poulakos, S., Smolic, A., and Gross, M. (2010). Nonlinear disparity mapping for stereoscopic 3d. *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 29(3).

Lang, M., Wang, O., Aydın, T., Smolic, A., and Gross, M. (2012b). Practical temporal consistency for image-based graphics applications. *ACM Trans. Graph.*, 31(4):34:1–34:8.

Lee, J. S. and Ebrahimi, T. (2012). Perceptual video compression: A survey. *IEEE Selected Topics in Signal Processing*, 6(6):684–697.

Lee, Y. J., Ghosh, J., and Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. *IEEE CVPR*, pages 1346–1353.

Lipton, L. (1982). *Foundations of the Stereoscopic Cinema: A Study in Depth*. Van Nostrand Reinhold Company Inc.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., and Shum, H.-Y. (2011). Learning to detect a salient object. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(2):353–367.

Ma, Y.-F. and Zhang, H.-J. (2003). Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, MULTIMEDIA '03, pages 374–381, New York, NY, USA. ACM.

Mahadevan, V. and Vasconcelos, N. (2010). Spatiotemporal saliency in dynamic scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(1):171–177.

Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W. (2011). HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4):40:1–40:14.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company, NY.

Marr, D. and Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, 194(4262).

Marr, D. and Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 204(1156).

Masaoka, K., Hanazato, A., Emoto, M., Yamanoue, H., Nojiri, Y., and Okano, F. (2006). Spatial distortion prediction system for stereoscopic images. *Journal of Electronic Imaging*, 15(1):1–12.

Mascelli, J. V. (1965). *The Five C's of Cinematography*. Silman-James Press, Beverly Hills, CA.

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

McKee, S. P. and Mitchison, G. J. (1988). The role of retinal correspondence in stereoscopic matching. *Vision Research*, 28(9).

Mendiburu, B. (2009). *3D Moving Making: Stereoscopic Digital Cinema from Script to Screen*. Focal Press.

Mitchison, G. J. and McKee, S. P. (1985). Interpolation in stereoscopic matching. *Nature*, 315.

Müller, B. and Basler, H. (1993). *Kurzfragebogen zur aktuellen Beanspruchung (KAB)*. Beltz-Test GmbH.

Murray, N., Vanrell, M., Otazu, X., and Parraga, C. A. (2011). Saliency estimation using a non-parametric low-level vision model. *IEEE CVPR*, pages 433–440.

Neuman, R. (2009). Bolt 3d: A case study. *Proc SPIE*, 7237.

Nguyen, D., Vedamurthy, I., and Schor, C. M. (2008). Cross-coupling between accommodation and convergence is optimized for a broad range of directions and distances of gaze. *Journal of Vision*, 48:893–903.

Nienborg, H., Bridge, H., Parker, A. J., and Cumming, B. G. (2004). Receptive field size in v1 neurons limits acuity for perceiving disparity modulation. *The Journal of Neuroscience*, 24(9):2065–2076.

Niu, Y., Geng, Y., Li, X., and Liu, F. (2012). Leveraging stereopsis for saliency analysis. *IEEE CVPR*, pages 454–461.

Ogle, K. N. (1932). An analytical treatment of the longitudinal horopter, its measurement and application to related phenomena, especially to the relative size and shape of the ocular images. *Journal of the Optical Society of America*, 22(12).

Ogle, K. N. (1950). *Researches in Binocular Vision*. W.B. Saunders Company, Philadelphia.

Ogle, K. N. and Schwartz, J. T. (1959). Depth of focus of the human eye. *Journal of the Optical Society of America*, 49(3):273.

Ohno, Y. (2000). Cie fundamentals for color measurements. *Proceedings of IS&T NIP 16 Conference*.

Ohzawa, I. (1998). Mechanisms of stereoscopic vision: the disparity energy model. *Current Opinion in Neurobiology*, 8(4):509 – 515.

Ohzawa, I., DeAngelis, G. C., and Freeman, R. D. (1990). Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science*, 249(4972):1037–1041.

Oruç, I., Maloney, L. T., and Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision Research*, 43:2451–2468.

Parker, A. J. (2007). Binocular depth perception and the cerebral cortex. *Nature Reviews Neuroscience*, 8:379–391.

Patterson, R. (2007). Human factors of 3-d displays. *Journal of the Society for Information Display*, 15:861–871.

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

Perazzi, F., Krähenbühl, P., Pritch, Y., and Hornung, A. (2012). Saliency filters: Contrast based filtering for salient region detection. *IEEE CVPR*, pages 733–740.

Perlin, K. and Hoffert, E. M. (1989). Hypertexture. *SIGGRAPH Comput. Graph.*, 23:253–262.

Qi, S. and Ho, J. (2013). Shift-map based stereo image retargeting with disparity adjustment. In *Proceedings of the 11th Asian Conference on Computer Vision - Volume Part IV*, ACCV'12, pages 457–469, Berlin, Heidelberg. Springer-Verlag.

Ramadan, S. (2009). New stereo 3d discovery: Women view men 20% larger than men view themselves. Retrieved December 2013 from http://goarticles.com/article/New-Stereo-3D-Discovery-Women-view-Men-20-larger-than-Men-view-themselves/1648164/.

Ramanarayanan, G., Ferwerda, J., Walter, B., and Bala, K. (2007). Visual equivalence: towards a new standard for image fidelity. SIGGRAPH '07. ACM.

Rapantzikos, K., Tsapatsoulis, N., Avrithis, Y., and Kollias, S. (2009). Spatiotemporal saliency for video classification. *Signal Processing: Image Communication*, 24(7):557–571.

Ren, Z., Hu, Y., Chia, L.-T., and Rajan, D. (2010). Improved saliency detection based on superpixel clustering and saliency propagation. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 1099–1102, New York, NY, USA. ACM.

Richards, W. (1971). Anomalous stereoscopic depth perception. *Journal of the Optical Society of America*, 61(3).

Schrater, P. and Kersten, D. (2000). How optimal depth cue integration depends on the task. *International Journal of Computer Vision*, 40(1):71–89.

Schreiber, K. M., Hillis, J. M., Filippini, H. R., Schor, C. M., and Banks, M. S. (2008). The surface of the empirical horopter. *Journal of Vision*, 8(3).

Seo, H. J. and Milanfar, P. (2009). Nonparametric bottom-up saliency detection by self-resemblance. In *Computer Vision and Pattern Recognition Workshops*, pages 45–52.

Sheskin, D. J. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC.

Shibata, T., Kim, J., Hoffman, D. M., and Banks, M. S. (2011). The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of Vision*, 11(8):1–29.

Siegel, M. W. and Nagata, S. (2000). Just enough reality: Comfortable 3-d viewing via microstereopsis. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(3).

Smit, F. A., van Liere, R., and Froehlich, B. (2007). Three extensions to subtractive crosstalk reduction. In *Eurographics Symp. on Virtual Env.*, pages 85–92.

Spottiswoode, R. and Spottiswoode, N. L. (1953). *The Theory of Stereoscopic Transmission*. University of California Press, Berkeley & Los Angeles.

Stefanoski, N., Wang, O., Lang, M., Greisen, P., Heinzle, S., and Smolic, A. (2013). Automatic view synthesis by image-domain-warping. *IEEE Transactons on Image Processing*, 22(9):3329–3341.

Steinman, S. B., Steinman, B. A., and Garzia, R. P. (2000). *Foundations of binocular vision: a clinical perspective*. McGraw-Hill Professional.

Tsirlin, I., Wilcox, L. M., and Allison, R. S. (2010). Monocular occlusions determine the perceived shape and depth of occluding surfaces. *Journal of Vision*, 10(6).

Tsirlin, I., Wilcox, L. M., and Allison, R. S. (2011a). The effect of crosstalk on depth magnitude in thin structures. In Woods, A. J. and Holliman, N., editors, *Stereoscopic Displays and Applications XXII*, volume 7863 of *Proceedings of the SPIE-IS&T Electronic Imaging*.

Tsirlin, I., Wilcox, L. M., and Allison, R. S. (2011b). The effect of crosstalk on the perceived depth from disparity and monocular occlusions. *Broadcasting, IEEE Transactions on*, 57(2):445–453.

Tyler, C. W. (1975). Spatial organization of binocular disparity sensitivity. *Vision Research*, 15(5):583 – 590.

Ukai, K. and Howarth, P. A. (2008). Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations. *Displays*, 29(2):106 – 116.

Valyus, N. A. (1966). *Stereoscopy*. London: Focal Press.

van Baar, J., Poulakos, S., Jarosz, W., Nowrouzezahrai, D., Tamstorf, R., and Gross, M. (2011). Perceptually-based compensation of light pollution in display systems. In *Proceedings of the 2011 ACM Symposium on Applied Perception in Graphics and Visualization*, New York, NY, USA. ACM.

Vasudevan, B., Ciuffreda, K. J., and Wang, B. (2007). Subjective and objective depth-of-focus. *Journal of Modern Optics*, 54(9):1307–1316.

Vig, E., Dorr, M., and Cox, D. (2012). Space-variant descriptor sampling for action recognition based on saliency and eye movements. *Proc. of ECCV*, pages 84–97.

Wang, B. and Ciuffreda, K. J. (2006). Depth-of-focus of the human eye: Theory and clinical implications. *Survey of Ophthalmology*, 51(1).

Wang, L., Teunissen, K., Tu, Y., Chen, L., Zhang, P., Zhang, T., and Heynderickx, I. (2011a). Crosstalk evaluation in stereoscopic displays. *J. of Display Technology*.

Wang, M., Konrad, J., Ishwar, P., Jing, K., and Rowley, H. (2011b). Image saliency: From intrinsic to extrinsic context. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 417–424.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4).

Ware, C. and Mitchell, P. (2008). Visualizing graphs in three dimensions. *ACM Trans. Appl. Percept.*, 5(1):2:1–2:15.

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

Watson, A. B., Ahumada, A. J., and Farrell, J. E. (1986). Window of visibility: a psychophysical theory of fidelity in time-sampled visual motion displays. *J. Opt. Soc. Am. A*, 3(3).

Watson, A. B. and Solomon, J. A. (1997). Model of visual contrast gain control and pattern masking. *J. Opt. Soc. Am. A*, 14(9):2379–2391.

Watt, S. J., Akeley, K., Ernst, M. O., and Banks, M. S. (2005). Focus cues affect perceived depth. *Journal of Vision*, 5(10):834–862.

Werlberger, M., Pock, T., and Bischof, H. (2010). Motion estimation with non-local total variation regularization. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA.

Woods, A. J., Docherty, T., and Koch, R. (1993). Image distortions in stereoscopic video systems. *Proc SPIE*, 1915.

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

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

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

Yano, S., Shinji, I., Tetsuo, M., and Thwaites, H. (2002). A study of visual fatigue and visual comfort for 3d hdtv/hdtv images. *Displays*, 23.

Yeh, Y. and Silverstein, L. (1990). Limits of fusion and depth judgment in stereoscopic color displays. *Human Factors*, 32(1):45–60.

Zhai, Y. and Shah, M. (2006). Visual attention detection in video sequences using spatiotemporal cues. *ACM Multimedia*, pages 815–824.

Zilly, F., Kluger, J., and Kauff, P. (2011). Production rules for stereo acquisition. *Proceedings of the IEEE*, 99(4):590–606.

Zilly, F., Muller, M., Eisert, P., and Kauff, P. (2010). The stereoscopic analyzer — an image-based assistance tool for stereo shooting and 3d production. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 4029 –4032.