TOWARDS PERCEPTUALLY REALISTIC GAZE-CONTINGENT VIRTUAL AND AUGMENTED REALITY DISPLAYS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Brooke Krajancich

June 2023

Abbreviations

VR Virtual reality

AR Augmented reality

HVS Human visual system

FOV Field of view

MAR Minimum angle of resolution

IPD Inter-pupillary distance

CSF Contrast sensitivity function

CFF Critical flicker frequency

Abstract

Virtual and augmented reality (VR/AR) display systems strive to generate perceptually realistic user experiences, that is, the ability to render digital objects that a human observer cannot distinguish from a real object. Current-generation systems have made significant progress towards this goal, however still struggle to match the spatio-temporal and depth sensing capabilities of human vision, constrained by the limited compute budgets, hardware, and transmission bandwidths of wearable computing systems. Gaze-contingent rendering and display paradigms have emerged as a promising solution. Enabled by recent developments in wearable eye tracking systems, this suite of techniques utilizes real-time eye movement sensing to adjust content to reduce bandwidth requirements or improve visual experience.

In this dissertation, we introduce and evaluate several approaches that aim to create more perceptually realistic VR and AR experiences using gaze-contingent display techniques. We model both the spatio-temporal sensitivity of the human visual system and how perception is affected by the distribution of visual attention. In both cases we demonstrate that exploiting these effects could significantly improve potential bandwidth savings of existing approaches. Next, we demonstrate how these bandwidth savings are directly dependent on the latency of the gaze-contingent display system, showing that reducing the latency of current generation systems could enable up to double the bandwidth savings. Finally, we describe how gaze-contingent stereoscopic rendering can improve the accuracy of disparity and depth rendering, including improving shape distortion in VR and alignment of physical and digitally rendering objects in AR.

Acknowledgements

I would like to thank a number of people for their support and guidance over these past 5 years. First and foremost, my advisor, Prof. Gordon Wetzstein who welcomed me into his research group at Stanford, taught me the power of Heilmeier and helped me find a research area I love. Likewise, I am grateful to my colleagues at the Stanford Computational Imaging Lab. Their expertise, sense of humor, and generosity participating in user studies, not only kept me sane, but helped me find purpose and excitement in my research. In particular, I would like to acknowledge Petr Kellnhofer, who was not only key to the success of all but one of the works presented in this dissertation, but also in keeping things fun and reminding me to not take things too seriously. To my friends at Stanford, particularly Leehi and Sonia, for helping me settle in and enjoy living so far from home. I'm also grateful to my wonderful and supportive family and friends back home, who have supported me in all of my pursuits from behind the scenes since day one. Last, but definitely not least, I wish to thank my partner, Jye, who anchored me through all the ups and downs of this challenging degree, even at times from the other side of the world.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

> *"The screen is a window through which one sees a virtual world. The challenge is to make that world look real, act real, sound real, feel real."*
>
> - Ivan Sutherland, *1965*

Virtual and augmented reality, or VR and AR, are next-generation display systems that promise new ways to interact with digital content, feeding us information or placing us in environments that would otherwise be infeasible or dangerous. While entertainment in the form of video games, video streaming, or artistic experiences may have been the most common use case for VR and AR to date, these systems have also begun to enable innovative solutions in various other industries. For example, in education and business, VR has been used to promote diversity, equity and inclusion with immersive empathy training [1]. While in medicine, VR has enabled modern medical training solutions [2] and AR has helped give surgeons "x-ray vision" – the ability to visualize pre-operative scans superimposed on the patient [3, 4].

For many of these applications it is desirable, if not critical, that the display systems synthesize *perceptually realistic* visual experiences. To quantify this, the research community has adopted the *visual Turing test* as the criterion for success. Similar to the Turing test, devised by Alan Turing in 1950, to evaluate whether a computer can pass for a human, Wetzstein and Lanman [5] popularized the visual Turing test to evaluate whether a display can recreate digital content that a human observer cannot distinguish from the real world. It's a subjective test, and one that no VR or AR technology can pass today. While VR can provide amazingly immersive experiences and AR can render impressive world-locked content, neither is at the level where anyone would wonder whether what they're looking at is real or virtual. One of the greatest challenges is simply resolution. To be able to pass the visual Turing test, we need to be able to form an image on the retina such that the observer cannot tell that the photons are emitted from discrete pixel locations, or from only a small region in space. So we need a display that covers at least the field of view (FOV) of the human visual system (HVS), with pixels smaller than our spatial resolving capacity, or around 13,000 pixels just in the horizontal direction[1] – far more than any existing consumer display. Short of this, we have to make a trade-off between resolution and FOV with the number of pixels we have available with current display technology, resulting in visual artifacts and detracting from perceptual realism. But even if current display technology could meet such specifications, drive and operate that many pixels, GPU performance budgets are far from that required to render such an enormous number of pixels, not to mention compressing and transmitting the 17 billion[2] pixels per second needed to run such a display. To put that into perspective, we'd need around 20 HDMI cables to move the ~ 50 GB/s.

However, many of these pixels are squandered due to fact that our perceptual capabilities vary across our visual field. For one, only a small central region of our retina, about 4° in diameter, called the fovea, is responsible for our spatial resolving capacity [6]. Outside the fovea, this ability falls off very rapidly. In fact, by 20° of eccentricity, it's already dropped by 90%, and if we built a display that covered our entire visual field, over 95% of pixels would be outside this boundary [3].

So with recent developments in wearable eye-tracking devices [9], it's no surprise that perceptually-aware computer graphics approaches have drawn extensive attention from various applications. In particular, gaze-contingent rendering and display has emerged as a promising solution to the bandwidth challenges of VR and AR, attempting to balance the amount of information displayed against the visual information processing capacity of the observer through real-time eye movement sensing; see [10, 11] for a review of this area. Based on the assumed knowledge of the instantaneous location of the observer's gaze position, content can be imperceptibly adjusted through several display processing approaches, including foveated rendering [7, 12, 13], compression [14, 15] and display [16, 17], to reduce rendering bandwidth or transmission requirements. In Chapter 2 we provide a detailed review of these works, along with a brief overview of terminology and techniques used to quantify visual perception relevant for later chapters.

However, while foveated graphics to date has typically focused on exploiting the variation in spatial sensitivity across the retina, the human visual system has other limitations that could be utilized to improve bandwidth savings. In Chapters 3 and 4 we describe two different approaches to doing exactly this, by modelling the spatio-temporal sensitivity and attentional effects of the human visual system, respectively. Each demonstrate the potential for significant compression gains over existing spatial-only approaches. In Chapter 5 we make the claim that latency – the time between a change in the viewer's gaze and the resulting change in the display's pixels – of a gaze-contingent display system is directly connected to the bandwidth savings afforded by foveated graphics. With a custom, low-latency foveated compression system, we demonstrate that up to double the bandwith

[1] Assuming a horizontal FOV of 220° and a spatial resolving capacity of 1 arcmin

[2] Assuming a conservative visual acuity of 40 cpd, a CFF of 90 hz and a FOV of $220° \times 135°$

[3] Based on a visual field size of $220° \times 135°$ and the acuity model of Geisler et al. [7], fit with parameters from Robson et al. [8]

savings could be achieved, just by lowering the system latency.

The emergence of wearable eye tracking in VR and AR display systems also enables improved perceptual realism with the rendering of dynamic gaze effects. Gaze-contingent display can be used to support focus cues, pupil steering, ocular parallax [18] and improve distortion correction [19] (see Chapter 2 for a detailed review). In Chapter 6 we describe how gaze-contingent stereo rendering could improve the accuracy of disparity and depth rendering by accounting for dynamic shifts of the optical center of the human eye. Our findings demonstrate significant improvements of disparity and shape distortion in a VR setting, and consistent alignment of physical and digitally rendered objects across depths in AR.

While VR and AR display systems have made significant progress towards the ultimate goal of passing the visual Turing test, the limited compute budgets, hardware, and transmission bandwidths of wearable computing systems make matching the spatio-temporal and depth sensing capabilities of the human vision challenging. In this dissertation, we introduce and evaluate several approaches that work towards reducing this challenge by utilizing the gaze-contingent rendering and display paradigm.

Included Publications

The research presented in this dissertation is compiled from several earlier publications. The works have been restructured to have a shared background chapter and supplemental material has been incorporated into the main text or appendices. The following is list of the works and where they appear in this document:

- *A Perceptual Model for Eccentricity-dependent Spatio-temporal Flicker Fusion and its Applications to Foveated Graphics*[4] is reproduced and adapted in Chapter 3 and Appendix A, along with large parts of Sections 2.1.5, 2.1.6 and 2.2.2. **Author contributions:** G.W. conceived the idea. B.K. conducted all user studies and P.K. fit the models to the data. B.K. analyzed the data and wrote the manuscript with input from all authors. G.W. supervised the project.

- *Towards Attention-aware Foveated Rendering*[5] is reproduced and adapted in Chapter 4 and Appendix B, along with Sections 2.1.7 and 2.2.2. **Author contributions:** B.K. conceived the idea, designed and conducted all user studies. B.K. fit the model, P.K. wrote and analyzed the foveation study. B.K. wrote the manuscript with input from all authors. G.W. supervised the project.

- *Towards Retina-Quality VR Video Streaming: 15ms Could Save You 80% of Your Bandwidth*[6] is reproduced and adapted in Chapter 5, along with a minor part of Section 2.2.2. **Author contributions:** K.W. conceived the idea. L.H. designed and built the low-latency system. B.K. and L.H. designed and conducted the latency measurement experiment. B.K. designed the user experiment and L.H. conducted it. L.H. wrote the manuscript with input and figures from B.K. G.W., P.L. and K.W. supervised the project.

- *Optimizing Depth Perception in Virtual and Augmented Reality through Gaze-contingent Stereo Rendering*[7] is reproduced and adapted in Chapter 6 and Appendix C, along with minor parts of Sections 2.1.8 and 2.2.3. **Author contributions:** G.W. conceived the idea. B.K. conducted all experiments and analyzed the data with help from P.K. B.K. wrote the manuscript with input from all authors. G.W. supervised the project.

[4] **Brooke Krajancich**, Petr Kellnhofer, and Gordon Wetzstein. “A perceptual model for eccentricity-dependent spatio-temporal flicker fusion and its applications to foveated graphics.” ACM Transactions on Graphics (TOG) 40.4 (2021): 1-11. © 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

[5] **Brooke Krajancich**, Petr Kellnhofer, and Gordon Wetzstein. “Towards Attention-aware Foveated Rendering.” ACM Transactions on Graphics (TOG) 42.4 (2023): 1-9. © 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

[6] Luke Hsiao, **Brooke Krajancich**, Philip Levis, Gordon Wetzstein, and Keith Winstein. “Towards retina-quality VR video streaming: 15ms could save you 80% of your bandwidth.” ACM SIGCOMM Computer Communication Review 52.1 (2022): 10-19. © 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

[7] **Brooke Krajancich**, Petr Kellnhofer, and Gordon Wetzstein. “Optimizing depth perception in virtual and augmented reality through gaze-contingent stereo rendering.” ACM Transactions on Graphics (TOG) 39.6 (2020): 1-10. © 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Chapter 2

Background

Gaze-contingent displays use eye tracking to update displayed content in real-time. This is usually to exploit variations in perceptual capabilities to save bandwidth or compute, or to render dynamic gaze effects that improve depth rendering or visual comfort. In this chapter, we define some relevant terms from vision science, explore how some relevant perceptual capabilities vary across the visual field, and describe how previous works have used this knowledge to create gaze-contingent rendering and display paradigms.

2.1 Perceptual Abilities Relevant for VR and AR

The HVS is limited in its ability to sense variations in light intensity over space and time. Furthermore, these abilities are dependent on many factors, including where the light lands on the retina and how the brain chooses to process the resulting signal. Understanding and quantifying these perceptual abilities is critical to improving the perceptual realism of VR and AR devices, including developing new gaze-contingent display techniques. In this section we provide a brief overview of the HVS, describing terminology used in later chapters to quantify perceptual abilities and relevant models used for display techniques in VR and AR.

2.1.1 The Human Eye and Vision

The HVS is made up of the eye and brain, connected by the optic nerve and optic tract. The eye serves as the camera, while the brain is responsible for processing the information. When a light ray hits the eye, it first passes through the cornea and undergoes refraction. The refracted ray passes through the aqueous humor, the iris, the lens, the vitreous humor, and finally, reaches the retina – the image sensor of the human eye (see Figure 2.1). The light photons reaching the retina are then detected and converted to electrical signals by two types of photoreceptors: rods and cones, which is then transmitted to the central neural system by ganglion cells for further processing. While we have 120 million rods, six million cones, the number of ganglion cells present in the retina is only around 1.2 million [20, 21], yet still help the retina transmit data at roughly the rate of an Ethernet connection (∼10 Mbit/s) [22].

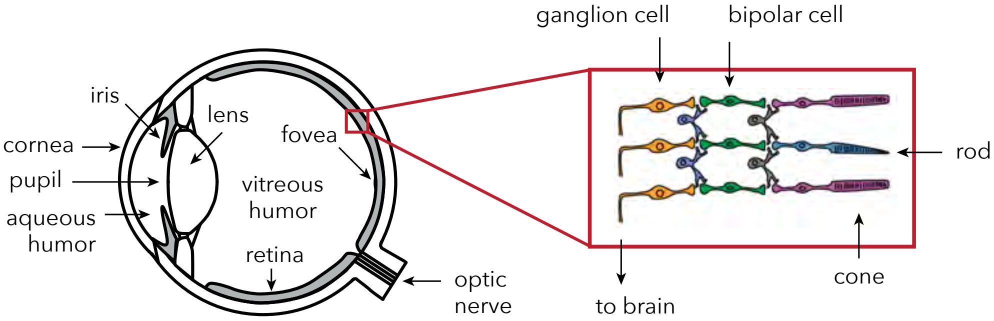

Figure 2.1: Anatomy of the human eye. When a light ray hits the eye, it first passes through the cornea and undergoes refraction. The refracted ray passes through the aqueous humor, the iris, the lens and the vitreous humor, to the retina. The light photons are then detected and converted to electrical signals by rods and cones and transmitted by ganglion cells to the brain for further processing. Retinal cells image from Casey Henley CC BY-NC-SA.

However, these photoreceptors are not evenly distributed across the retina. Cones peak in high density at the fovea, a small area of the retina about 1.5 mm in diameter [23]. Ganglion cell density is also concentrated around the fovea, such that the number of photoreceptors connected to a single ganglion cell increases with distance from the fovea, leading to decreased visual sensitivity (see Section 2.1.4).

2.1.2 Eye Movements

Since optimal vision of small details is only obtained by a small region of the retina, the eyes make short, rapid movements called saccades to scan visual scenes with the high-resolution fovea. While these ballistic-like movements can occur at speeds of up to ∼900°/s [24], perception is largely, although not completely supressed (referred to as saccadic suppression), so that the associated retinal image movement does not cause disturbance. Since this occurs a short period before, during, and after the eye movement (totaling ∼50/,ms [25]), this can reduce some of the challenges they pose to gaze-contingent systems. Though physiologically the eyes can rotate up to around 45° from the straight ahead position during a saccade [26], in practice, this rarely exceeds 20°, as head movement is usually used to fixate on targets outside this.

Smooth pursuit movements are much slower tracking movements of the eyes designed to keep a moving stimulus on the fovea [27]. Even so, the eyes are not fully steady, instead showing a variety of small-amplitude shifts which can be broken into three components: microsacaades, drift, and tremor (see [24, 28] for a detailed review). Fine detail is explored with a slower (velocities up to 1°/s) random-walk-like pattern called drift, where every so often the fixation position is corrected with microsaccades (saccades with amplitude less than 1°). This motion also consists of a small (less than 0.01°/s) continuous high frequency spectrum (up to 200 Hz) called tremor.

2.1.3 The Visual Field

The visual field, also referred to as the FOV of the visual system, describes the range of angles that can be seen by the human eye while fixating straight ahead. The normal[1] *monocular* human visual field, that is the region seen by a single eye, extends to approximately 60° nasally (toward the nose, or inward) from the vertical meridian in each eye, to 110° temporally (away from the nose, or outwards) from the vertical meridian, and approximately 60° above and 75° below the horizontal meridian [29, 30]. Thus the *total* visual field is approximately 220° horizontally by 135° vertically (illustrated in Figure 2.2), given by the union of the monocular visual fields. For comparison, current generation commercial VR and AR headsets can provide a FOV up to approximately 115° horizontally by 90° vertically [2] (outlined in red in Figure 2.2).

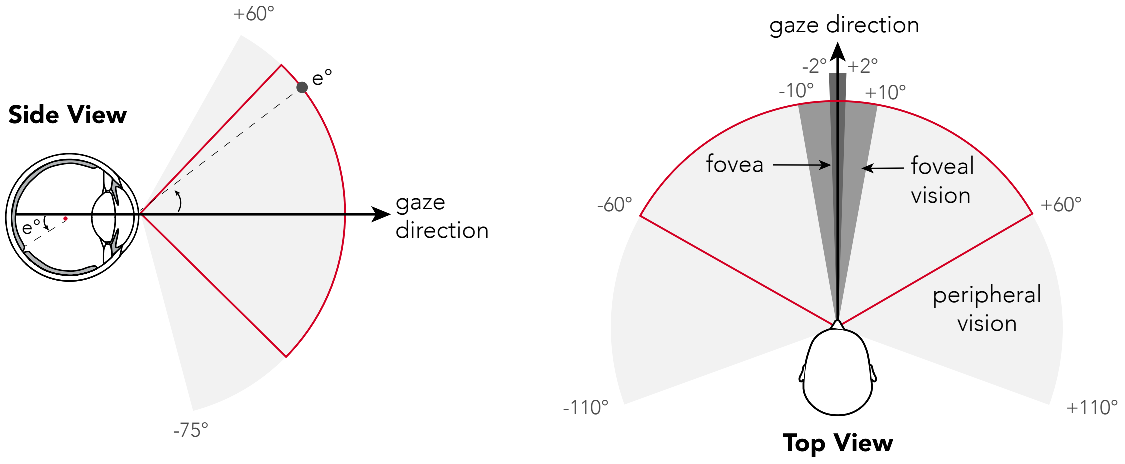

Figure 2.2: Illustration of the human visual field (also referred to as the human FOV). (left) The vertical visual field extends 135°, 60° above and 75° below the horizontal meridian. The eccentricity of a point e is the angular distance of that point from the fovea, or equivalently, from the gaze direction. (right) The total visual field extends 220° horizontally. This is divided into foveal and peripheral regions for the purposes of this dissertation. The FOV of current-generation commercial VR displays is illustrated in red.

While these display systems target covering the total visual field, it is also useful to note that much of this area is seen by only one eye – effectively reducing depth perception in these regions (see Section 2.1.8). The area seen by both eyes, referred to as the *binocular* visual field, is approximately

[1] Significant individual differences exist [6].
[2] Based on the HTC VIVE Pro 2.

120° horizontally by 135° vertically, given by the intersection of the monocular visual fields.

Gaze-contingent display solutions (and this dissertation), usually divide the visual field (or equivalently, the retina) into "foveal" and "peripheral" regions in order to more easily refer to their differing characteristics [12, 13, 31]. Defined by eccentricity, e, or the angular distance of that point from the fovea, these regions are illustrated in Figure 2.2. Even though the fovea actually represents a much smaller area, the VR and AR community usually refer to the central ∼10° region (referred to as the perifovea by vision scientists [23]) as foveal vision, since this is where most of the rendering effort is concentrated. Beyond that is referred to as the periphery.

2.1.4 Visual Acuity and the Minimum Angle of Resolution

The most widely used measure of the visual resolution of the HVS is visual acuity, which measures the size of the smallest detail in a visual target that permits some criterion level of identification or detection performance [32]. The smaller the size of this critical detail, the better the vision of the observer. Visual acuity is highest in the fovea, where photoreceptors are packed most densely. Here, each cone is $\sim 0.5'$ (0.5 arcminutes or 1/120 of a degree of visual angle) in size [33,34], limiting acuity to ∼60 cpd (cycles per degree) by the Nyquist sampling theorem. However, empirical population measurements suggest that the population average is actually closer to 40–50 cpd [8, 12, 35], largely attributed to distortions before the light reaches the retina.

As illustrated in Figure 2.3 (left), 20/20 vision means that a distance of 20 ft the individual can just recognize letters on a Snellen chart whose limbs subtend $1'$ (equivalent to an acuity of 30 cpd). Currently, this is considered normal (or average) visual acuity and is used as the target for display design [36], though currently, commercial VR displays can hardly provide 20/90 visual acuity [3] i.e. $4.5'$ pixel size at 20 ft.

Visual acuity drops off rapidly outside the fovea, in fact, by 20° of eccentricity, it's already dropped by 90% (see Figure 2.3). This falloff has been well described in the literature by many equivalent models [7,37,38] and attributed primarily to the drop in retinal ganglion cell densities [39] but also lens aberrations [35, 40]. In Chapter 3, we use the model of Geisler and Perry [7] shown in Figure 2.3 (right panel, middle plot) to extend our eccentricity-dependent spatio-temporal flicker fusion model data to higher spatial frequencies.

As will be described in Section 2.2.2, many techniques in foveated graphics [12, 13, 42–44] base their sampling distribution on the reciprocal of visual acuity, the minimum angle of resolution (MAR). Then normal (or 20/20) visual acuity is equivalent to a MAR of 1'.

[3] Based on the HTC VIVE Pro 2.

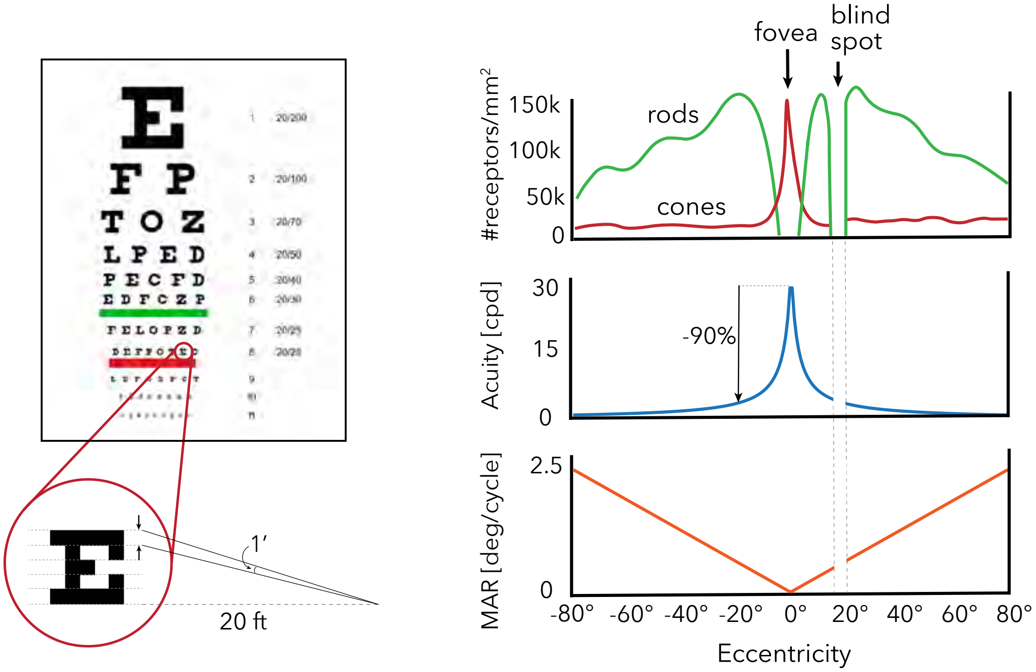

Figure 2.3: Visual acuity, 20/20 vision and the minimum angle of resolution. (left) Snellen charts are commonly used to measure visual acuity. They are designed such that a person with 20/20 vision can just recognize letters whose limbs subtend 1 arcmin (1') at a distance of 20 ft. (right) Top plot: Distribution of photoreceptors (rods and cones) across the retina (data from Pirenne [41]). Between 13.6° to 21.6° at the nasal side is a photoreceptor-free region called blind spot. Middle plot: Visual acuity fall-off model proposed by Geisler and Perry [7], display relative to the fovea. Outside the fovea visual acuity drops off rapidly – by 20° of eccentricity, it's already dropped by 90%. Bottom plot: Model of MAR across eccentricity by Guenter et al. [12].

Defined as the smallest angle at which two points are perceived as different [45], the MAR has been shown to increases approximately linearly with eccentricity, modelled by:

$$\omega(e) = me + \omega_0 \tag{2.1}$$

where w is the MAR (in degrees per cycle), e is the eccentricity angle (in degrees), $\omega_0 = 1/48°$ (in cycles per degree) is the smallest resolvable angle at the fovea, and $m \approx 0.022 - 0.034$ is the experimentally fit MAR slope (dependent on parameters of the display).

However, the objection to visual acuity (and equivalently, MAR) as an assessment of spatial vision is that it only measures ability to perceive sharp and clear outlines of very small (high frequency) objects. It fails to assess visual capability over the full range of spatial frequencies and luminance changes present in typical real-world scenes. Thus it has become common to measure the contrast sensitivity function.

2.1.5 Contrast Sensitivity

Contrast sensitivity measures the ability to detect identify minute differences in shadings and patterns [46]. Sinusoids provide a systematic way to describe our sensitivity to these variations, since we can constitute any waveform with sums of sinusoids with different amplitudes and frequencies. Schade [47] was the first to propose the use of sinusoidal gratings to measure the performance of the visual system with the eye's response in 1956. Soon after, Campbell and Robson [48] proposed the first model of contrast sensitivity, known as a contrast sensitivity function (CSF), as a complex and discrete function of the retina. One attractive feature of CSFs is that they are end-to-end models, which explain the response of the visual system (detection) for a given input (luminance pattern). Because of that, CSFs have found many applications in image/video visibility and quality metrics [49–51], compression codecs [52, 53], tone-mapping operators [54], foveated rendering [55] and many other areas of computer graphics.

The majority of recent CSF data comes from experiments with 2D Gabor wavelets (see Table 1 in [56]); a sinusoid multiplied by a Gaussian envelope to restrict spatial extent.

$$g(\mathbf{x}, \mathbf{x_0}, \theta, \sigma, f_s) = A \exp\left(-\frac{\mathbf{x} + \mathbf{x_0}^2}{2\sigma^2}\right) \cos\left(2\pi f_s \mathbf{x} \cdot [\cos\theta, \sin\theta]\right) + L_0, \tag{2.2}$$

where $\mathbf{x}$ denotes the spatial location on the display, $\mathbf{x_0}$ is the center of the wavelet, σ is the standard deviation of the Gaussian in visual degrees, and f_s and θ are the spatial frequency in cpd and angular orientation in degrees for the sinusoidal grating function. Figure 2.4 shows the 1D luminance profile of this function, together with a full 2D display at decreasing contrasts.

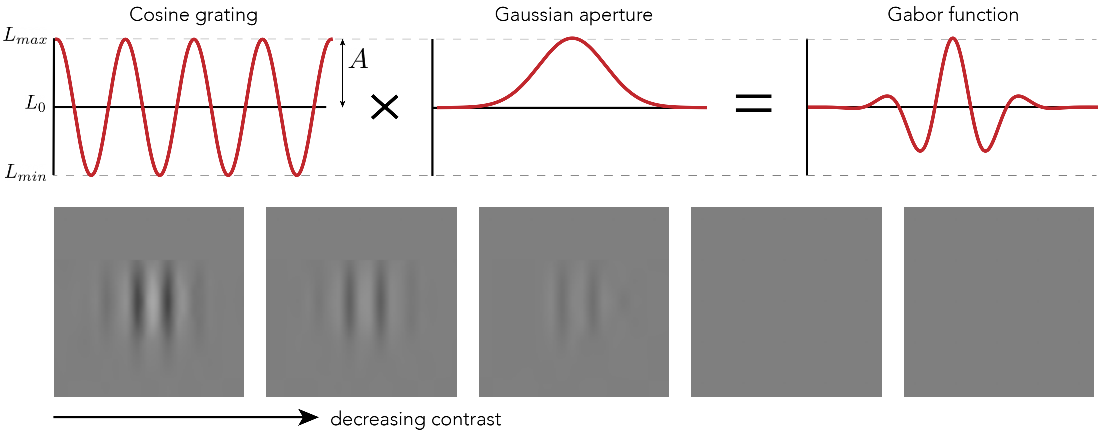

Figure 2.4: Gabor wavelets are most commonly used to measure contrast sensitivity. (top) A Gabor wavelet is constructed by multiplying a sinusoidal function by a Gaussian function, illustrated here in 1D. To find the contrast threshold (1/contrast sensitivity) for a given set of parameters, the contrast of a 2D Gabor wavelet is reduced (bottom) until the orientation (in this case 0°) can no longer be discriminated.

Being a type of wavelet, these functions conveniently minimize the uncertainty principle concerning time-frequency localization i.e. they measure response in a way that accounts for the fact that low frequencies cannot be well localized in space. For this reason, we also choose to use these wavelets to measure spatio-temporal flicker fusion thresholds in Chapter 3.

Contrast, c, can be defined as the degree of blackness to the whiteness of a particular object or a target. For detection gratings, it is typically reported in units of Michelson contrast:

$$c = \frac{L_{max} - L_{min}}{L_{max} + L_{min}} = \frac{m}{L_0}$$

where L_{max} and L_{min} are the luminances of the peaks and troughs of the spatial grating. For sinusoidal gratings, Michelson contrast is equivalent to the modulation amplitude m divided by the background luminance L_0. The contrast threshold, c_T, is then the minimum contrast required to perceive an object clearly. This threshold is typically measured using an n-alternative-force-choice (2AFC, 3AFC,...) protocol, in which n stimuli are shown sequentially or side-by-side and the observer needs to select the one that contained the stimulus or showed a different orientation from the rest [57]. The contrast sensitivity is then the reciprocal of the contrast threshold:

$$S = \frac{1}{c_T} = \frac{L_0}{m}$$

since sensitivity is high when the threshold is low, and visa versa.

Repeating and plotting this for different spatial frequencies is the CSF. The frequency at which the contrast is at its maximum value of 1, or the log of the CSF is zero, is assumed to be the boundary of visibility, and is related (but not identical to) the visual acuity as determined with high-contrast targets.

You can observe the overall shape of your own CSF by looking at Figure 2.5. It shows sinusoidal stripes of increasing spatial frequency along the horizontal axis and of decreasing contrast along the vertical axis. At threshold contrast, your ability to detect the grating disappears, and so the height at which you no longer see the stripes but just a gray background indicates your sensitivity to the gratings at the given spatial frequency (approximately indicated by the red dotted line). If the detection of contrast was dictated solely by image contrast, the alternating bright and dark bars should appear to have equal height everywhere in the image. However, the bars seem to be taller in the middle of the image, showing that people are most sensitive to intermediate spatial frequencies at about 4-5 cpd [6].

The major challenge of measuring and modeling CSFs is the large number of parameters describing the stimulus. Several papers have independently investigated many of these, including background luminance [58–60], size of the stimulus [61], eccentricity [62], orientation [63] and chromaticity [64] of the pattern.

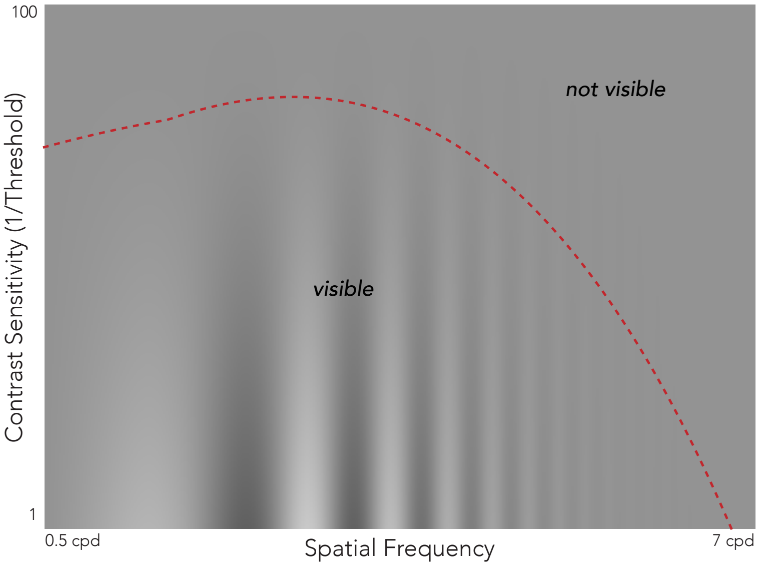

Figure 2.5: Demonstration of the shape of the contrast sensitivity function for luminance gratings. Spatial frequency increases continuously from left to right, and contrast increases from top to bottom. The envelope of the striped region (approximated by the red dotted line) should approximate the sCSF.

Furthermore, most visual scenes show variations in both spatial and temporal contrast. Just as it is possible to obtain the spatial CSF (sometimes referred to as the sCSF), it is also possible to determine the temporal CSF (tCSF), by instead varying the temporal frequency of the sinusoid. The frequency at which the tCSF is at its maximum value of 1, or the log of the CSF is zero, is assumed to be the boundary of visibility, and is referred to as the critical flicker frequency (see Section 2.1.6). Similar to the sCSF, several studies have independently investigated the effect of many parameters on the tCSF [65, 66].

However, spatial and temporal sensitivity are not independent. This relationship can be captured by varying both spatial and temporal frequency of the gratings to obtain the spatio-temporal contrast sensitivity function (stCSF) [65] (shown in Figure 2.6). Kelly [67] was the first to conduct a systematic study of spatio-temporal patterns and fit an analytical function to stCSF data, which was later adjusted by Daly [68] to better describe naturally observed stimuli shown on a cathode-ray tube display. Both models, however, were limited to the fovea and a single luminance. Watson and Ahumada [69] later devised the pyramid of visibility, a simplified model that can be used if only higher frequencies are relevant. While also modeling luminance dependence, this model was again not applicable to higher eccentricities. Subsequent work saw the refitting of this model for higher eccentricity data but the original equation was simplified to only consider stationary content and the prediction of sCSF [70]. However, at the time the work included in this dissertation was published, there existed no unified model parameterized by all axes of interest; spatial frequency, temporal frequency, eccentricity and luminance [71]. In Chapter 3, we address this gap. In particular, we measure critical flicker fusion (see Section 2.1.6) rather than stCSF to be more conservative in modeling the upper limit of human perception, for use in foveated graphics applications.

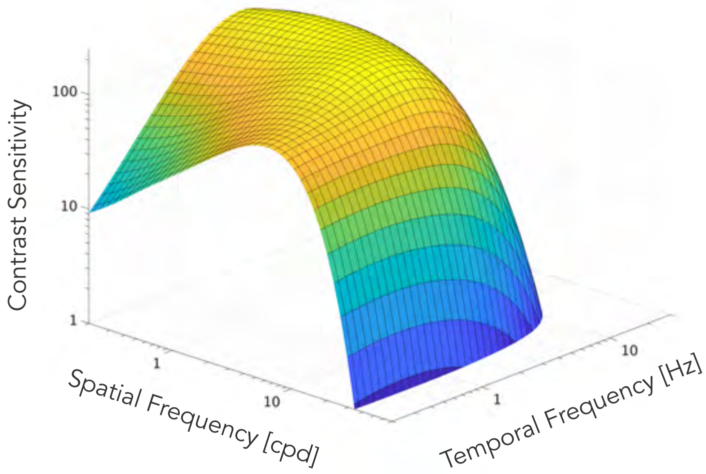

Figure 2.6: Combined spatio-temporal CSF, representing luminance contrast sensitivity for different combinations of spatial and temporal modulation (data from Kelly [67]). Human contrast sensitivity is not space-time separable.

Following growing interest in the topic, additional models capturing eccentricity dependence for the full spatio-temporal domain have since been published [56,72]. Mantiuk et al. [73] proposed a new video quality metric, FovVideoVDP, which used a CSF derived from the spatio-chromatic CSF from Mantiuk et al. [64], the cortical magnification model (see Appendix B.1), and Kelly-Daly's model [68] discussed above. Later Mantiuk et al. [56] proposed a unified model, StelaCSF, which accounts for all major dimensions of the stimulus: spatial and temporal frequency, eccentricity, luminance, and area by combining data from several previous papers. For completeness, we provide a comparison our eccentricity-dependent spatio-temporal flicker fusion model (described in Chapter 3) with StelaCSF in Section 3.5.

2.1.6 Critical Flicker Frequency

Analogous to visual acuity and spatial resolution, and complimentary to the tCSF function, the critical flicker frequency (CFF), or flicker fusion threshold, is a measure of the temporal resolution of the visual system. These measurements capture the temporal frequency at which tCSF drops to unity, or the boundary of visibility, but also represent the frequency at which an intermittent light stimulus appears to be completely steady to the observer.

While related, CFF and tCSF are used by vision scientists to quantify slightly different aspects of human vision. The CFF is considered to be a measure of conscious visual detection dependent on the temporal resolution of visual neurons, since at the CFF threshold, an identical flickering stimulus varies in percept from flickering to stable [74]. While contrast sensitivity is considered to be more of a measure of visual discrimination, as evidenced by the sinusoidal grating used in its measure requiring orientation recognition [75].

Unlike the tCSF, temporal sensitivity has been observed to peak in the periphery, somewhere between $20 - 50°$ eccentricity [76–78] at 70–120 Hz (depending on luminance), attributed to faster cone cell responses in the mid visual field by Sinha et al. [79].

Several studies measure data points for these functions independently [80–82], across eccentricity [76, 83], as a function of luminance [84] and color [85], however few present quantitative models, possibly due to sparse sampling of the measured data. Additionally, data is often captured at luminances very different from typical VR displays. The Ferry–Porter [86] and Granit–Harper [87] laws are exceptions to this, describing CFF as increasing linearly with log retinal illuminance and log stimulus area, respectively. Subsequent work by Tyler et al. [88] showed that the Ferry-Porter law also extends to higher eccentricities. We use this relationship in Chapter 3 to extend our eccentricity-dependent spatio-temporal model of flicker fusion to additional display luminances.

2.1.7 Visual Attention

Even while our eyes are fixated on a particular location, it does not appear that the visual system passively processes all the information available within the image. Rather, we selectively attend to different aspects of it at different times [6]. Sometimes we attend globally to the whole scene; at other times we attend to a selected object or set of objects; at still other times we attend locally to a specific object part. Our ability to engage in these flexible strategies for preferentially processing different information within the visual field is referred to as *visual attention* (or just *attention*).

There are several different types of attention [89] that can be used to explain many interesting phenomena, including change blindness [90] or tunnel vision [91]. However, for the purpose of this dissertation, we will focus discussion on *spatial* attention, which guides preferential processing to a particular location in the visual field. Often modeled as a "zoom" or "variable-power lens", that is, the attended region can be adjusted in size, spatial attention defines a trade-off between its allocation and processing efficiency because of the limited processing capacity of the brain [92]. As illustrated in on the left in Figure 2.7, this region is most often directed *overtly*, by moving our eyes towards the location, but we can also deploy attention to an area in the periphery *covertly*, via a mental shift. Recently it has been shown that combinations of these approaches can be used to split the "zoom lens" across two to four regions of the visual field when it is of use to the observer [93].

Physiologically, attention modulates neuronal responses and alters the profile and position of receptive fields near the attended location [100]. Behaviorally, it improves performance in various

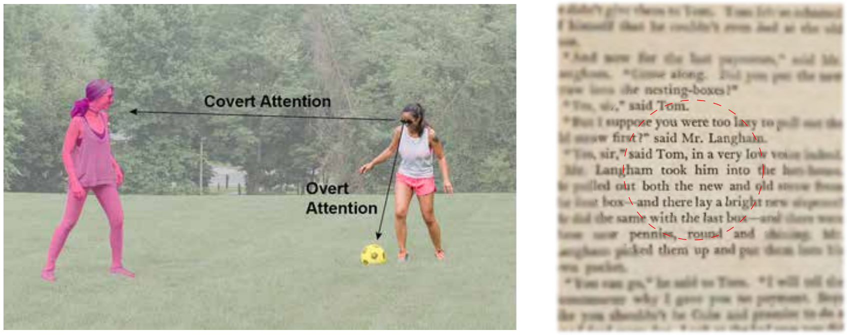

Figure 2.7: Spatial visual attention. (left) Illustration of attention allocation. While the soccer player is overtly attending to (focusing her eyes on) the soccer ball (yellow), she deploys some attention to her periphery to monitor her teammate (pink). Image credit: CC0 from National Eye Institute, NIH. right) Illustration of the "zoom lens" model of visual attention. Several studies have demonstrated that, under many conditions, increasing the amount of attention allocated to a visual task can enhance performance [94, 95], including contrast sensitivity [96, 97], visual acuity [98], and speed of information accrual [99], at a cost to the unattended regions.

visual tasks. One prominent effect of attention is the modulation of performance in tasks that involve the visual system's spatial resolving capacity [101]. In line with the "zoom lens" model, several studies have shown that covert attention enhances contrast sensitivity at the attended location at the cost of decreased sensitivity at unattended locations across the visual field, at different eccentricities and isoeccentric (polar angle) locations [89, 96, 102, 103]. In such studies, attention is typically modulated by visual means, e.g., pre-cueing the location of the visual target [102, 103], or by drawing attention away from the stimuli by use of a concurrent visual task presented elsewhere [97, 104]. Carrasco et al. [102] found that pre-cuing attention to the visual target enhanced contrast sensitivity between 0.05 and 0.1 log units over a broad range of spatial frequencies, and later, Ling and Carrasco [105] described this attention effect as equivalent to applying an effective *contrast gain* to the stimulus. A similar effect also occurs for visual acuity [106] and speed of information accrual [99].

Also in line with the "zoom lens" model, and most similar to our work in Chapter 4, Huang and Dobkins [97] showed that when attention is divided across several points in the visual field, this reduces its enhancement effect at each location. In particular, they showed that drawing attention to the fovea with a rapid serial visual presentation task reduced contrast discrimination performance in the periphery by up to a factor of 10. However, each of these studies measures a single position somewhere between $5-10°$ eccentricity and often only for 1 or 2 users. To the best of our knowledge, this effect has not been modeled over the visual field, nor is there any available data for the effect in the periphery ($> 10°$), which we rectify with our work in Chapter 4.

2.1.8 Stereoscopic Acuity

The fact that humans have two laterally separated eyes whose visual fields overlap in the central region of vision creates one of the most compelling experiences of depth. Since each eye views a different perspective of the environment, the two retinal images are slightly different. That is, the same point in the environment project to locations on the left and right retinae that are displaced in a way that depends on how much closer or farther the point is from the fixation point (illustrated on the left in Figure 2.8 by θ_L and θ_R). This relative angular displacement, i.e. $\theta_L - \theta_R$, is referred to as the binocular disparity. If the binocular disparity of an object is small and within Panum's fusion area then we perceive it as a single, fused object [107]. This process, known as stereopsis, delivers a sense of depth and solid shape, and is considered one of the strongest depth cues for objects within a few meters [108, 109].

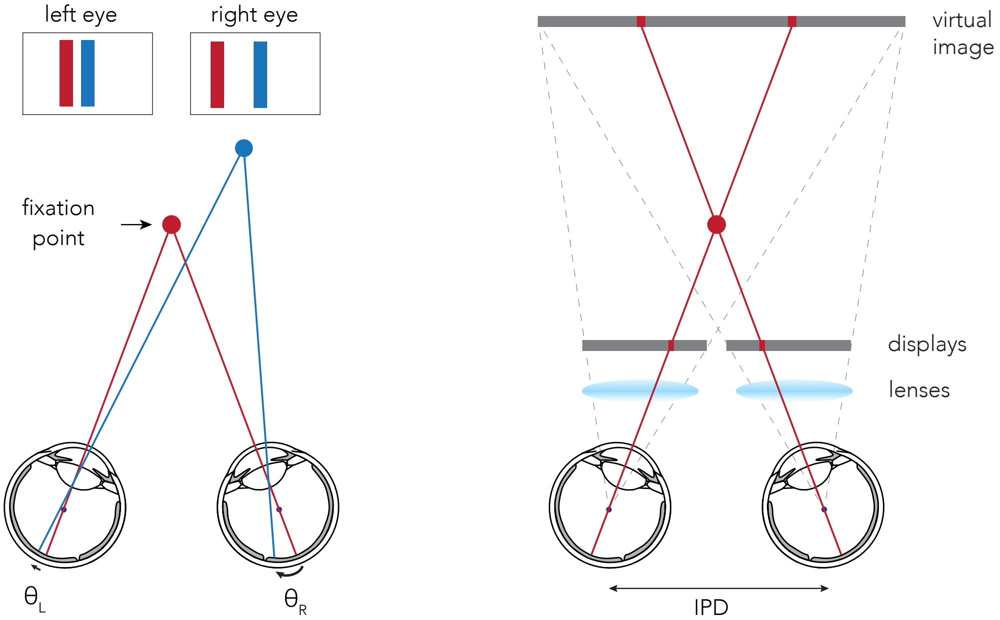

Figure 2.8: Binocular disparity in the real world and rendered by VR. (left) Since humans have laterally separated eyes, the same point in the environment project to different locations on the left and right retinae. The relative angular displacement, $\theta_L - \theta_R$, is referred to as the binocular disparity. (right) In VR and AR this emulated by near eye displays that show slightly different images to the left and right eyes, calculated using the geometry of the system, including the user's inter-pupillary distance (IPD).

Stereoscopic acuity measures the ability to resolve binocular disparity, i.e. that two objects at each eye differ slightly. The value of the difference in angular separation, when the separation in depth is at threshold is called the stereoscopic acuity, and is generally very high for humans; $\sim 20''$ (20 arcseconds or 1/3 of an arcminute) [110, 111]; which is smaller than the angle subtended by an individual photoreceptor in the fovea (see Section 2.1.4). However, this does decrease with retinal eccentricity [112].

As illustrated in Figure 2.8 (right), in order to emulate this depth cue in VR and AR, most conventional systems use some form of near-eye displays to show slightly different images to the left and right eyes. Since the displays are mounted close to the eyes to achieve a wearable form-factor, lenses are often used to create virtual images at some distance, often $1 - 2\,\mathrm{m}$ away from the viewer, where the images are more easily focused. Then to generate the illusion of an object at a particular depth, the geometry of the system (in the form of matrices, see Section 6.2.2), including the user's inter-pupillary distance (IPD) is used to calculate which pixels to light up to generate the retinal disparity that would occur during natural viewing. However, as we will discuss in Chapter 6, since humans have such high stereoscopic acuity, even small inaccuracies in this geometric calculation leads to objects being perceived at a different depth than intended.

2.2 Gaze-contingent Rendering and Display

The gaze-contingent rendering and display paradigm has enabled a variety of important computer graphics techniques that adapt to the user's gaze direction [10, 11]. In this section, we describe how recent developments in eye tracking has enabled such techniques to be used in VR and AR display systems. Then, we discuss several relevant gaze-contingent display works based on whether the gaze data is used to reduce computation or transmission bandwidth (Foveated Graphics), or support dynamic gaze effects.

2.2.1 Recent Developments in Eye Tracking

Eye tracking refers to the combination of hardware and software used to detect a user's eyes over time, usually to calculate where or what they are looking at [36]. The point in 3D space where the user is deemed to be looking is referred to in the literature (and this dissertation) as the gaze position. With the ability to track a user's attention (overt attention, see Sec 2.1.7) in real-time, eye tracking is useful for a wide range of applications, including vision research, saliency, marketing, and as an input mechanism for computing [113]. In VR and AR, it not only enables gaze-contingent display and rendering paradigms, but can also be used to enable user interactions [114], create more realistic virtual avatars [115] or enable biometric authentication [116].

Most modern eye trackers rely on an infrared light source and video cameras to track the pupil centers and the corneal reflection – a projection of the infrared rays from the outer surface of the cornea (see illustration in Figure 2.1). During eye movement, the pupil follows the gaze direction, while the corneal reflection remains unchanged. A personal calibration is usually required to find a mapping function that converts the coordinates reported by the eye tracker to the coordinates of the gaze position in the visual environment. When the gaze position is constrained to a 2D plane, as in Chapters 4 and 5, where we require gaze position on a computer monitor, a standard point-based calibration procedure is usually sufficient. During this procedure, the user is asked to sequentially fixate on a set of (5-16) small point targets displayed on the screen for a few seconds, the data from which is used to fit a polynomial mapping function [114]. However when the gaze position could be anywhere in 3D space, more calibration points are typically required to estimate line-of-sight intersection, and even then, accuracy is usually lower and a continuing challenge for the community [117, 118]. Table 2.1 lists typical accuracies[4] of some commercially-available systems, both VR devices with native eye tracking (above the divider) and the standalone devices used in this dissertation (below the divider). While calibration-free methods exist for multiple-camera settings [120], the highest accuracy for a single camera is usually provided with user-specific calibration [121]. See Liu et al. [118] for a detailed review of this area.

Table 2.1: Specifications of some commercially-available eye trackers, both VR devices with native eye tracking (above the divider) and the standalone devices used in this dissertation (below the divider). *Accuracy* is the average angular deviation between the true gaze position and the gaze position estimated by an eye tracker. *Sampling Rate* or sampling frequency of an eye tracker refers to how many times per second the eye is recorded by the eye tracker. *Latency* refers to the delay between eye movement events and the time these events are received by the computer system (results from Stein et al. [119]). The Eyelink 1000, used in Chapter 5, is often regarded as the industry "gold-standard", however it is a bulky, desktop-based system not usable for VR and AR. The PupilLabs Core is used in Chapter 4.

Company	Accuracy	Sampling Rate	Latency
Fove-0	0.25-1.15°	70 Hz	16 ms
Varjo VR-1	1°	98 Hz	36 ms
HTC VIVE Pro Eye	1.1°	90 Hz	50 ms
PupilLabs Core	0.6°	200 Hz	12 ms
Eyelink 1000	0.25°	1000 Hz	2.2 ms

Eye tracking technology has been commercially-available for decades on desktop displays, but it's been the recent emergence of small, high-resolution cameras that have seen these systems mounted near-eye, allowing use for VR and AR applications. We are already seeing more and more commercial devices include native support, including the Microsoft Hololens 2, Magic Leap One, Varjo VR-1, Fove-0, and HTC Vive Pro Eye, paving the way for widespread deployment of gaze-contingent rendering and display paradigms.

2.2.2 Foveated Graphics

Foveated graphics encompasses a suite of gaze-contingent techniques that exploit eccentricity-dependent aspects of human vision, such as acuity, to minimize the bandwidth of a graphics system by optimizing bit depth [122], color-fidelity [123], level-of-detail [38,124,125], tone mapping [126–128] or by simply reducing the number of vertices or fragments a graphics processing unit has to sample, ray-trace, shade, or transmit to the display; see [10, 11] for a review of this area.

[4] Specifications taken from published specifications and Stein et al. [119]. These are typically recorded under the most ideal conditions and thus accuracies are usually lower in practice.

Foveated rendering is perhaps the most well-known example of this class of algorithms [7, 12, 13, 129–132]; where rendering quality is progressively decreased toward the periphery to improve performance. See Wang et al. [133] for a recent review of this area. In Chapter 4, we approximate the pioneering approach of Guenter et al. [12] to validate our attention-aware CSF model. Assuming that a few discrete layers are sufficient to model the acuity fall-off in the HVS, Guenter et al. [12] used three blended layers rendered at different resolutions, centered at the gaze position or fovea, as shown on the left in Figure 2.9. The innermost, foveal layer, is rendered at the highest (or native display) resolution, while the middle and outermost layers are rendered at progressively lower resolutions. All three layers are then interpolated to the display resolution and blended smoothly. The MAR function (see Section 2.1.4 for more detail) is used to compute the size and resolution for the eccentricity layers as shown in on the right in Figure 2.9, where the slope m is estimated from user studies for the particular display parameters.

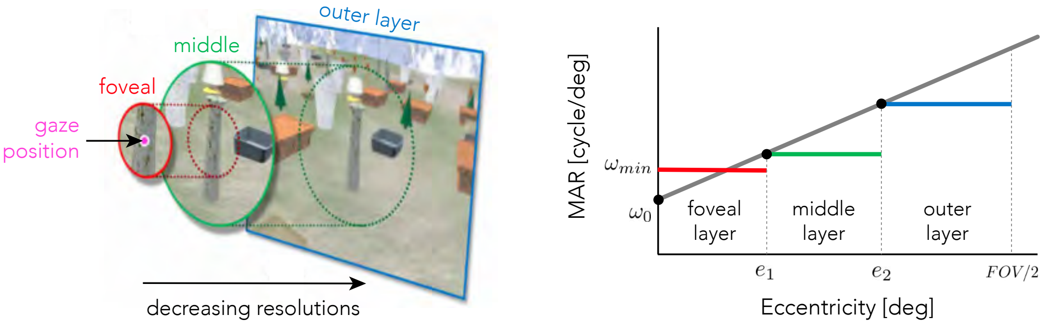

Figure 2.9: Foveated rendering greatly reduces the number of pixels shaded and overall graphics computation. (left) Guenter et al. [12] rendered three eccentricity layers (red border = foveal layer, green = middle layer, blue = outer layer) around the tracked gaze point (pink dot), shown at their correct relative sizes. These are interpolated to native display resolution and smoothly composited to yield the final displayed image. (right) The eccentricity later parameters are selected based on the MAR function. The foveal layer is always sampled at the smallest MAR the display supports, ω^{min}. The other angular radii, e_1 and e_2 are chosen by minimizing the sum of the pixels in all three layers. Images adapted from Guenter et al. [12].

Follow on works have explored adjustments to approximating the model of visual acuity fall-off [31, 134–136], different layer blending approaches [13, 137] and peripheral degradation techniques [136, 138, 139]. While foveated rendering has traditionally focused on exploiting the varying spatial sensitivity of the human eye, several works have demonstrated chromatic degradation can also be used [123, 140], exploiting the rapidly decreasing color sensitivity towards the periphery. Recently, Tursun et al. [55] expanded this concept by considering local luminance contrast across the retina, Xiao et al. [141] additionally consider temporal coherence, and Meng et al. [142] showed that leveraging ocular dominance to further improve performance. To the best of our knowledge, none of these methods exploit the eccentricity-dependent spatio-temporal characteristics of human vision, perhaps because no existing model of human perception accounts for these in a principled manner. In Chapter 3 we develop such a model which could enable significant bandwidth savings for all of the aforementioned techniques well beyond those enabled by existing eccentricity-dependent acuity models.

Foveated display is another approach to reduce bandwidth and optimize visual quality by optically manipulating light from one or more displays [16]. The foveation effect can be achieved by either using a single display [143], or two separate displays [17, 144–146]. Yoo et al. [143] proposed the first single-display-based near-eye foveated system based on temporal polarization multiplexing and provides two operating modes, whereas Tan et al. [145], Kim et al. [17], and Lee at al. [144] use two displays – one for the foveal region with a narrow FOV and the other for the peripheral region with a wide FOV.

Finally, these ideas have also been applied to video processing and compression. Traditional video compression removes temporal and spatial redundancy in a sequence of video frames. Foveated video compression builds on these techniques by using real-time gaze information to concentrate data allocation in an encoded video to the foveal region, achieving better compression in the periphery [14, 15, 131]. This includes techniques such as adapting encoding parameters [147], predicting a user's FOV [148] and upscaling highly compressed video using superresolution [149]. Instead of compressing a single video stream, recently Romero et al. [150] demonstrated a system that stores a video in two resolutions, low and high. A client first fetches the low-resolution stream, and then streams only the cropped, high-resolution segments based on a viewer's current gaze. Similarly, Jeppsson et al. [151] divided a video into many small blocks and pre-encodes each block in many different resolutions. Then, when streaming, the resolutions are chosen on the server based on gaze data and stitched together at the client into three levels of resolution. Foveated video compression can achieve bitrates that are $25 - 60\%$ of the bitrates of traditional compression algorithms with similar visual quality. In Chapter 5 we demonstrate an approach that achieves 80% compression by lowering the system latency to sub 15 ms.

2.2.3 Supporting Dynamic Gaze Effects

The emergence of wearable eye tracking in VR and AR display systems also enables the rendering of dynamic gaze effects, including focus cues, pupil steering, ocular parallax and improved distortion correction. In Chapter 6 we discuss how they also enable us improve the perception of depth via gaze-contingent disparity rendering.

Focus Cue Supporting Displays

Conventional displays cannot reproduce a critical aspect of 3D vision in the natural environment: changes in stereoscopic depth are also associated with changes in focus. For users with normal vision, this creates an unnatural condition known as the vergence–accommodation conflict [152], leading to double vision, compromised visual clarity, visual discomfort, and fatigue [153,154]. Moreover, a lack of accurate focus also removes cues that are important for depth perception [108].

Several different solutions have been proposed, many utilizing gaze-contingent display paradigms. Disparity manipulation is one such approach, where the disparities of a stereo image are remapped to fit into the zone of comfort of a 3D display to mitigate the vergence–accomodation conflict and improve user comfort [154–157]. This amounts to shifting the 3D scene forward or back according to gaze depth such that the fixated object appears well within the zone of comfort, close to the physical screen. Depth of field rendering is another computational approach, where gaze depth is used to calculate and render the correct focus blur in the displayed scene [158–160]. While improving perceptual realism, these approaches do not stimulate accommodation and thus, have not been found to improve perception and comfort [158]. To address the issue of simulating correct accommodative distances in a gaze-contingent manner, the focus distance needs to be shifted. Another type of gaze-contingent display, referred to as varifocal displays, can address this by dynamically shifting the image plane according to the gaze depth [161–164].

Pupil Steered Displays

Some types of near-eye displays, most notably light-field [165–167] and holographic [168–170] displays, can in theory provide natural parallax within their respective eye box volume. Similarly, multifocal displays can be used to render focus blur by approximating the scene volume by decomposition across a few virtual planes [171–173]. While these displays don't rely on tracking gaze position, the eye boxes of current demonstrations are too small to support a significant rotational range of the eye [173]. This not only prevents realistic parallax, but can also easily destroy the visual percept itself. One approach to exit pupil expansion is pupil steering [174–176], a gaze-contingent display paradigm where a user's eyes are tracked and the small exit pupils of the displays are optically steered towards the user's pupils.

Ocular Parallax Rendering

The gaze-contingent stereoscopic rendering approach we present in Chapter 6, builds off work from Konrad et al. [18] on ocular parallax rendering. They show that the offset between the centers of rotation and projection of the human eye creates an important monocular depth cue, ocular parallax. While they demonstrated that using gaze positions to render this effect in VR and AR improves ordinal depth perception and portrayal of realistic depth in monocular viewing conditions, we are the first to show significant effects on absolute depth perception and also digital–physical object alignment in VR and AR applications.

Optical Distortion Correction

The lenses used in conventional VR headsets typically cause a number of visual artifacts that are not seen with direct-view displays, such as significant optical distortion. In practice, this is mitigated by counter-warping the displayed image. However, if the rendered distortion correction does not update with eye rotation, the apparent optical distortion will continually change – a phenomenon known as pupil swim [177]. Several approaches to mitigate this effect have been proposed using gaze-contingent rendering, both in simulation [178, 179] and demonstration [19].

Chapter 3

Towards Spatio-temporal Foveated Rendering

With the goal of showing digital content that is indistinguishable from the real world, VR and AR displays strive to match the perceptual limits of human vision. That is, a resolution, framerate, and FOV that matches what the human eye can perceive. However, the required bandwidths for graphics processing units to render such high-resolution and high-framerate content in interactive applications, for networked systems to stream, and for displays and their interfaces to transmit and present this data is far from achievable with current hardware or standards.

Foveated graphics techniques have emerged as one of the most promising solutions to overcome these challenges (see Section 2.2.2). While these methods have typically built on the insight that visual acuity, or spatial resolution, varies across the retina, common wisdom also suggests that so does temporal resolution. This suggests that further bandwith savings, well beyond those offered by today's foveated graphics approaches could be enabled by exploiting such perceptual limitations with novel gaze-contingent hardware or software solutions. However, as described in Section 2.1.5 and illustrated in Table 3.1, at the time this work was published, no perceptual model for eccentricity-dependent spatio-temporal aspects of human vision existed (see Section 3.5 for further discussion on new work).

In this chapter, we experimentally measure user data and computationally fit models that adequately describe the eccentricity-dependent spatio-temporal aspects of human vision. Specifically, we design and conduct user studies with a custom, high-speed VR display that allows us measure data to model CFF in a spatially-modulated manner. In this chapter, we operationally define CFF as a measure of spatio-temporal flicker fusion thresholds. As such, and unlike current models of foveated vision, our model is unique in predicting what temporal information may be imperceptible for a certain eccentricity, spatial frequency and luminance (illustrated in Figure 3.1).

Table 3.1: Existing models of CSF and CFF. Note that CSF models perception continuously across a range of conditions while CFF only models the limit of temporal perception at high temporal frequencies. Unlike ours, none of these models accounts for spatial and temporal variation as well as eccentricity and luminance.

Model	Spat.	Temp.	Eccentr.	Lum.	Prop.
Tyler [86]	✗	✓	✗	✓	CFF
Kelly [67]	✓	✓	✗	✗	stCSF
Watson [69]	✓	✓	✗	✓	stCSF
Watson [70]	✓	✗	✓	✓	sCSF
This Chapter	✓	✓	✓	✓	CFF

Using our model, we predict potential bandwidth savings of factors up to 3,500× over unprocessed visual information and 7× over existing foveated models that do not account for the temporal characteristics of human vision.

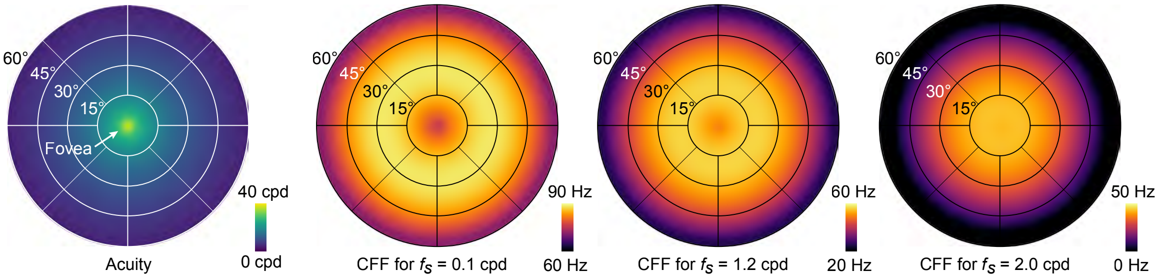

Figure 3.1: Foveated graphics techniques rely on eccentricity-dependent models of human vision. While such models are well understood for spatial acuity (left, [7]), our work is the first to experimentally derive a more comprehensive model for the spatio-temporal aspects over the retina under conditions close to VR and AR applications. As seen in the three plots on the right, we model CFF thresholds in an eccentricity-dependent manner. The CFF varies with spatial frequency f_s and luminance and it exhibits an anti-foveated effect, with the highest thresholds observed in the near–mid periphery of the visual field. Our perceptual model and its experimental validation could provide the foundation of future spatio-temporally foveated graphics systems.

Specifically, in this chapter we: (1) design and conduct user studies to measure and validate eccentricity-dependent spatio-temporal flicker fusion thresholds with a custom display, (2) fit several variants of an analytic model to this data and also extrapolate the model beyond the space of our measurements using data provided in the literature, including an extension for varying luminance, and (3) analyze bandwidth considerations and demonstrate that our model may afford significant additional savings for foveated graphics.

3.1 Estimating Flicker Fusion Thresholds

To develop an eccentricity-dependent model of flicker fusion, we need a display that is capable of showing stimuli at a high framerate and over a wide FOV. In this section, we first describe a custom high-speed VR display that we built to support these requirements. We then proceed with a detailed discussion of the user study we conducted and the resulting values for eccentricity and spatial frequency–dependent CFF we estimated.

3.1.1 Display Prototype

Our prototype display is designed in a near-eye display form factor to support a wide FOV. As shown on the left in Figure 3.2, we removed the back panel of a View-Master Deluxe VR Viewer and mounted a semi-transparent optical diffuser (Edmund Optics #47-679) instead of a display panel, which serves as a projection screen. This View-Master was fixed to an SR Research headrest to allow users to comfortably view stimuli for extended periods of time. To support a sufficiently high framerate, we opted for a Digital Light Projector (DLP) unit (Texas Instruments DLP3010EVM-LC Evaluation Board) that rear-projects images onto the diffuser towards the viewer. A neutral density (ND16) filter was placed in this light path to reduce the brightness to an eye safe level, measured to be $380\,\mathrm{cd/m^2}$ at peak.

The DLP has a resolution of 1280 × 720, and a maximum frame rate of 1.5 kHz for 1-bit video, 360 Hz for 8-bit monochromatic video, or 120 Hz for 24-bit RGB video. We positioned the projector such that the image matched the size of the conventional View-Master display. Considering the magnification of the lenses, this display provides a pixel pitch of 0.1' (arc minutes) and a monocular FOV of 80° horizontally and 87° vertically.

To display stimuli for our user study, we used the graphical user interface provided by Texas Instruments to program the DLP to the 360 Hz 8-bit grayscale mode, deemed sufficient for our measurements of CFF, which unlike CSF, does not require precise contrast tuning. The DLP was unable to support the inbuilt red, green and blue light-emitting diodes (LEDs) being on simultaneously, so we chose to use a single LED to minimize the possibility of artifacts from temporally multiplexing colors. Furthermore, since the HVS is most sensitive to mid-range wavelengths, the green LED (OSRAM: LE CG Q8WP) of peak wavelength 520 nm and a 100 nm full width at half maximum, was chosen so as to most conservatively measure the CFF thresholds. We used Python's PsychoPy toolbox [180] and a custom shader to stream frames to the display by encoding them into the required 24-bit RGB format that is sent to the DLP via HDMI.

Stimuli The flicker fusion model we wish to acquire could be parameterized by spatial frequency, rotation angle, eccentricity (i.e., distance from the fovea), direction from the fovea (i.e., temporal, nasal, etc.) and other parameters. A naive approach may sample across all of these dimensions,

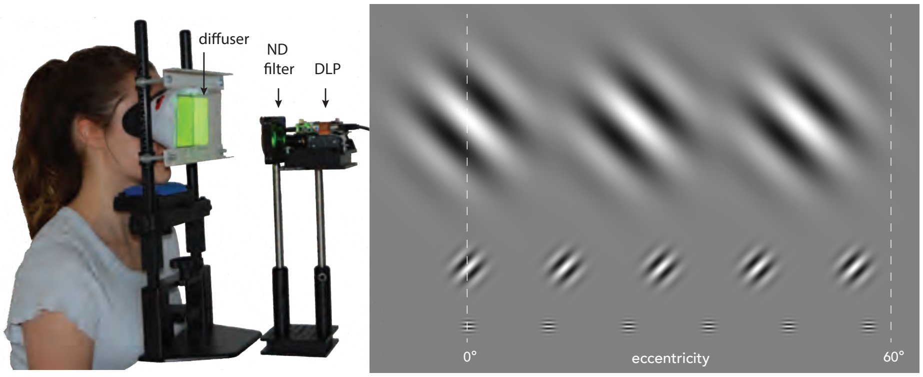

Figure 3.2: Our set-up for measuring user CFF thresholds. A photograph of a user in our custom VR display prototype is shown on the left. An ND filter is used to reduce the brightness of a DLP which projects onto a semi-transparent diffuser. On the right we show an illustration of the last 3 orders of test Gabor wavelets (described in Table 3.2) shown at a contrast of 1. Eccentricities were chosen to cover a FOV of 60° at each scale. Orientation was chosen randomly from $45^\circ, 90^\circ, 135^\circ$ and 180° at the point of beginning a QUEST staircase.

but due to the fact that each datapoint needs to be recorded for multiple subjects and for many temporal frequencies per subject to determine the respective CFFs, this seems infeasible. Therefore, similar to several previous studies [76, 82, 181], we make the following assumptions to make data acquisition tractable:

1. The left and right eyes exhibit the same sensitivities and monocular and binocular viewing conditions are equivalent. Thus, we display the stimuli monocularly to the right eye by blocking the left side of the display.

2. Sensitivity is rotationally symmetric around the fovea, i.e. independent of nasal, temporal, superior, and inferior direction, thus being only a function of absolute distance from the fovea. It is therefore sufficient to measure stimuli only along the temporal direction starting from the fovea.

3. Sensitivity is orientation independent. Thus, the rotation angle of the test pattern is not significant.

These assumptions allow us to reduce the sample space to only two dimensions: eccentricity e and spatial frequency f_s. Later we also analyze retinal illuminance l as an additional factor.

It is also worth observing that an eccentricity-dependent model that varies with spatial frequency must adhere to the uncertainty principle. That is, low spatial frequencies cannot be well localized in eccentricity. For example, the lowest spatial frequency of 0 cpd is a stimulus that is constant across the entire retina whereas very high spatial frequencies can be well localized in eccentricity. This behavior is appropriately modeled by wavelets. As such, we select our stimuli to be a set of 2D Gabor wavelets. Described in detail in Section 4.1.1, we use a scaled and shift version of the general form to be suitable for display, such that the such that the pattern modulates between 0 and 1, with an average gray level of 0.5. We also use the display specifications to convert the spatial locations on the screen x and x_0 from Equation A.1 to be defined in terms of eccentricities e and e_0 (described in Appendix A.1) The resulting stimulus exhibits three clearly visible peaks and smoothly blends into the uniform gray field which covers the entire field of view of our display and ends sharply at its edge with a dark background.

This choice of Gabor wavelet was motivated by many previous works in vision science, including the standard measurement procedure for the CSF [75] (see Section 2.1.5), and image processing [182, 183], where Gabor functions are frequently used for their resemblance to neural activations in human vision. For example, it has been shown that 2D Gabor functions are appropriate models of the transfer function of simple cells in the visual cortex of mammalian brains and thus mimicking the early layers of human visual perception [184, 185].

As a tradeoff between sampling the parameter space as densely as possible while keeping our user studies to a reasonable length, we converged on using 18 unique test stimuli, as listed in Table 3.2. We sampled eccentricities ranging from 0° to almost 60°, moving the fixation point into the nasal direction with increasing eccentricity, such that the target stimulus is affected as little as possible by the lens distortion. We chose not to utilize the full 80° horizontal FOV of our display due to lens distortion becoming too severe in the last 10°. We chose to test 6 different spatial frequencies, with the highest being limited to 2 cpd due to the lack of a commercial display with both high enough spatial and temporal resolution. However, as described in Chapter 2, human visual acuity is considerably higher; 60 cpd based on peak cone density [33] and 40–50 cpd based on empirical data [8, 12, 35]. Later we use existing data to an extension for these higher spatial frequencies (see Section 3.2.2). The Gaussian windows limiting the extent of the Gabor wavelets are scaled according to spatial frequency, i.e., $\sigma = 0.7/f_s$ such that each stimulus exhibits the same number of periods, defining the 6 wavelet orders. Finally, eccentricity values were chosen based on the radius of the wavelet order to uniformly sample the available eccentricity range, as illustrated on the right in Figure 3.2.

The wavelets were temporally modulated by sinusoidally varying the contrast from $[-1, 1]$ and added to the background gray level of 0.5. In this way, at high temporal frequencies the Gabor wavelet would appear to fade into the background. The control stimulus was modulated at 180 Hz, which is far above the CFF for all of the conditions, and thus appears invisible (non-flickering) to the observer.

3.1.2 User Study

Participants Nine adults participated (age range 18–53, 4 female). Due to the demanding nature of our psychophysical experiment, only a few subjects were recruited, which is common for similar low-level psychophysics (e.g. [13]). Furthermore, previous work measuring the CFF of 103 subjects

Table 3.2: Parameters of 18 test Gabor wavelets. We define 6 orders by spatial frequency (and radii) of the stimuli. The number and eccentricity locations per order were chosen based on radius to uniformly sample the available eccentricity range. f_s: spatial frequency, σ: wavelet standard deviation, e: eccentricity. (*) Note that in practice the extent was limited by our display FOV and $f_s = 0.0055$ cpd is used for the analysis in Section 3.2.1.

Order	f_s (cpd)	σ (°)	e (°)
0	0.000(*)	inf(*)	0.0
1	0.011	63.0	0.0
2	0.041	17.2	0.0, 19.2
3	0.154	4.6	0.0, 24.5, 48.2
4	0.571	1.2	0.0, 14.8, 29.2, 42.7, 55.0
5	2.000	0.5	0.0, 12.3, 24.4, 35.9, 46.8, 56.8

found a low variance [186], suggesting sufficiency of a small sample size. All subjects in this and the subsequent experiment had normal or corrected-to-normal vision, no history of visual deficiency, and no color blindness, but were not tested for peripheral-specific abnormalities. All subjects gave informed consent. The research protocol was approved by the Institutional Review Board at Stanford University.

Procedure To start the session, each subject was instructed to position their chin on the headrest such that several concentric circles centered on the right side of the display were minimally distorted (due to the lenses). The threshold for each Gabor wavelet was then estimated in a random order with a two-alternative forced-choice (2AFC) adaptive staircase designed using QUEST [187]. The orientation of each Gabor wavelet was chosen randomly at the beginning of each staircase from $0°, 45°, 90°$ and 135°. At each step, the subject was shown a small (1°) white cross for 1.5 s to indicate where they should fixate, followed by the test and control stimuli in a random order, each for 1 s. For stimuli at 0° eccentricity, the fixation cross was removed after the initial display so as not to interfere with the pattern. The screen was momentarily blanked to a slightly darker gray level than the gray background to indicate stimuli switching. The subject was then asked to use a keyboard to indicate which of the two randomly ordered patterns (1 or 2) exhibited more flicker. The ability to replay any trial was also given via key press and the subjects were encouraged to take breaks at their convenience. Each of the 18 stimuli were tested once per user, taking approximately 90 minutes to complete.

Results Mean CFF thresholds across subjects along with the standard error (vertical bars) and extent of the corresponding stimulus (horizontal bars) are shown in Figure 4.4. The table of measured values is included in Appendix A.3. The measured CFF values have a maximum above 90 Hz. This relatively large magnitude can be explained by the Ferry–Porter law [86] and the high adaptation luminance of our display. Similarly large values have previously been observed in corresponding conditions [88]. As expected, the CFF reaches its maximum for the lowest f_s values. This trend follows the Granit–Harper law predicting a linear increase of CFF with stimuli area [76]. For higher f_s stimuli, we observe an increase of the CFF from the fovea towards a peak between 10° and 30° of eccentricity. Similar trends have been observed by Hartmann et al. [76], including the apparent shift of the peak position towards fovea with increasing f_s and decreasing stimuli size. Finally, our subjects had difficulty to detect flicker for the two largest eccentricity levels for the maximum $f_s = 2\,\text{cpd}$. This is predictable as acuity drops close to or below this value for such extreme retinal displacements [7].

3.2 Fitting the Model

The measured CFFs establish an envelope of spatio-temporal flicker fusion thresholds at discretely sampled points within the resolution afforded by our display prototype. Practical applications, however, require these thresholds to be predicted continuously for arbitrary spatial frequencies and eccentricities. To this end, we develop a continuous eccentricity-dependent model for spatio-temporal flicker fusion that is fitted to our data. Moreover, we extrapolate this model to include spatial frequencies that are higher than those supported by our display by incorporating existing visual acuity data and we account for variable luminance adaptation levels by adapting the Ferry–Porter law [88].

3.2.1 A Model of Spatio-temporal Flicker Fusion

Each of our measured data points is parameterized by its spatial frequency f_s, eccentricity e, and CFF value averaged over all subjects. Furthermore, it is associated with a localization uncertainty determined by the radius of its stimulus u. In our design, u is a function of f_s and, for 13.5% peak contrast cut-off, we define $u = 2\sigma$ where $\sigma = 0.7/f_s$ to be the standard deviation of our Gabor wavelets.

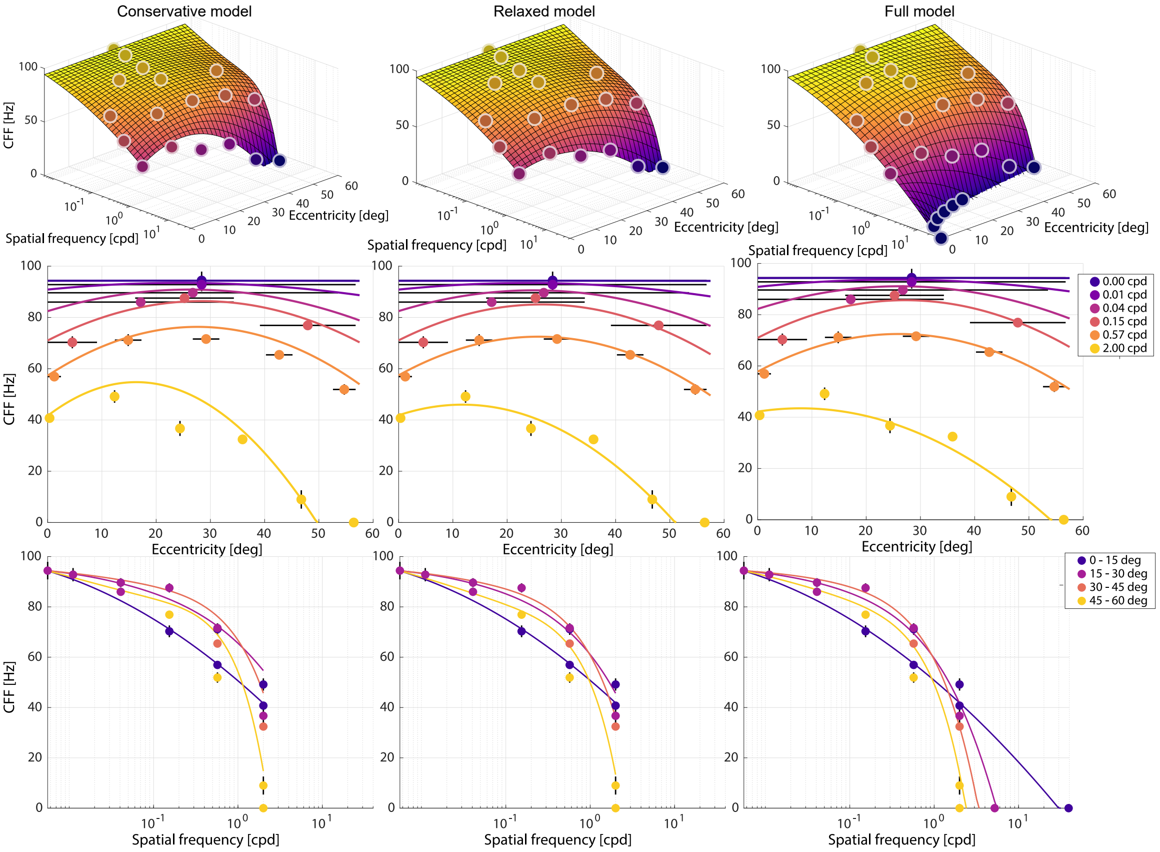

Figure 3.3: Three different parameter fits for our flicker fusion model $\Psi(e, f_s)$ (columns). Orthographic views of the eccentricity–CFF (second row) and spatial frequency–CFF (third row) planes show subject-averaged measured points along with their variances (standard error of means, vertical bars). The horizontal bars in the eccentricity dimension denote spatial extents of respective stimuli. The curves represent corresponding cross sections of the fitted models.

We formulate our model as:

$$
\begin{gathered}
\Psi(e, f_s) = \max(0, p_0\tau(f_s)^2 + p_1\tau(f_s) + p_2 \\
+ \ (\mathrm{p}_3\tau(f_s)^2 + p_4\tau(f_s) + p_5) \cdot \zeta(f_s)e \\
+ \ (\mathrm{p}_6\tau(f_s)^2 + p_7\tau(f_s) + p_8) \cdot \zeta(f_s)e^2) \\
\zeta(f_s) = \exp(p_9\tau(f_s)) - 1 \\
\tau(f_s) = \max(\log_{10} f_s - \log_{10} f_{s_0}, 0),
\end{gathered}
$$

where $\mathbf{p} = [p_0, \ldots, p_9] \in R^{10}$ are the model parameters (listed in Table 3.3), $\zeta(f_s)$ restricts eccentricity effects for small f_s and $\tau(f_s)$ offsets logarithmic f_s relative to our constant function cut-off.

We build on three domain-specific observations to find a continuous CFF model $\Psi(e, f_s) : R^2 \rightarrow R$ that fits our measurements. First, both our measurements and prior work indicate that the peak CFF is located in periphery, typically between 20° and 50° of eccentricity [76–78]. For both the fovea and far periphery the CFF drops again forming a convex shape which we model as a quadratic function of e. Second, because the stimuli with very low f_s are not spatially localized, their CFF does not vary with e. Consequently, we enforce the dependency on e to converge to a constant function for any f_s below $f_{s_0} = 0.0055$ cpd. This corresponds to half reciprocal of the full-screen stimuli visual field coverage given our display dimensions. Finally, following common practices in modeling the effect of spatial frequencies on visual effects, such as contrast [188] or disparity sensitivities [189], we fit the model for logarithmic f_s.

Before parameter optimization, we need to consider the effect of eccentricity uncertainty. The subjects in our study detected flicker regardless of its location within the stimuli extent $\mathbf{m} = [e \pm u]$ deg. Therefore, Ψ achieves its maximum within $\mathbf{m}$ and is upper-bounded by our measured flicker frequency $f_t \in R$. At the same time, nothing can be claimed about the variation of Ψ within $\mathbf{m}$ and therefore, in absence of further evidence, a conservative model has to assume that Ψ is not lower than f_t within $\mathbf{m}$. These two considerations delimit a piece-wise constant Ψ. In practice, based on previous work [76, 78] it is reasonable to assume that Ψ follows a smooth trend over the retina and its value is lower than f_t value at almost all eccentricities within $\mathbf{m}$.

Table 3.3: Parameters $p_{0...9}$ for our model fitted for conservative and relaxed assumptions as well as the full modeled extended using acuity data. The degrees-of-freedom adjusted R^2 shows the fit quality.

Model	Parameters										R^2
	p_0	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	
Conservative	94.4,	-10.1,	-4.08,	0.431,	-0.280,	0.0484,	-0.00912,	0.00679,	-0.00140,	1.56	0.889
Relaxed	94.3,	-10.1,	-4.06,	0.430,	-0.282,	0.0464,	-0.00929,	0.00672,	-0.00129,	1.58	0.938
Full	94.3,	-10.1,	-4.06,	0.435,	-0.281,	0.0440,	-0.00877,	0.00613,	-0.00111,	1.58	0.938

Consequently, in Table 3.3 we provide two different fits for the parameters. Our conservative model strictly follows the restrictions from the measurement and tends to overestimate the range of visible flicker frequencies which prevents discarding potentially visible signal. Alternatively, our relaxed model follows the smoothness assumption and applies the measured values as upper bound. To fit the parameters, we used the Adam solver in PyTorch initialized by the Levenberg–Marquardt algorithm and we minimized the mean-square prediction error over all extents $\mathbf{m}$. The additional constraints were implemented as soft linear penalties. To leverage data points with immeasurable CFF values, we additionally force $\Psi(e, f_s) = 0$ at these points. This encodes imperceptibility of their flicker at any temporal frequency. Figure 4.4 shows that the fitted Ψ represents the expected effects well. The eccentricity curves (row 2) flatten for low f_s and their peaks shift to lower e for large f_s. The conservative fit generally yields larger CFF predictions though it does not strictly adhere to the stimuli extents due to other constraints.

3.2.2 Extension for High Spatial Frequencies

Due to technical constraints, the highest f_s measured was 2 cpd. At the same time, the acuity of human vision has an upper limit of 60 cpd based on peak cone density [33] and 40–50 cpd based on empirical data [8, 12, 35] (see Section 2.1.4). To minimize this gap and to generalize our model to other display designs we extrapolate the CFF at higher spatial frequencies using existing models of spatial acuity.

For this purpose, we utilize the acuity model of Geisler and Perry [7]. It predicts acuity limit A for e as

$$A(e) = \ln(64)\frac{2.3}{0.106 \cdot (e + 2.3)}, \tag{3.1}$$

with parameters fitted to measurements of Robson and Graham [8]. Their study of pattern detection rather than resolution is well aligned with our own study design and conservative visual performance assessment. Similarly, their bright adaptation luminance of 500 cd/m^2 is also close to our display.

$A(e)$ predicts limit of spatial perception. We reason that at this absolute limit flicker is not detectable and, therefore, the CFF is not defined. We represent this situation by zero CFF values in the same way as for imperceptible stimuli in our study and force our model to satisfy $\Psi(e, A(e)) = 0$. In combination with our relaxed constraints we obtain our final full model as shown in Figure 4.4. It follows the same trends as our original model within the bounds of our measurement space and intersects the zero plane at the projection of $A(e)$.

3.2.3 Accounting for Adaptation Luminance

Our experiments were conducted at half of our display peak luminance $L = 380$ cd/m^2. This is relatively bright compared to the 50–200 cd/m^2 luminance setting of common VR systems [190]. Consequently, our estimates of the CFF are conservative because the Ferry–Porter law (see Section 2.1.6) predicts the CFF to increase linearly with logarithmic levels of retinal illuminance [88]. While the linear relationship is known, the actual slope and intercept varies with retinal eccentricity [88]. For this reason, we measured selected points from our main experiment for two other display luminance levels.

Four of the subjects from the model experiment performed the same procedure for a subset of conditions with a display modified first with one and then two additional ND8 filters yielding effective luminance values of 23.9 and 3.0 cd/m^2. We then applied a formula of Stanley and Davies [191] to

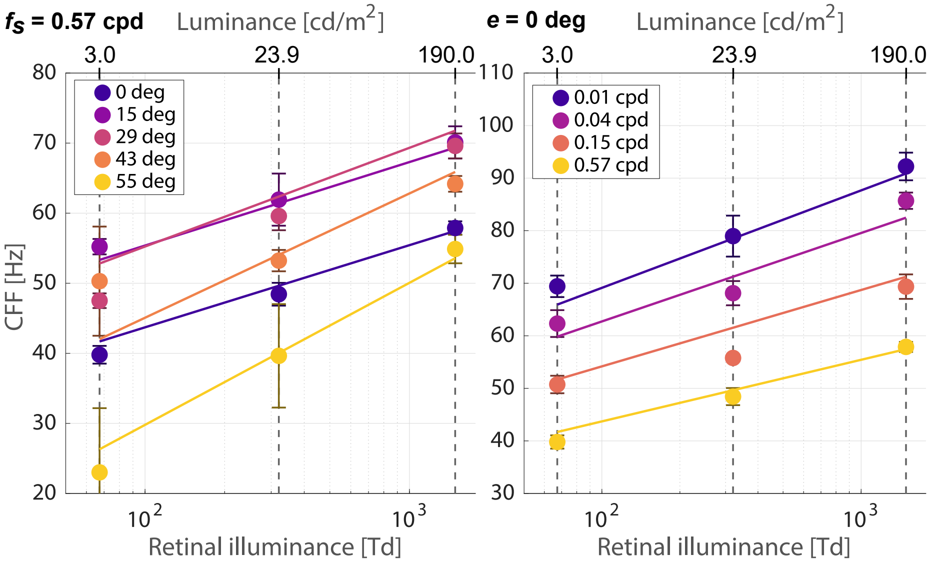

Figure 3.4: Luminance scaling of our full model fitted for $l_0 = 1488\,\mathrm{Td}$. The points represent mean measured CFF values and the lines the prediction of the transformed model. Vertical bars are standard errors. The left plot shows varying slopes across e (with $f_s = 0.57\,\mathrm{cpd}$). The right plot shows varying intercepts across f_s (with $e = 0\,\mathrm{deg}$).

compute the pupil diameter under these conditions as

$$d(L) = 7.75 - 5.75\left(\frac{(La/846)^{0.41}}{(La/846)^{0.41}+2}\right), \tag{3.2}$$

where $a = 80 \times 87 = 6960\,\mathrm{deg}^2$ is the adapting area of our display. This allows us to derive corresponding retinal illuminance levels for our experiments using $l(L) = \pi d(L)^2/4 \cdot L$ as 67.3, 321 and 1488 Td and obtain a linear transformation of our original model $\Psi(e, f_s)$ to account for L with an eccentricity-dependent slope as

$$\hat{\Psi}(e, f_s, L) = (s(e, f_s) \cdot (\log_{10}(l(L)/l_0)) + 1)\Psi(e, f_s) \tag{3.3}$$

$$s(e, f_s) = \zeta(f_s)(q_0 e^2 + q_1 e) + q_2 \tag{3.4}$$

where $l_0 = 1488$ Td is our reference retinal illuminance, $\zeta(f_s)$ encodes localization uncertainty for low f_s as in Equation 3.2.1 and $\mathbf{q} = [5.71 \cdot 10^{-6}, -1.78 \cdot 10^{-4}, 0.204]$ are parameters obtained by a fit with our full model.

Figure 3.4 shows that this eccentricity-driven model of Ferry–Porter luminance scaling models not only the slope variation over the retina but also the sensitivity difference over range of f_s well (degree-of-freedom-adjusted $\mathrm{R}^2 = 0.950$).

3.3 Model Validation

Our eccentricity-dependent spatio-temporal model is unique in that it allows us to predict what temporal information may be imperceptible for a certain eccentricity and spatial frequency. One possible application for such a model is in the development of new perceptual video quality assessment metrics (VQMs). Used to guide the development of different video codecs, encoders, encoding settings, or transmission variants, such metrics aim to predict subjective video quality as experienced by users. While it is commonly known that many existing metrics, such as peak signal-to-noise ratio (PSNR) and structural similarity index metric (SSIM), do not capture many eccentricity-dependent spatio-temporal aspects of human vision well, in this section we discuss a user study we conducted which shows that our model could help better differentiate perceivable and non-perceivable spatio-temporal artifacts. Furthermore, we also use the study to test the fit derived in the previous section, showing that our model makes valid predictions for different users and points beyond those measured in our first user study.

The study was conducted using our custom high-speed VR display with 18 participants, 13 of which did not participate in the previous user study. Users were presented with videos, made up of a single image frame perturbed by Gabor wavelet(s) such that when modulated at a frequency above the CFF they become indistinguishable from the background image. Twenty-two unique videos were tested, where the user was asked to rank the quality of the video from 1 ("bad") to 5 ("excellent"). We then calculate the difference mean opinion score (DMOS) per stimulus, as described in detail by Seshadrinathan et al. [192], and 3 VQMs, namely PSNR, SSIM and one of the most influential metrics used today for traditional contents: the Video Multimethod Assessment Fusion (VMAF) metric developed by Netflix [193]. The results of which are shown in Figure 3.5. We acknowledge that neither of these metrics has been designed to specifically capture the effects we measure. This often results in uniform scores across the stimuli range (see e.g. Figure 3.5 III). We include them into our comparison to illustrate these limits and the need for future research in this direction.

The last row shows a comparison of all measured DMOS values with each of the standard VQMs and our own flicker quality predictor. This is a simple binary metric that assigns videos outside of the CFF volume where flicker is invisible a good label while it assigns a bad label to others. Our full $\Psi(e, f_s)$ model was used for this purpose. We computed Pearson Linear Correlation Coefficients (PLCC) between DMOS and each of the metrics while taking into account their semantics. Only our method exhibits statistically significant correlation with the user scores ($r(20) = 0.988$, $p < 0.001$ from a two-tail t-test).

Further, we broke our stimuli into 5 groups to test the ability of our model to predict more granular effects.

1. *Unseen stimulus.* We select $e = 30°$ and $f_s = 1.00$ cpd, a configuration not used to fit the model, and show that users rate the video to be of poorer quality (lower DMOS) when it is

modulated at a frequency lower than the predicted CFF (see panel I in Figure 3.5).

2. *f_s dependence.* We test whether our model captures spatio-temporal dependence. Lowering f_s with fixed radius makes the flicker visible, despite not changing f_t (panel II).

3. *e change.* By moving the fixation point towards the outer edges of the screen, thereby increasing e, we show that our model captures and predicts the higher temporal sensitivity of the mid-periphery. Furthermore, by moving the fixation point in both nasal and temporal directions, we validate the assumed symmetry of the CFF. As a result, users were able to notice flicker that was previously imperceptible in the fovea (panel III).

4. *Direction independence.* We confirm our assumption that sensitivity is approximately rotationally symmetric around the fovea. A single visible stimulus from Group 1 was also added for the analysis as a control (panel IV).

5. *Mixed stimuli.* We show that the model predictions hold up even with concurrent viewing of several stimuli of different sizes, eccentricities and spatial frequencies (panel V).

The perturbation applied to the videos in Groups 1–3 only varied in modulation frequency or eccentricity with respect to a moving gaze position. In response, the dependency of other metrics on f_t is on the level of noise, since changes related to variation of f_t or f_s of a fixed-size wavelet average out over several frames, and moving the gaze position does not change the frames at all. Our metric is able to capture the perceptual effect, maintaining significant correlation with DMOS ($r(3) = 0.994$, $p = 0.001$, $r(2) = 0.983$, $p < 0.05$ and $r(3) = 0.977$, $p < 0.001$ respectively).

Groups 4 and 5 were designed to be accumulating sets of Gabor wavelets, but all with parameters that our model predicts as imperceptible. We observe that VQMs drop with increasing total distortion area, while the DMOS scores stay relatively constant and significantly correlated with our metric ($r(2) = 1.000$, $p < 0.001$ and $r(3) = 0.993$, $p < 0.01$ respectively). These results indicate that existing compression schemes may be able to utilize more degrees of freedom to achieve higher compression, without seeing a drop in the typical metrics used to track perceived visual quality. The exact list of chosen parameters, along with the CFFs predicted by our model are listed in Appendix A.4.

In conclusion, the study shows that our metric predicts visibility of temporal flicker for independently varying spatial frequencies and retinal eccentricities. While existing image and video quality metrics are important for predicting other visual artifacts, our method is a novel addition in this space that significantly improves perceptual validity of quality predictions for temporally varying content. In this way, we also show that our model may enable existing compression schemes to utilize more degrees of freedom to achieve higher compression, without compromising several conventional VQMs.

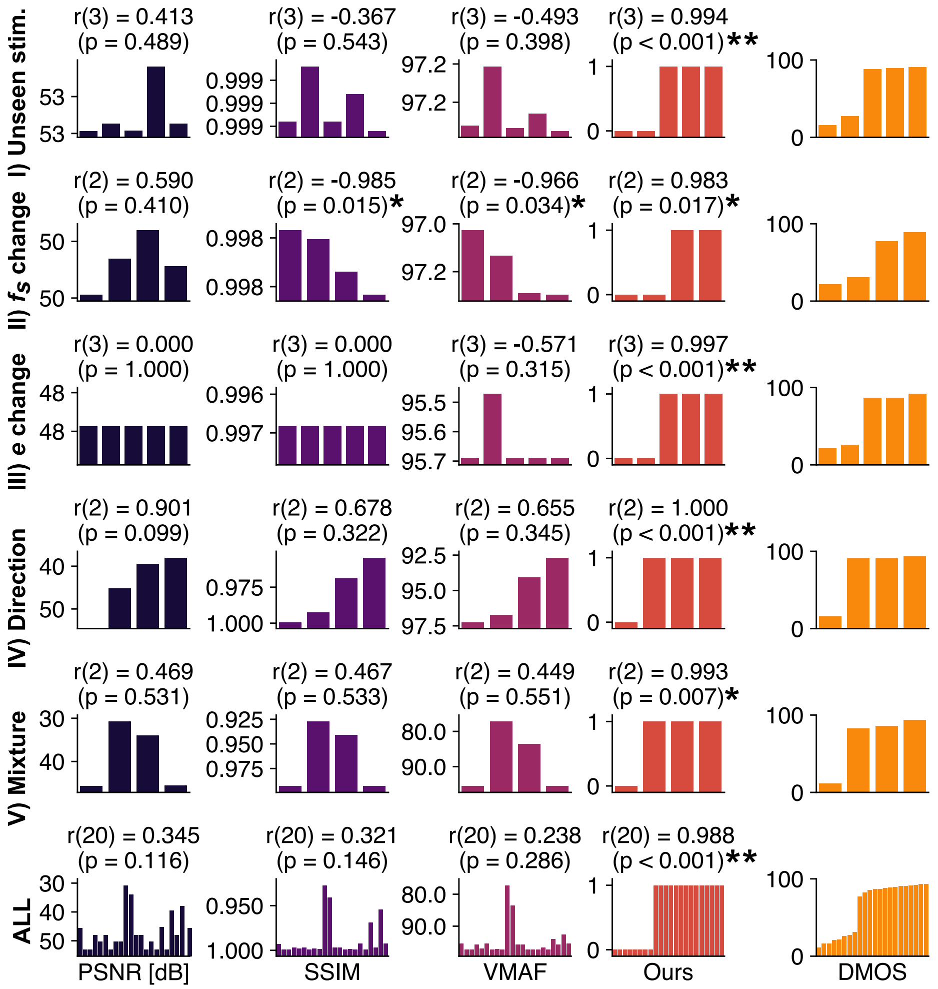

Figure 3.5: Results of our validation study. Each row represents one of the 5 tested effects along with a cumulative plot in the last row. Results of different metrics (vertical columns) are compared to DMOS (right column). Each bar represents the score for one stimulus rescaled so that a higher bar indicates a better quality. PLCC of each metric and DMOS are shown along the p-values (t-test). (*) and (**) mark statistical significance at $p = 0.05$ and $p = 0.001$ levels respectively.

3.4 Analyzing Bandwidth Considerations

Our eccentricity-dependent spatio-temporal flicker fusion model defines the gamut of visual signals perceivable to the HVS. Hence any signal outside of this gamut can be removed to save bandwidth in computation or data transmission. In this section we provide a theoretical analysis of the compression gain factors this model may potentially enable for foveated graphics applications when used to allocate resources such as bandwidth. In practice, developing foveated rendering and compression algorithms holds other nuanced challenges and thus the reported gains represent an upper bound.

To efficiently apply our model, a video signal should be described by a decomposition into

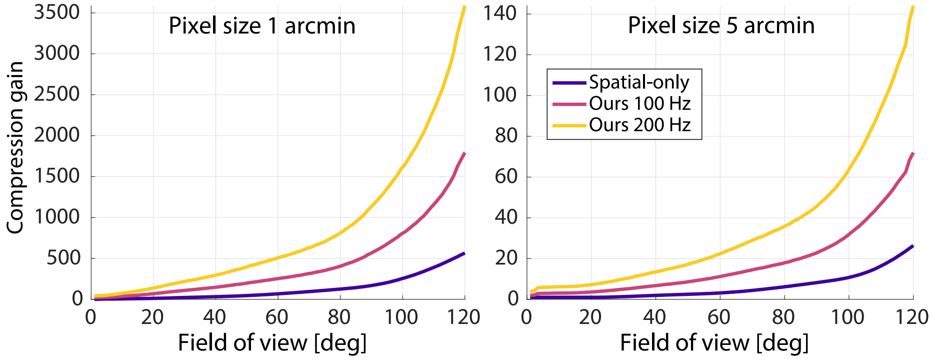

Figure 3.6: Compression gain analysis as a fraction of original and retained decomposition coefficients depending on the screen pixel density (left and right plots) and the covered visual field (horizontal axis). Note the difference in the gain axes scales.

spatial frequencies, retinal eccentricities, and temporal frequencies. We consider discrete wavelet decompositions (DWT) as a suitable candidate because these naturally decompose a signal into eccentricity-localized spatio-temporal frequency bands. Additionally, the multitude of filter scales in the hierarchical decomposition closely resembles scale orders in our study stimuli.

For the purpose of this thought experiment, we use a biorthogonal Haar wavelet, which, for a signal of length N, results in $\log_2(N)$ hierarchical planes of recursively halving lengths. The total number of resulting coefficients is the same as the input size. From there our baseline is retaining the entire original set of coefficients yielding compression gain of 1.

For traditional spatial-only foveation, we process each frame independently and after a 2D DWT we remove coefficients outside of the acuity limit. We compute eccentricity assuming gaze in the center of the screen and reject coefficients for which $\Psi(e, f_s) = 0$. From our model definition, such signals cannot be perceived regardless of f_t as they lie outside of the vision acuity gamut. The resulting compression gain compared to the baseline can be seen for various display configurations in Figure 3.6.

Next, for our model we follow the same procedure but we additionally decompose each coefficient of the spatial DWT using 1D temporal DWT. This yields an additional set of temporal coefficients f_t and we discard all values with $\Psi(e, f_s) < f_t$.

For this experiment, we assume a screen with the same peak luminance of $380\,\mathrm{cd/m^2}$ as our display prototype and a pixel density of 60 pixel per degree (peak $f_s = 30\,\mathrm{cpd}$) for approximately retinal display or a fifth of that for a display closer to existing VR systems. The tested aspect ratio of 165:135 approximates the human binocular visual field [29]. We consider displays with a maximum framerate of 100 Hz (peak reproducible $f_t = 50\,\mathrm{Hz}$) and 200 Hz, capable of reaching the maximum CFF in our model. The conventional spatial-only foveation algorithm is not affected by this choice.

Figure 3.6 shows that this significantly improves compression for both small and large FOV displays regardless of pixel density. This is because the spatial-only compression needs to retain all temporal coefficients in order to prevent the worst-scenario flicker at e and f_s with highest CFF. Therefore, it only discards signal components in the horizontal $e \times f_s$ plane. On the other hand, our scheme allows to also discard signal in the third temporal dimension closely copying the shape of the Ψ which allows to reduce retained f_t for both fovea and far periphery in particular for high f_s coefficients which require the largest bandwidth.

3.5 Discussion

The experimental data we measure and the models we fit to them further our understanding of human perception and lay the foundation of future spatio-temporal foveated graphics techniques. Yet, several important questions remain to be discussed.

Limitations and Future Work First, our data and model work with CFF, not the CSF. CFF models the temporal thresholds at which a visual stimulus is predicted to become (in)visible. Unlike the CSF, however, this binary value of visible/invisible does not model relative sensitivity. This makes it not straightforward to apply the CFF directly as an error metric, for example to optimize foveated rendering or compression algorithms. Obtaining an eccentricity-dependent model for the spatio-temporal CSF is an important goal for future work, yet it requires a dense sampling of the full 3D space spanned by eccentricity, spatial frequency, and time with user studies. The CFF, on the other hand only requires us to determine a single threshold for the 2D space of eccentricity and spatial frequency. Considering that obtaining all 18 sampling points in our 2D space for a single user already takes about 90 minutes, motivates the development of a more scalable approach to obtaining the required data in the future.

Although we validate the linear trends of luminance-dependent behavior of the CFF predicted by the Ferry–Porter law, it would be desirable to record more users for larger luminance ranges and more densely spaced points in the 2D sampling space. Again, this would require a significantly larger amount of user studies, which seems outside the scope of this work. Similarly, we extrapolate our model for f_s above the 2 cpd limit of our display. Future work may utilize advancements in display technology or additional optics to confirm the model extension. We also map the spatio-temporal thresholds only for monocular viewing conditions. Some studies have suggested that such visual constraint lowers the measured CFF in comparison to binocular viewing [194, 195]. Further work is required to confirm the significance of such an effect in the context of displays.

It should also be noted that our model assumes a specific fixation point. We did not account for the dynamics of ocular motion, which may be interesting for future explorations. By simplifying temporal perception to be equivalent to the temporal resolution of the visual system we also do not account for effects caused by our pattern sensitive system, such as our ability to detect spatio-temporal changes like local movements or deformations. Similarly, we do not model the effects of crowding in the periphery, which has been shown to dominate spatial resolution in pattern recognition [196]. Finally, all of our data is measured for the green color channel of our display and our model assumes rotational invariance as well as orientation independence of the stimulus. A more nuanced model that studies variations in these dimensions may be valuable future work. All these considerations should be taken into account when developing practical compression algorithms, which is a task beyond the scope of this paper.

Finally, we validate that existing error metrics, such as PSNR, SSIM, and VMAF, do not adequately model the temporal aspects of human vision. The correlation between our CFF data and the DMOS of human observers is significantly stronger, motivating error metrics that better model these temporal aspects. Yet, we do not develop such an error metric nor do we propose practical foveated compression or rendering schemes that directly exploit our CFF data. These investigations are directions of future research.

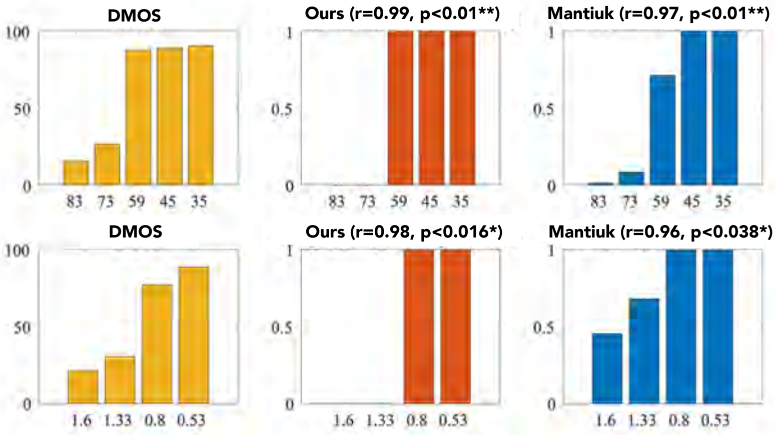

Figure 3.7: Mantiuk et al. [56] adapted a flicker detection flicker algorithm [197] using their unified CSF model, stelaCSF, to compared against the results of our validation study (Section A.4). The top row shows condition I, unseen stimulus, where a fixed Gabor is presented flickering at various frequencies. The bottom row shows condition II, f_s dependence, where flickering Gabors of fixed size but varying spatial frequency are shown at a single temporal frequency. They reproduce the mean DMOS score of the user study on the left, show the results obtained by our model in the center, and then the results of their modified algorithm on the right. Image adapted from Mantiuk et al. [56].

New Developments Since this work was published, we've seen a high interest and more researchers inspired to build even more comprehensive models. In particular, Mantiuk et al. [56] proposed the first unified CSF model, StelaCSF, which accounts for all major dimensions of the stimulus: spatial and temporal frequency, eccentricity, luminance, and area by combining data from several previous papers. Unlike our model, the CSF can model relative sensitivity, rather than just binary visibility. In their analysis, the authors use their model to modify the flicker predictor by Denes and Mantiuk [197] to compare against the prediction of our model. As shown in Figure 3.7, with the introduction of stelaCSF, the algorithm is able to accurately predict the detection of flicker for the conditions studied in our validation experiment (Section 3.3).

3.6 Summary

At the convergence of applied vision science, computer graphics, and wearable computing systems design, foveated graphics techniques will play an increasingly important role in emerging VR and AR display systems. With our work, we hope to contribute a valuable foundation to this field that helps spur follow-on work exploiting the particular characteristics of human vision we model.

Chapter 4

Towards Attention-aware Foveated Rendering

Foveated graphics present a promising solution to the bandwidth challenges preventing perceptually realistic VR and AR displays. As we describe in the previous chapter, these techniques have the potential to achieve higher gain factors by exploiting not just limitations in spatial sensitivity across the retina, but also other limitations of the HVS. For instance, the fact that our perceptual abilities also depend on higher-level cognitive processing. In fact we rarely see what we are looking at unless we direct sufficient cognitive resources [198], explaining many phenomena including change blindness [90]. *Visual attention* refers to a set of cognitive operations that helps us selectively process the vast amounts of information with which we are confronted, allowing us to focus on a certain location or aspect of the visual scene, while ignoring others [89]. Most often, we direct our attention *overtly*, by moving our eyes towards a location, but we can also direct attention to an area in the periphery *covertly*, via a mental shift. As described in Chapter 2, several studies have demonstrated that, under many conditions, increasing the amount of attention allocated to a visual task can enhance performance [94, 95]. In a similar manner, dividing attention between tasks can reduce visual peformance, resulting in lower contrast sensitivity [96,97], visual acuity [98], and speed of information accrual [99].

However, as illustrated in Figure 4.1, existing models of contrast sensitivity and visual acuity across the visual field are built on experiments where subjects are asked to covertly direct high levels of visual attention to a discrimination task in the periphery. Thus, for most scenarios in the real world, as well as VR and AR, where most of our attention is directed overtly (at our gaze position), we are likely overestimating our perceptual abilities in the periphery. Consequently, current efficacies of foveated graphics are too conservative in most real-use cases.

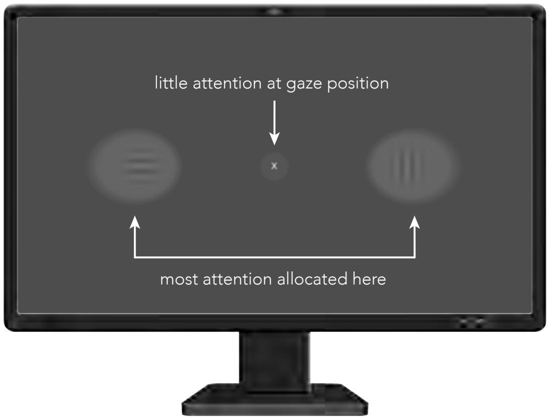
little attention at gaze position
x
most attention allocated here

Figure 4.1: Existing eccentricity-dependent CSF models are built on experiments where users maintain central fixation, but covertly direct most of their attention to the discrimination task in the periphery. The size and intensity of the spotlights illustrate visual attention distribution.

In this chapter, we propose to account for the effect of covert attention when modeling human contrast sensitivity. To this goal, we investigate the effect of modulating the amount of attention allocated to the contrast discrimination task in the periphery, by forcing attention to the fovea with a visually demanding task. Specifically, we compare the standard approach to measuring contrast sensitivity, where a low amount of attention is directed to the fovea ("low"), to scenarios where part or most of the attention is directed there ("medium" and "high"). As illustrated on the left of Figure 4.2, we show that in such instances, peripheral contrast discrimination thresholds elevate significantly and introduce the first attention-aware CSF model. Furthermore, we motivate the development of future foveation models with another user study, demonstrating that tolerance for foveation (i.e., peripheral blur) is significantly higher when the user is concentrating on a task in the fovea (illustrated on the right of Figure 4.2). Analysis of our model predicts potential bandwidth savings over 7 times higher than those afforded by current models.

Specifically, in this chapter we: (1) design and conduct user studies to measure and validate eccentricity-dependent effects of attention on contrast discrimination and foveation efficacy, (2) introduce the first analytic model of contrast sensitivity across eccentricity under varying attention, and (3) analyze bandwidth considerations and demonstrate that our model may afford significant bandwidth savings over existing foveated graphics techniques.

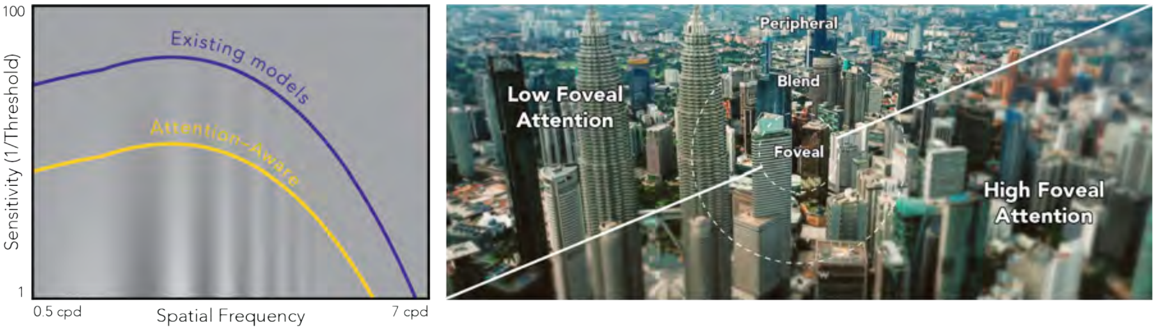

Figure 4.2: Foveated graphics techniques rely on eccentricity-dependent models of human vision. However, existing models of contrast sensitivity (left, purple line, shown for a fixed eccentricity) do not take into account allocation of visual attention across the visual field. Our work is the first to experimentally derive a model for eccentricity-dependent attention-aware sensitivity (left, yellow line). As illustrated on the right, when the user is focused on a task in the fovea, less attention is directed to the periphery and a higher level of foveation (i.e., peripheral blur) is possible without impacting the perceived visual quality. (City image by Pok Pie from Pexels: https://www.pexels.com/photo/4847105/)

4.1 A Model for CSF under Divided Attention

While modulating the amount of attention has been shown to affect contrast discrimination thresholds (see Section 2.1.7), insufficient data and a lack of existing models prevent this effect from being applied to existing CSF models and hence foveated graphics. In this section, we provide a detailed discussion of the user study we conducted and the model we fit to predict the effect of modulating peripheral attention on the CSF.

4.1.1 Measuring CSF

Similar to Chapter 4, the CSF model we wish to acquire could be parameterized by temporal frequency, spatial frequency, rotation angle, eccentricity (i.e., distance from the fovea), direction from the fovea (i.e., temporal, nasal, etc.) and other parameters. Here, we make the sampling tractable by nominally select 3 points across the eccentricity (e) available with our display (see Section 4.1.3), a spatial frequency (f_s) of 2 cpd, and a diameter of 5° for the furthest point. We then use the *cortical magnification factor* to scale the spatial frequencies and diameters at the other retinal positions (see stimuli No. 1–3 in Table 4.1) such that the discrimination thresholds should be approximately the same (see Appendix B.1 for more detail).

In order to obtain the contrast discrimination thresholds, we use the standard approach described in Chapter 2, including use of 2D Gabor wavelets. A detailed description of this stimuli is included in Section 2.1.5. During each trial, two wavelets are simultaneously presented for 500 ms, centered at a given eccentricity, to the left and right side of the central fixation position. Each grating is randomly orientated either horizontally or vertically and the user is asked to discriminate whether the wavelets are of the same or different orientations.

Table 4.1: Parameters of tested Gabor patches. For measuring the model (shown above the divider), we chose a diameter of 5° at the highest eccentricity of 21° to utilize the full field of view of our display and a spatial frequency f_s of 2 cpd, then use the cortical magnification factor to scale these parameters at 7° and 14° eccentricity. Gabor's sigma was defined as 20% of the diameter. For validation, we chose 2 sets of Gabor parameters (shown below the divider) used to fit StelaCSF [56], namely measurements taken by Virsu and Rovamo [62] and Wright and Johnson [199]. Stimulus No. 5 was also tested at two additional adaption luminances, 58 and 116 cd/m^2 (No. 6 and 7).

No.	Eccentricity (°)	Diameter (°)	Spatial Freq. (cpd)	Adaptation Lum. (cd/m^2)
1	7	2.16	4.62	28
2	14	3.58	2.79	28
3	21	5	2	28
4	9.25	1.7	2	28
5	15	5	4	28
6	15	5	4	58
7	15	5	4	116

4.1.2 The Attention-modulating Task

Inspired by Huang and Dobkins [97], we present a rapid serial visual presentation (RSVP) at the fixation cross in order to modulate the amount of attention paid to the peripheral contrast discrimination task. The RSVP stimulus consists of N $1° \times 1°$ letters, each lasting $500/N$ ms with 0 ms blank in between, such that the task lasts the total display duration of the peripheral Gabor wavelets. The color of the letters alternate between red and green (scaled to be approximately isoluminant with the background), where the initial color is randomized across trials, and the user is asked to identify the color of the "target letter" (the letter "T", which appears only once in a given sequence). Increasing N increases the difficulty of the task and should force more attention to the fovea, at the cost of reduced attention to the periphery. Consequently, three task levels were chosen to have an N of 1 (easy), 4 (medium) and 6 (hard), to force "low" , "medium" , and "high" levels of attention to the fovea. The target letter "T" was also adjusted such that for the "medium" and "high" attention tasks it would not appear in the first 3rd of letters to avoid users obtaining the color early enough to shift their attention to the periphery before the trial ended.

4.1.3 User Study

Setup Due to the need to display high resolution stimuli across a wide field of view, we conduct our study using a 34 inch, 144 Hz Dell Curved Gaming Monitor (Model No. S3422DWG, see Figure 4.3). This display has an adjustable backlight, allowing us to tune luminance. For this study we use a setting that gives a minimum and maximum luminance of 0.6 cd/m^2 and 104 cd/m^2, respectively, and a gamma of 1.89. The neutral gray background triggered luminance adaptation to 28 cd/m^2. A 2×2 spatial dithering was used to avoid visible color banding in the low-contrast stimuli. We used Python's PsychoPy toolbox [180] and a custom shader to stream frames to the display by wired HDMI connection. All subjects were tested in a well-lit room and viewed the video display binocularly from an SR Research headrest situated 94 cm away, thus giving a field of view of $46° \times 20°$ and a resolution of 71 ppd (pixels per degree of visual angle). Pupil Labs Core eye trackers were mounted to the headrest to verify central gaze fixation throughout all studies (see Section 2.2.1 for specifications).

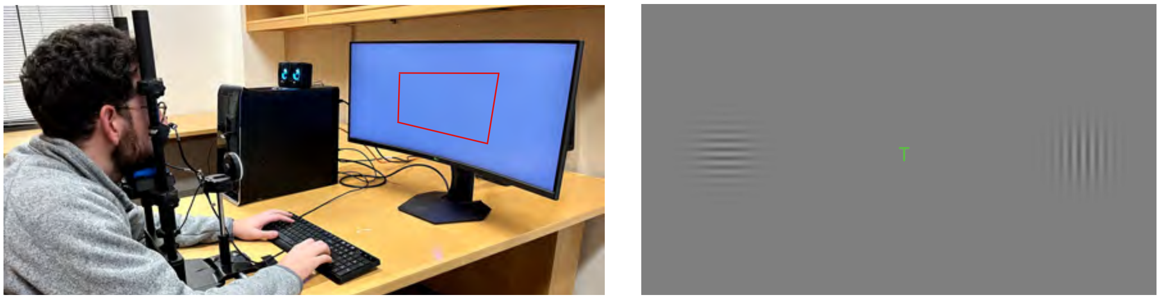

Figure 4.3: Our setup for measuring contrast thresholds under differing attention conditions. A photograph of the user study setup is shown on the left. An enlarged illustration of the stimulus on the screen (outlined in red) is shown on the right. The central RSVP letter task with the Gabor wavelets centered at e to the left and right. The brightness of the letter "T" and the contrasts of the Gabors have been exaggerated for visibility.

Subjects Ten adults participated (age range 23–29, 2 female). Due to the demanding nature of our psychophysical experiment, only a few subjects were recruited, which is common for similar low-level psychophysics (see e.g. [13]). All subjects in this and subsequent experiments had normal or corrected-to-normal vision, no history of visual deficiency, and no color blindness, but were not tested for peripheral-specific abnormalities. All subjects gave informed consent. The research protocol was approved by the Institutional Review Board at Stanford University.

Procedure To begin the study, subjects were set up in a comfortable position on the headrest and the eye tracker was calibrated using a 5-point screen calibration [200]. The thresholds for each contrast condition (stimuli No. 1–3 in Table 4.1) was then estimated in a random order. For each condition, a two-alternative forced-choice (2AFC) adaptive staircase designed using QUEST [187] was used to measure the contrast discrimination threshold for each attention condition, starting with the "low" , then the "medium" and ending with the "high" foveal attention condition. At each step, the subject was shown a small (1°) white fixation cross for 1.2 s to indicate where they should fixate, followed by the attention-modulating task and contrast stimuli for 500 ms, then a Gaussian white noise screen for 1 s (to reduce after images). The subject was then given 10 s to indicate via different sets of marked buttons on a keyboard whether the target letter "T" was red or green, followed by whether the contrast patterns were of the same or different orientations. If the subject failed to answer during that time, the trial would be replayed. Each of the 3 test conditions at each of the 3 attention conditions were tested twice per subject, taking approximately 90 minutes, with subjects encouraged to take breaks between staircases.

Results Mean contrast thresholds across subjects are shown in Figure 4.4a (see Appendix B.2 for a table of mean measured values). It can be seen that the contrast thresholds are almost identical for the "low" attention condition, agreeing with the theory of *cortical magnification* described by Virsu and Rovamo [62, 201]. For the "medium" attention condition, however, the contrast thresholds do increase significantly with eccentricity ($p < 0.05$, paired t-test between neighboring eccentricities), almost 2× for 7° and over 3× for 21° (see Figure 4.4b). Similarly, the "high" attention condition exhibits up to 4× threshold increase within our measured eccentricity range ($p < 0.05$). The increase in gain factors with eccentricity is consistent with work by Staugaard et al. [202] who showed a decrease in attentional capacity with increasing stimulus eccentricity, when stimuli are scaled in size to account for cortical magnification. Furthermore, we observe significant differences between individual attention modes across the eccentricity ranges ($p < 0.01$ for most pairs, $p < 0.05$ for the "medium" and "high" attention, paired t-test with Bonferroni correction). This confirms our assumption that increasing the task difficulty will shift attention towards the fovea at a cost to sensitivity in the periphery. On the other hand, despite the considerable difference between the "low" and "medium" condition gradients, the gradients of the "medium" and "high" conditions are surprisingly similar, suggesting that the effect of attention modulation is non-linear.

4.1.4 Per-condition Model

The observed increase in thresholds for larger eccentricities is nearly linear with a small distortion which we describe using the square root of eccentricity to fit attention-dependent contrast threshold models:

$$t_a(e) = p_0\sqrt{e} + p_1 \tag{4.1}$$

where a is denotes one of our foveal attention conditions ("low" , "medium" or "high"). See Figure 4.4a for plots and Table 4.2 for parameters.

As contrast sensitivity varies among observers, we are primarily interested in the relative attention gain represented by threshold elevations defined with respect to the "low" attention baseline

Table 4.2: Fitted parameters of our attention-aware contrast threshold model $t_a(e)$. R^2 is the coefficient of determination.

a	p_0	p_1	R^2
"low"	$9.672 \cdot 10^{-4}$	$2.741 \cdot 10^{-2}$	0.705
"medium"	$2.737 \cdot 10^{-2}$	$-1.620 \cdot 10^{-2}$	1.000
"high"	$2.714 \cdot 10^{-2}$	$1.612 \cdot 10^{-2}$	0.956

condition as:

$$g_a(e) = \frac{t_a(e)}{t_{\text{low}}(e)} \tag{4.2}$$

Assuming orthogonality of the attention effect and other independent parameters of the stimulus, we can formulate the attention-aware contrast sensitivity as:

$$S_a(e, \cdots) = S(e, \cdots)\frac{1}{g_a(e)} \tag{4.3}$$

where S is any of the CSF models discussed in Chapter 2 (Section 2.1.5). In Section 4.1.6 we use the StelaCSF [56] model.

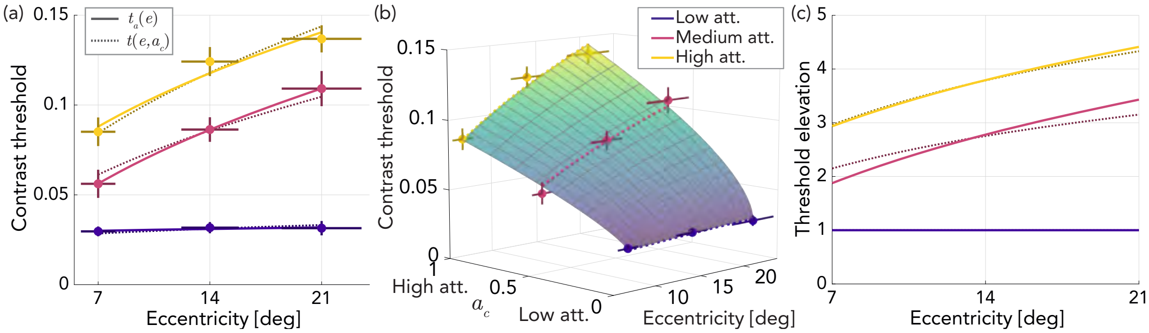

Figure 4.4: Main study: (a) The mean measured contrast thresholds and the fitted attention curves for the per-condition model $t_a(e)$ (Equation 4.1, full lines) and the unified model $t(e, a_c)$ (Equation 4.4, dotted lines). The horizontal bars display extent of the Gabors. The vertical error bars show standard error. (b) A continuous attention-eccentricity fit of the unified model $t(e, a_c)$ (Equation 4.4). (c) The attention gains $g_a(e)$ relative to the "low" foveal attention condition computed for each of the two models.

4.1.5 Unified model

Additionally, we explore a speculative model unifying the eccentricity e with a continuous interpretation of the attention condition $a_c \in [0, 1]$ where $\{$"low" $\to 0$, "medium" $\to 0.5$, "high" $\to 1\}$. We design this model as an attention-dependent sweep between the per-attention curves, parameterized relative to our lowest eccentricity of 7°. We model the dependency for the slope and intercept separately using two gamma curves $a_c^{\gamma_s}$ and $a_c^{\gamma_i}$ to account for the non-linear perception of the different attention conditions. Due to the extreme non-linearity of the slope development we constrain γ_s to 0.5 and fit:

$$t(e, a_c) = \Psi\left(s_0, s_1, a_c^{\gamma_s}\right) \cdot \left(\sqrt{e} - \sqrt{7}\right) + \Psi\left(i_0, i_1, a_c^{\gamma_i}\right) \tag{4.4}$$

to our measured data. Here, $\{s_0, s_1, i_0, i_1, \gamma_i\} = \{0.00243,\ 0.0307,\ 0.0285,\ 0.0844,\ 0.771\}$ are the fitted parameters (DoF-adjusted $\mathrm{R}^2 = 0.973$) and $\Psi(\alpha, \beta, w) = \alpha(1-w)+\beta w$ is a linear interpolation function.

We compare the resulting unified model (shown in Figure 4.4b) to our per-condition models $t_a(e)$ (in Figure 4.4a). Despite the lower parameter count, the unified model fits the measured data within the measurement errors. While the unified model allows for convenient interpolation, we argue for fitting task-specific models in practice, because the connection between the task and attention is highly individual and not well understood. Hence, we use our per-condition models $t_a(e)$ (Equation 4.1) throughout the rest of this paper wherever not explicitly specified otherwise.

4.1.6 Validation Experiments

In Equation 4.3, we apply attention correction as a multiplicative factor under an assumption of orthogonality between the two functions. If this assumption holds, the difference between new thresholds predicted by our model and their measured values should be low. We test this by measuring the attention gains for four new stimuli with a different cortical magnification and adaptation luminance levels than in our main study. We then compare the thresholds obtained by direct measurement with the thresholds predicted by our attention gain $g_a(e)$.

Experiment We use the same experiment procedure as for the main study, except with two parameter sets used to fit StelaCSF [56], a recently demonstrated unified model of CSF (see stimuli No. 4 and 5 in Table 4.1). These points were selected from the only 2 datasets measured using stationary stimuli outside the fovea, with spatial frequency, eccentricity and size as different as possible to the stimuli used to fit our attention-aware CSF model. Additionally, we test effect of varying luminance adaptation on one of these datapoints by adjusting the backlight of our display (see stimuli No. 6 and 7) .

Subjects Eleven adults participated (age range 23-29, 5 female), six for stimuli No. 4 and 5 and five for stimuli No. 6 and 7. Only three of these subjects participated in the main study.

Results We measured mean thresholds for the validation stimuli (No. 4–7) and the “low” attention condition as 0.032, 0.045, 0.57 and 0.51, which we use as baselines for a relative multiplicative adjustment of our measurements to corresponding predictions of StelaCSF. We compute Interquartile

Range (IQR) of this multiplicative factor to detect outliers. We treat each of the 2 per-user repetitions as a single data sample and we remove a total of 3 strong outliers with offset of 4 or more IQR from the quartiles. We then apply this base adjustment consistently to all individual measurements to remove variability of the base sensitivity performance among users and instead focus on relative gains between attention conditions (see Figure 4.5b). The resulting adjusted measurements are then compared with the contrasts predicted by the original attention-unaware StelaCSF model and our derived attention-aware CSF model $S_a(e, \cdots)$. We compute the error of both models in Figure 4.5c.

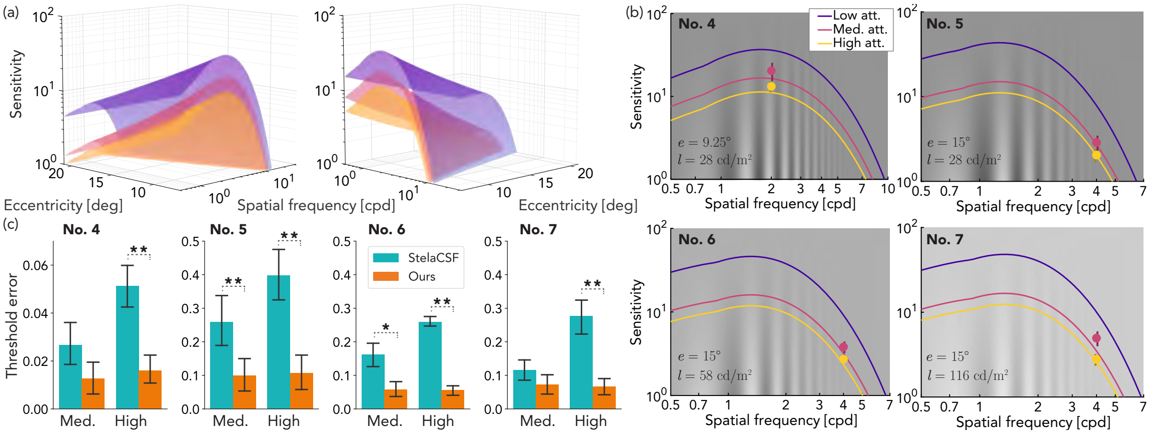

Figure 4.5: Validation studies: (a) Two different views of an eccentricity vs. spatial frequency plot for the original StelaCSF [56] model (the top surface, in purple) and our scaled models $S_{\mathrm{medium}}(e, f_s)$ (in magenta), $S_{\mathrm{high}}(e, f_s)$ (in yellow) for a static stimulus with an area of $1\,\mathrm{deg}^2$ and an adaptation luminance of $28\,\mathrm{cd/m}^2$ (same as our model study in Section 4.1.3). (b) Slices of the same models describing dependency on spatial frequency for the conditions used in our validation study (see No. 4–7 in Table 4.1). The points denote directly measured sensitivities scaled relative to the baseline. The bars are 95% confidence intervals. (c) Corresponding threshold prediction errors of StelaCSF vs. our model (lower is better). The error bars are 95% confidence intervals and significance is indicated at the $p < 0.05$ and 0.01 levels with * and ** respectively (Wilcoxon test).

For the 2 stimuli isoluminant with our model data (No. 4 and 5), we observe statistically significantly lower error between our and the original StelaCSF predictions with respect to the experimentally measured thresholds in all conditions except for one. The lower observed difference between "low" and "medium" attention for the lower frequency stimulus (No. 4) points to an overestimation of the gain by our model here. Notably, even in this worst case, the prediction error is still lower than that of the baseline StelaCSF.

As a practical example, our measurements indicate that with "high" attention and a $28\,\mathrm{cd/m}^2$ display, a spatial pattern with $f_s = 2\,\mathrm{cpd}$ shown at eccentricity of 15° will be just discriminable if rendered with an amplitude value of 32 (for a 0–255 signal range of an 8-bit display with gamma of 2.2) while our model would yield amplitude of 30 and the baseline model would adhere to the original stelaCSF prediction of amplitude 8.

The favorable performance of our model also holds for the other 2 luminance levels (stimuli No. 6

and 7). Despite this, we observe a trend of attention gain reduction with increasing luminance which is significant for the "medium" attention at 116 cd/m^2 ($g_a : 3.03 \rightarrow 2.15$, $p < 0.05$, Mann-Whitney U test) and "high" attention at 58 cd/m^2 ($g_a : 4.12 \rightarrow 3.37$). This compression could be caused by the overall increase of sensitivity under such conditions and should be considered by users of our model.

To summarize, our experiment suggests that while the assumption of full orthogonality is unlikely to hold everywhere, the relative benefit of including the attention model may still be stronger relative to the cost of this simplification.

4.2 Attention-aware Foveated Rendering

The goal of foveated rendering is to reduce computational cost without introducing perceptible artifacts by exploiting the reduction of vision performance in the periphery, typically by adjusting sampling rate with respect to peripheral acuity drop (see Chapter 2, Section 2.2.2 for a review of these techniques). The quality of such foveation can be assessed by visual difference predictors, for example, FovVideoVDP [73], a state-of-the-art metric that models the spatial, temporal, and peripheral aspects of perception. In this section, we experimentally validate whether integration of our attention-aware perceptual model improves the performance of FovVideoVDP in predicting visibility of foveation artifacts under varying attention conditions. To that end, we emulate a simple foveated renderer and we separately calibrate the foveation intensity for three different attention regimes in a user study. We then compare the perceptual errors of the calibrated stimuli predicted by FovVideoVDP with and without our attention-aware model to assess their agreement with human judgment.

4.2.1 Measuring Imperceptible Foveation

Similar to the contrast discrimination task in Section 4.1, we create a space-multiplexed comparison of foveated images. In particular, we split the screen into left and right sides, apply foveation to one side only (randomly selected) and ask subjects which of the sides appeared more visually degraded (see Figure 4.6a for an illustration). Then, by modifying the parameters of the foveated side, we can find the threshold for which the foveation is nearly imperceptible to the subject. The central transition around the fixation was replaced by a neutral background vertical bar with 6° width and Gaussian fall-off (standard deviation of 0.5°) and the attention-modulating RSVP task used in the previous studies, was then displayed centrally.

For the foveation, we base our approach on the work of Guenter et al. [12] (described in Chapter 2, Section 2.2.2) and the linear MAR model (described in Section 2.1.4) describing the reciprocal of acuity as:

$$\omega(e) = me + \omega_0 \tag{4.5}$$

where the bias $\omega_0 = 1/48°$, as in the original work, and the slope m is a free variable measured as a threshold in our study. Rather than choosing three distinct layers, the peripheral resolution decrease was simulated as a continuous function of eccentricity by an approximated Gaussian filter with spatially varying standard deviation:

$$\sigma(e) = \frac{\omega/\omega_s - 1}{2\sigma_c} \tag{4.6}$$

where $\omega_s = 0.0283°$ is the peak MAR of our display and $\sigma_c = 2$ is the chosen cut-off determining the assumed bandwidth of the filter. Note that this particular choice primarily affects the absolute value of our slopes and not the relative ratios between conditions.

4.2.2 User Study

Setup We used the same experimental setup as in Section 4.1. The stimuli consisted of one of four foveated images displayed across the entire screen with the gaze fixation directed to the center.

Subjects Thirty adults participated (age range 23–29, 10 female). Subjects were first shown the original images, to avoid exploratory saccades during the study, and then shown an example of the foveation effect.

Procedure As in the studies in Section 4.1, subjects were instructed to always fixate on the central RSVP task and observe the foveation task concurrently in their periphery. The thresholds for each image were estimated in a random order for each subject. For each image, a 2-AFC adaptive staircase using QUEST [187] was used to measure the threshold of the foveation slope m for each attention condition, starting with the "low" , followed by the "medium" and ending with the "high" foveal attention condition. Similar to the previous studies, at each step, the subject is shown the attention-modulating task and foveation detection task for 500 ms. The subject then indicated the color of the target letter "T", and if correct, were asked which side of the image (left or right) was more visually degraded. Whenever the subject incorrectly answered the RSVP task, they were forced to start the step again. Each subject viewed 2 images, either "Tulips" and "City" or "Mountain" and "Forest" (see Figure 4.6a) at each of the 3 attention conditions, to keep the study duration to approximately 45 minutes (including breaks).

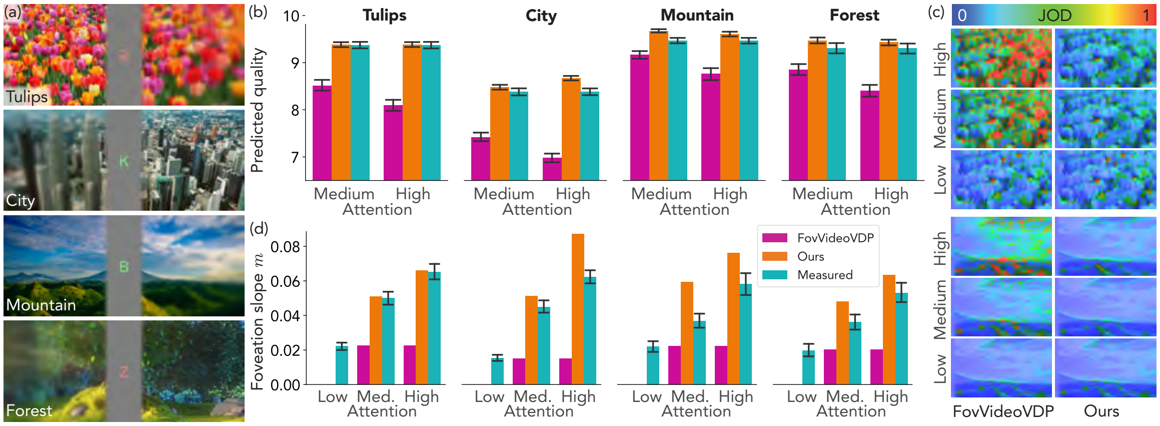

Figure 4.6: Foveation study: (a) Stimuli from our study showing the attention-modulating RSVP task in fovea and the peripheral foveation detection task. One side (randomly selected) is foveated while the other is left at full resolution. The foveation effect and the color and size of the RSVP task are exaggerated for visibility. (b) Quality scores predicted by the original FovVideoVDP metric vs. our modified predictor (closer to *Measured* is better) for the foveated images calibrated in our user study. The error bars show 95% confidence intervals. The *Measured* quality refers to the actual quality threshold measured in the "low" condition. (c) Comparison of visual difference maps produced by the original FovVideoVDP metric vs. our modified predictor for the calibrated MAR slopes. Colors visualize Just-Objectionable-Differences (JOD) with respect to the original "Tulips" and "Mountain" images (small section from the right periphery shown). (d) Comparison of MAR slopes (intensities) predicted by the original FovVideoVDP metric vs. our modified predictor for each image compared to the *Measured* slopes (closer to *Measured* is better). The legend is shared with panel (b). The error bars show 95% confidence intervals of the measured values. Note that the model-based slope predictors do not yield variance (no error bars shown). (Tulips image by Alex Ohan from Pexels: https://www.pexels.com/photo/11644862/, City image by Pok Rie: https://www.pexels.com/photo/4847105/, Mountain image by Archie Binamira from Pexels: https://www.pexels.com/photo/672451/ and Forest by Blender Foundation from: https://peach.blender.org/about/)

Results In Figure 4.6d, we show the measured MAR slopes m averaged across the users (labeled "*Measured*"). We applied the same IQR procedure as in Section 4.1.6 but we did not detect any outliers. As expected, for all images the slope significantly increases ($p < 0.001$, paired t-test with Bonferroni correction) for both "medium" and "high" compared to "low" attention conditions. This means that a more aggressive foveation becomes acceptable as attention shifts from periphery towards the fovea. Furthermore, we note that there are statistically significant differences between slopes measured for at least some image pairs with "low" (one-way ANOVA, $F(3, 116) = 4.99$, $p = 0.003$), "medium" ($F(3, 116) = 10.26$, $p < 0.001$) and "high" ($F(3, 116) = 3.87$, $p = 0.011$) attention. This points to a content-dependent nature of the problem. In the next section, we discuss whether a visual difference predictor could be used to predict foveation parameters for a specific image.

4.2.3 Foveated Quality Prediction

Setup Acuity-driven foveation algorithms conservatively account for the worst-case scenario of the smallest detectable image detail [132]. As seen in our results, distribution of contrast in specific images affects the acceptable foveation intensity. This is modeled by visual difference predictors such as FovVideoVDP [73], which decomposes an image into spatio-temporal frequency bands and models their visibility by utilizing the CSF and a contrast masking model. However, while explicitly modeling retinal eccentricity, the original CSF does not account for attention. We experimentally modify the authors' implementation and integrate our model as an orthogonal scaling factor of the CSF component. We then apply the original and the modified predictor to assess the quality of the foveated images with the per-image calibrated slopes from Section 4.2.2. Furthermore, we evaluate whether an inverse process could be used to optimize the foveation intensity.

Quality metric In Figure 4.6b, we display quality scores produced by the original FovVideoVDP metric compared to our modified predictor obtained by computing visual difference between a foveated image calibrated by each individual subject and the full-quality reference. The scores were averaged for each image, for each of the "medium" and "high" attention conditions. We compare these to the expected value (labeled "*Measured*") which was obtained by FovVideoVDP for foveated images calibrated with the "low" attention condition. We assume that this represents the personal subject-specific threshold of the perceived quality for the given image and that it should remain constant under varying attention.

Following our previous results, we expect that images with larger objective degradation should be judged as having equivalent quality and that a successful prediction should reflect that. Consequently, we observe that the error measured as a relative difference between our prediction and the "*Measured*" value is consistently lower than that for FovVideoVDP ($p < 0.001$, Wilcoxon test). This suggests that our modified predictor is better aligned with the attention-modulated perception.

In Figure 4.6c, we additionally compare maps of Just-Objectionable-Differences (JOD) produced by both predictors for the mean calibrated slopes at each condition. FovVideoVDP indicates a strong increase of perceived artifacts even as attention towards the fovea ("high" attention condition). Our method instead predicts errors on the boundary of visibility for all conditions which is consistent with our assumption.

Finally, we note that the "low" attention score predicted by FovVideoVDP based on the directly measured slopes is significantly different between at least some of the images (9.375 for "Tulips", 8.383 for "City", 9.470 for "Mountain" and 9.314 for "Forest", $F(3, 116) = 150.5$, $p < 0.001$, one-way ANOVA). Since this is the baseline condition, this discrepancy is orthogonal to our primary objective of exploring the overall impact of attention, and thus we defer its investigation to future work.

MAR slope prediction The visual difference prediction potentially allows us to optimize foveation parameters by posing it as a constrained problem:

$$\Theta = \arg\min_{\Theta} C(\Theta) \quad \text{subject to} \quad Q(\Theta) \geq Q_{\text{thr}} \tag{4.7}$$

where Θ is a set of rendering parameters, $C(\Theta)$ is the cost of the rendering (typically time and power consumption), $Q(\Theta)$ is an image quality predictor and Q_{thr} the required threshold. In our case $\Theta = m$, $C(\Theta)$ is a monotonically decreasing function of m, $Q(\Theta)$ is provided by our visual difference predictor (with access to the reference image) and Q_{thr} is obtained from the "low" foveal attention condition in Section 4.2.2. The resulting problem of one variable can be efficiently solved by bisection.

Ideally, we could use a single Q_{thr} for any image. However, due to the significant difference between "low" attention thresholds obtained for our images, we opt to use scene specific Q_{thr} of 9.375 for "Tulips", 8.383 for "City", 9.470 for "Mountain" and 9.314 for "Forest". This simulates a correction function that calibrates the underlying predictor for content-dependent effects and development of which is outside of the scope of this work.

Figure 4.6d compares the MAR slopes obtained by solving the inverse problem with $Q(\Theta)$ implemented using the original FovVideoVDP metric and our modified predictor. While our predicted slopes do not always match the directly measured values, for the "high" attention the errors are consistently lower than those from FovVideoVDP ($p < 0.05$ for the "Mountain", $p < 0.01$ for the rest, non-parametric Wilcoxon test). This is remarkable given the large domain gap between the model and foveation stimuli. It suggests that our model is useful for attention-aware foveated rendering. Similarly, we observe statistically lower prediction errors of our model with the "medium" attention for the "Tulips" and "City" images ($p < 0.01$) while no statistically significant differences were measured for the rest.

The remaining error could originate from a multitude of sources, among them the orthogonal assumption of CSF scaling as well as other higher level effects not accounted for in either model. To illustrate the impact, in the worst case of the "City" image with "high" attention and the extreme periphery of 46° in our experimental display setup, this error would lead to removal of spatial details in the 0.25–0.35 cpd band which might be noticeable based on our measured data.

Importantly, we observe that the bias towards overestimation of the slopes is consistent. This hints to feasibility of fine tuning for a specific foveation algorithm. Even without such treatment, the relative preference of our model over the baseline is most prominent for the "high" attention which is particularly relevant for many applications where users focus at a specific target on the screen.

4.2.4 Analyzing Bandwidth Considerations

In this section we analyze the additional computation gain that could theoretically be obtained by using our attention-aware model when the user is focused on a task in the fovea. While we could analyze the bandwidth by decomposing the image into frequency bands and discarding the signal following the CSF predictions, we decided on a more conservative approach that instead uses our direct perceptual measurements of vision performance under the specific foveal (RSVP) task. Therefore, we model the foveation algorithm by Guenter et al. [12] together with the global mean MAR slopes m obtained for the "low" , "medium" and "high" attention conditions as 0.0198, 0.0420 and 0.0596.

Unlike the discrete segmentation in the original algorithm, we simplify the analysis by assuming that sampling rate of each pixel can be controlled independently and hence we directly map the local MAR, $\omega(e)$ (Equation 4.5), to the computational gain Ψ derived from local areal sampling density as:

$$\Psi(\mathrm{FOV}) = \left(\int_{\mathrm{FOV}} 1 \,\mathrm{d}x\right) \cdot \left(\int_{\mathrm{FOV}} \max\left(\frac{\omega(x)}{\omega_s}, 1\right)^{-2} \mathrm{d}x\right)^{-1} \tag{4.8}$$

where ω_s is the peak MAR of a given display with a 2D field of view FOV.

In Figure 4.7, we display computational gains obtained as a function of display FOV for a common 20 ppd and future high-density 60 ppd displays as an upper bound for the analyzed algorithm. It must be noted that gains in real applications are influenced by efficiency of a particular renderer. As our contributions are independent of such design choices, more advanced foveation approaches such as noise-based enhancement [132] can be considered for additional gains.

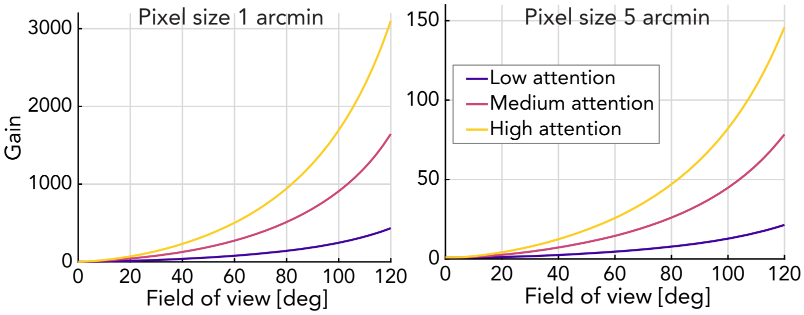

Figure 4.7: Computational gain analysis as a fraction of original and retained pixel sampling density depending on pixel size and the covered visual field (horizontal axis). Note the difference in the gain axes scales.

4.3 Discussion

The experimental data we measure and the models we fit to them further our understanding of human perception and lay the foundation of future attention-aware foveated graphics techniques. Yet, several important questions remain to be discussed.

Limitations and Future Work While our studies clearly demonstrate that modulating attention distribution between the periphery and fovea strongly impacts contrast sensitivity and foveation efficacy, we do not propose a method to measure attention. In contrast to overt attention, which is readily measurable using eye tracking, covert attention is much more challenging. A promising direction for exploration is the relation between pupil dilation and attentional effort [203] and in some scenarios pupillary light response [204]. One might also investigate the combination of eye tracking with image salience [205] or other metrics. It should also be noted that training and practice can significantly improve the ability to split attention between the fovea and periphery [206]. This could lead to decrease in attention dedicated to the foveal RSVP task. To mitigate this, we randomize trial order and encourage sufficient rest time, yet such an approach is time consuming and limits the CSF gamut that can be measured in one sitting. While we show that our orthogonal scaling approach still leads to favorable performance when compared to baselines, we emphasize that our attention model should not be extrapolated outside of the measured eccentricities. Finally, rather than proposing a novel foveation algorithm, we focus on demonstration of the perceptual effect as a whole. Future work should investigate more advanced foveation algorithms and explore the effect of attention in the temporal domain.

4.4 Summary

At the convergence of applied vision science, computer graphics, and wearable computing system design, foveated graphics techniques will play an increasingly important role in future VR and AR systems. With our work, we hope to motivate the importance of cognitive science in human perception and inspire a new axis of approaches within foveated graphics.

Chapter 5

Low-latency Foveated Display

Foveated graphics shows great potential for reducing the data needed to achieve perceptually realistic experiences in VR and AR. This suite of techniques rely on accurate, real-time gaze information to center the high quality content on the fovea, while decaying quality in the periphery. In practice, however, these systems have update (motion-to-photon or end-to-end) latencies, i.e. the time delay between the user's eyes moving and the pixels of the display updating. For current-generation commercial VR and AR displays that include native eye tracking, this is 45-81 ms in the best case scenario [119], and more with complicated graphics pipelines for foveated rendering or other display manipulation techniques [133]. This delay can cause discomfort (i.e., simulation sickness, fatigue) [207], but also introduces uncertainty in a viewer's gaze position[1]. This is illustrated in Figure 5.1. If the system can't update the foveal region fast enough, then the degraded peripheral region will be incident on the fovea and the user will observe visual artifacts.

Despite this, the effects of latency on these displays are not well characterized. In foveated rendering, a number of prior works have measured the maximal tolerable system latency to be between 42 ms to 91 ms, depending on the size of the full resolution foveal image that follows the gaze, the degree of degradation applied to the image, and the type of degradation method used [12, 43, 209, 210]. Similarly, Loschky et al. [211] also observed that detection of image artifacts due to foveation in gaze-contingent, multiresolution displays did not change if latency was kept under 60 ms. However, to the best of our knowledge, the impact of latency on the potential bandwidth savings afforded by these techniques has not been described in the literature.

Display systems with larger latencies introduce larger uncertainty about the viewer's gaze position. So when saccades can occur at speeds of up to $\sim 900°/s$ [24] (see Section 2.1.2), these systems will require that foveated graphics techniques have a larger foveal region to avoid the perception of the degradation applied in the periphery – when the system cannot update its position quickly

[1] Inaccuracy in an eye tracking device also contributes to this uncertainty but is out of scope of this work.

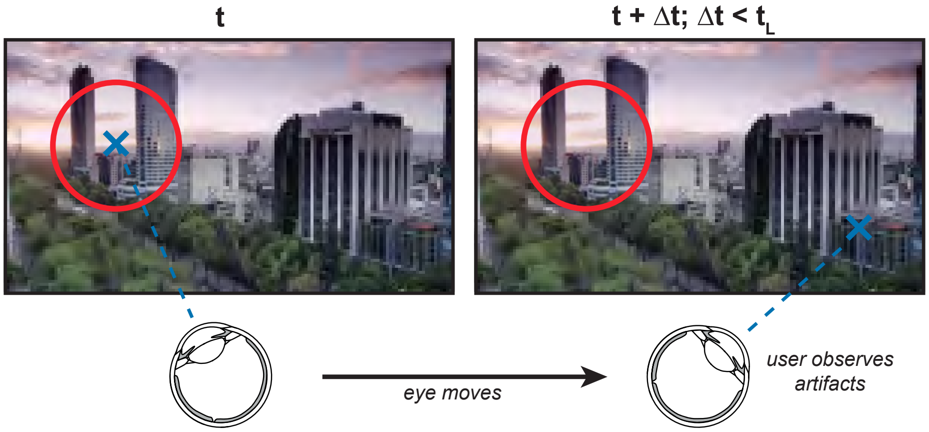

Figure 5.1: Illustration of how system latency can cause visual artifacts. (left) The gaze position used by the system matches the actual gaze position, and the periphery can be highly compressed. (right) However, if the user's eye moves fast enough to escape the foveal region in a shorter time, Δt, than the system latency, t_L, then the degraded peripheral region will be incident on the fovea and the user will observe visual artifacts. (City image source: Derf's collection [208])

enough. Consequently, there is tension between minimally sizing the foveal region for better compression and sizing it large enough to ensure a viewer does not see visual artifacts.

In this work, we propose latency reduction as a method for improving the bandwidth saving potential of foveated graphics. To this achieve goal, we design and build a custom, low-latency foveated compression display system, and carefully characterize the breakdown of the sub-15 ms latency across the end-to-end system. Finally, we use this system to conduct a user study, demonstrating that up to 2× extra bandwidth savings are possible just by reducing the latency of the gaze-contingent display system. As such, we motivate the importance of latency as not just a systems engineering problem, but as an axis crucial to the success of foveated graphics.

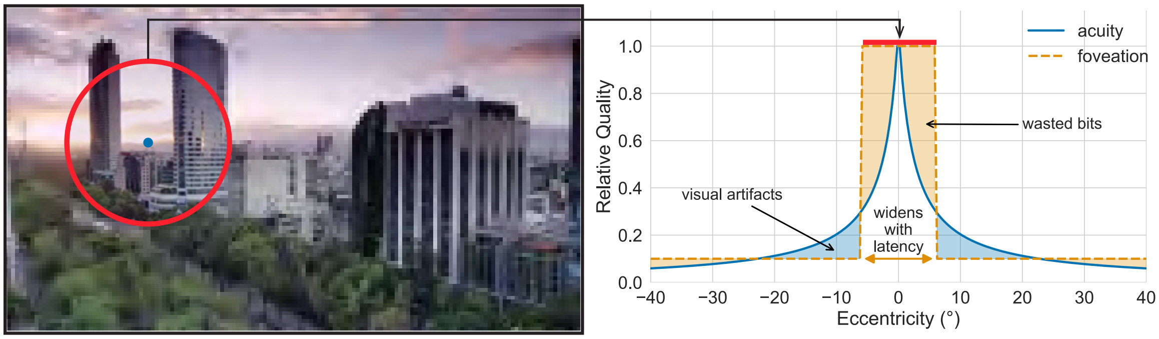

Figure 5.2: One reason VR display systems today cannot yet deliver retina-quality visual experiences is due to bandwidth limitations. To reduce data rates, foveated compression techniques exploit the decay of visual acuity across the visual field, keeping a small region of high resolution while decaying quality in the periphery (illustrated on the left). We show that decreasing system latency benefits foveated compression and enables minimally-sized regions of high resolution (illustrated on the right).

Specifically, in this chapter we: (1) build and characterize a low-latency video streaming system using foveated video compression that reacts to gaze changes within 15 ms, over 3× lower than previously demonstrated in VR display systems, and (2) design and conduct a user study that shows that just by reducing the system latency, can as much as double the bandwidth savings of foveated compression without noticeable quality degradation.

5.1 Low-latency System Prototype

Understanding the impact of latency on foveated video compression requires a system with very low latencies. However, current-generation commercial VR and AR displays systems have system latencies over 45 ms [119]. In addition, these devices do not have sufficiently high resolutions (i.e., less than 4K) to be an ideal test bed for studying the impact of latency on compression of retina-quality video[2]. Consequently, we built a custom ulta low-latency foveated display system; a desktop-based system as a proxy for future VR display systems. Doing so allows us to focus on the impact of latency without the limitations of current devices.

5.1.1 Architecture

We designed our system based on a typical video-streaming architecture with a client and server model. However, rather than the client only receiving encoded video frames from the server to decode and display, the client also sends the viewer's current gaze position each time a frame is received (see Figure 5.3). This gaze sample allows the server to encode the next video frame foveated on the viewer's gaze position. To minimize system latency, the server and client run as separate processes on the same machine and communicate using message passing, implemented with shared memory.

5.1.2 Two-Stream Approach

We model the decay in visual acuity across eccentricity (see Chapter 2, Section 2.1.4) with a simple step function, implemented using two traditionally compressed video streams – one for the cropped high-resolution foveal region and one for the low resolution background. For example, rather than processing a full 4K (3840×2160 px) video, a two-stream approach might process a small 480×480 px foveal region and a downscaled 768 × 432 px background, which combines to be less than 7 % of the original 4K pixels. While this approach results in a course approximation of spatial sensitivity, it results in much less processing than other methods e.g. Gaussian fall-off [212], and focuses on the impact of reducing the latency spent on encoding and decoding.

The server sequentially reads uncompressed frames at the frame rate of the input video. Then, for each gaze sample it receives from the client, it compresses up to two versions of the current

[2] At the time of publication. Recent VR displays, such as the Vive Pro 2, do include 4k or higher resolution displays.

frame[3]. First, it downscales the video frame to a significantly lower resolution. Second, it crops the video frame to a small area around the viewer's gaze location. The resolution of both the downscale and the crop are configurable. It then encodes these two frames to send to the client. At the client, the reverse process occurs. First, the client decodes and upscales the background frame to the size of its display. Next, it decodes the foreground frame and positions it at the corresponding gaze position with a blend. In particular, we set the alpha channel (opacity) to a 2D Gaussian in order to fade out the hard, square edges of the foreground. The parameters of the Gaussian are chosen empirically. Finally, it displays this composed frame.

As is typical with compression techniques, this approach trades off increased computation (real-time encoding per client) for reduced bitrate. While the server can pre-encode the background, the foreground must be encoded in real-time using the viewer's gaze.

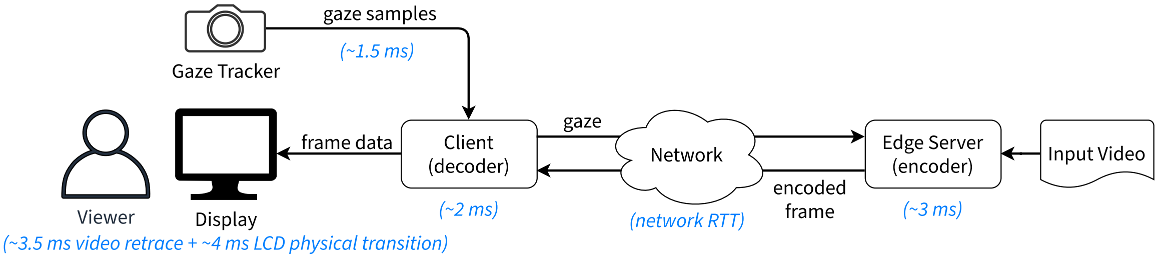

Figure 5.3: Overview of our low-latency, desktop-based foveated compression prototype system. This system allows us to focus on the effects of latency on foveated video compression by avoiding the limitations complexities of current VR display systems.

5.1.3 System Details

We implemented our system in Rust, using SDL2, FFmpeg, and x264. Our workstation runs Pop!_OS 20.04 and contains an AMD Ryzen 7 3700X CPU, 16 GB of 3200 MHz memory, and an NVIDIA GeForce RTX 2070 SUPER GPU. Our display is an LG 27GN95B-B (4K at 144 Hz, 7.6 ms input latency). We use an Eyelink 1000 to minimize the latency between a viewer's eyes moving and receiving the data in software. Although lower-latency eye trackers are continually being developed [213], the Eyelink 1000 provides the best accuracy, latency[4], and sampling rate among those that are commercially available (see Chapter 2, Section 2.2.1 for specifications).

[3] We also skip both background frames when the current video frame has not changed and foreground frames if the gaze has not not changed.

[4] We disable the built-in filters to further minimize latency

5.1.4 Measuring Latency

To ensure precision and repeatability, we carefully designed an automated system for measuring the end-to-end system latency of our foveated compression prototype (illustrated in Figure 5.4). This system is based on the previously used approach of using an artificial saccade generator (ASG) and photodiode circuit [214–216]. An ASG emulates the user moving their eyes to trigger the eye tracker, removing the need to rely on a human. Unable to find a suitable commercial ASG, we designed our own based on an Arduino circuit. Since Eyelink 1000 uses the relative position of the pupil and corneal reflection (CR) to track gaze position (see Section 2.2.1), we mounted a printed pattern of an eye with 2 infrared LEDs placed behind pinholes in the pupil. The LEDs are connected in a circuit such that a TTL (transistor-transistor logic) signal switches which one is on, where the other is off. The eye tracker then acquires the pupil and detects the LED that is powered on as a CR.

To measure end-to-end system latency, our automated setup conducts the following steps:

1. The control computer sends a TTL pulse which triggers the ACG.

2. The ACG switches between the on/off LEDs, moving the CR position, emulating an instantaneous saccade.

3. The eye tracker detects this new gaze position and sends it to our decoder.

4. Our system uses this gaze position to process (including scaling, cropping, encoding and decoding) and display the next video frame, including a small white portion.

5. A photodiode is attached to the surface of the display, over the position of the white portion, such this display update (from black) increases the luminance of the area underneath and triggers the photodiode circuit.

6. The output from the photodiode is then used by the control computer to calculate the latency between the onset of the saccade and the physical display change (specifically, the mid-point of the rise slope).

In some cases, it is possible for the saccade to be triggered and the display pixels to change within the one refresh cycle of the monitor. However, if the saccade does not line up with the frame clock, then it may take up to an additional refresh cycle to update. To take this into account, we run this measurement for 300 repetitions and plot the empirical cumulative distribution function (ECDF) in Figure 5.4. The minimum observed latency is under 11 ms, with the majority of samples falling under 16 ms, with the average being ∼14 ms. An approximate breakdown of where time is spent is annotated in Figure 5.3.

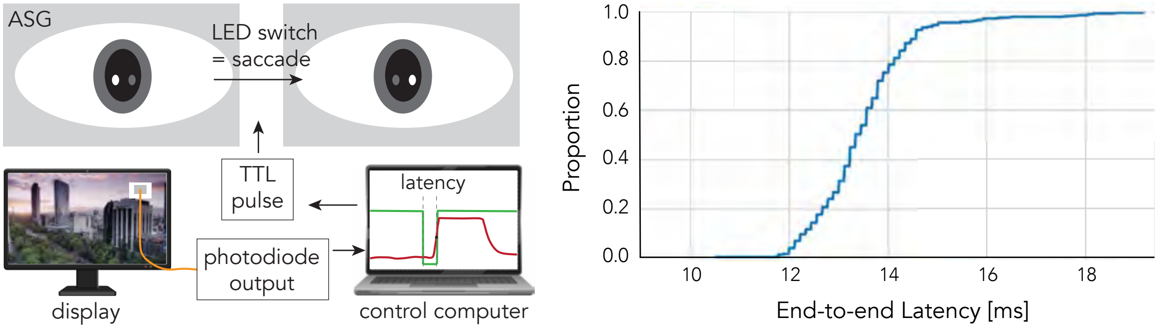

Figure 5.4: Our automated system for measuring end-to-end system latency. (left) The control computer sends a TTL pulse which triggers the ACG, emulating a saccade. The eye tracker detects the change and triggers our system to update the frame. The display change triggers a photodiode attached to the surface of the display which is sent back to the control computer. The system latency is then calculated as the time difference between the onset of the TTL and the mid-point of the photodiode rise slope. (right) ECDF for 300 measurements of end-to-end system latencies, averaging ∼14 ms.

5.2 Characterizing Latency vs. Bitrate

We conduct a user study to characterize the relationship between system latency and how much compression can be achieved while maintaining similar visual quality. In particular, we use our low-latency prototype (described in Section 5.1) to measure how reducing latency may affect tolerable compression bitrates.

5.2.1 Stimuli

Two 4K videos from Derf's collection [208] were selected based on having different content but both consisting of two sub-scenes. The first, `barscene`, shows one sub-scene with strong bokeh and another with dialogue between two individuals that naturally guides a viewer's gaze. The second, `square_timelapse`, shows one sub-scene of a busy crowd of people where viewers' gaze typically jumps around the scene, and another of a city skyline with many hard edges and natural scenery.

We evaluate these two videos at three system latency conditions, to keep the study to a reasonable duration. First, we evaluate our system at its unmodified latency (∼14 ms). Then, we add artificial delay to match the minimum and maximum latencies of commercially available display systems (as measured by Stein et al. [119]: 45 ms and 81 ms). For each video and latency combination, we prepare a set of compression configurations starting at lower resolutions with higher compression, and moving to higher resolutions with lower compression. We also encode the video at 28 Mbit s^{-1} as a proxy for the video quality of streaming platforms like YouTube[5]. In this way, we conduct a user study to assess which compression level is of most similar visual quality to the 28 Mbit s^{-1} "YouTube" version, for a given system latency.

[5] Specifically, we use `x264 --preset veryfast --bitrate 28000`.

5.2.2 User Study

Setup We use the system detailed in Section 5.1.3. As shown in Figure 5.5, the participant's head is stabilized using an SR Research headrest, and the eye tracker is placed between the monitor and participant. The display is placed 57 cm away from the headrest, giving a horizontal FOV of 55° and above retina resolution.

Procedure To start the session, the eye tracker was calibrated and the tracking accuracy validated for each participant. Then, participants viewed two different videos and performed the same task on each. The order in which the videos were presented was equally divided among participants. For each video, participants were asked to perform four trials of a matching task. Three of the trials correspond to each latency points (14 ms, 45 ms and 81 ms), and the last was one of these points randomly selected, to provide a check for consistency. The order of the trials was also randomized. For each trial, participants were shown a reference video (the 28 $\mathrm{Mbit\,s^{-1}}$ or "YouTube" encoding) and then asked to select which (if any) of ten comparison videos (ordered by increasing quality) was the most similar in quality. Participants were allowed to take as long as they wished to make their selection and could freely navigate and re-watch any of the videos. Each participant did 8 trials, resulting in a study duration of ∼45 min.

Participants We recruited 13 participants[6]. All participants provided written consent before taking part in the study, and the methods were approved by Stanford's institutional review board. Before each experiment, the participants were briefed about the purpose of the study and their task. Of these 13 participants, we excluded 2 participants' data from the results because we were unable to achieve a maximum calibration accuracy error $<10°$ (unacceptably large compared to the size of the foveal region).

Results The mean compressed bitrate as a percentage of the 28 $\mathrm{Mbit\,s^{-1}}$ baseline along with the 95 %-confidence interval for each latency are shown in Figure 5.5. Firstly, we can appreciate the power of gaze-contingent rendering and display paradigms. Even with a simple two-stream approach at system latencies comparable to commercial VR devices [119], can compress these videos to $\sim 20\%$ of the baseline while maintaining a similar visual quality. We can also see that reducing that system latency from 81 ms to 45 ms does not significantly improve the required video bitrate (t-test, $t = 0.10$, $p = 0.92$). It is not until we push the system's latency to below that of commercially available display systems that we see a significant, additional ∼2 × compression benefit (t-test, $t = 2.76$, $p = 0.008$). This finding also suggests this relationship is not simply a question of making the foveal region larger as the delay increases – we suspect there is a distinct phenomenon (and a compression opportunity) at low latencies.

[6] The COVID-19 pandemic limited the number of participants available for this study.

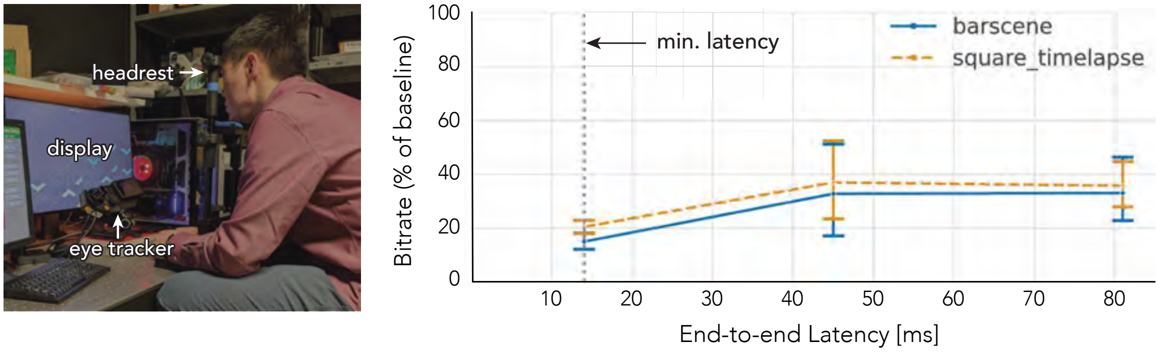

Figure 5.5: Latency vs. compression, with 95 %-confidence intervals. We improve compression by 5 × using a simple two-stream approach. However, the full benefit comes only at latencies lower than demonstrated by current VR HMDs.

5.3 Discussion

Despite being key to their success, the system latency of gaze-contingent displays is rarely discussed. With this work, we present latency reduction as a method for improving foveated video compression and validate its potential by implementing a prototype, ultra-low-latency video streaming system. Our findings suggest not only that driving down system latency can result in significant compression gains without changing the compression algorithm itself, but also that these gains might only be realized with system latencies much lower than previously proposed budgets. Yet, several important questions remain to be discussed.

Limitations and Future Work This work focuses on evaluating the impact of motion-to-photon latency on the compression ratio of gaze-contingent compression. As a result, our low-latency prototype has a few important limitations. First, our system excludes the latency introduced by separating the client and server with a realistic network. The need for low server-to-client latencies means that a video encoder would need to be near the client at the network edge (a potential use case for edge computing). Our prototype also uses an encoder-in-the-loop approach to perform video compression and streaming in real-time. This approach has a higher computational cost than those that pre-encode video since it instead requires the server to encode video for each viewer. Additionally, we evaluate our system using an eye tracker and display that are among the fastest available today; comparable performance is unavailable on the consumer market or in current head-mounted displays.

With regards to our user study, it would be useful to know exactly what latency target is required to realize a significant reduction in transmitted bitrate through gaze-contingent encoding; that is to say: if reducing latency from current levels (∼ 80 ms) to ∼ 45 ms is not helpful, but reducing to 15 ms is very helpful, then what does the curve look like between 15 and 45ms? Because of the limitations of our study (which was conducted during the COVID-19 pandemic), we cannot answer these questions today, but encourage the community to further investigate the trade-offs between lowering latency in gaze-contingent video transmission and resulting bitrate reductions.

5.4 Summary

Foveated graphics shows great potential for reducing the data needed to achieve perceptually realistic experiences in VR and AR. With this work, we hope to highlight the importance of latency as not just a systems engineering problem, and spur follow-on work investigating other implications for VR/AR.

Chapter 6

Gaze-contingent Stereo Rendering

Accurate stereoscopic (or just *stereo*) rendering is one of the key requirements for perceptual realism. As described in Chapter 2 (Section 2.1.8), humans are very sensitive to changes in disparity and so even small inaccuracies in the rendering would lead to objects being perceived at a different depth than intended. Especially in the case of optical see-through AR displays, where we want to place and anchor digital objects physical environments, even small amounts of disparity distortion negatively affect the experience and would destroy the seamless blending of virtual and real content.

Current stereo rendering algorithms used in VR and AR fall short of accurately modeling the HVS. As described in Section 2.1.8, the geometry of the system, including the user's IPD is used to calculate and display the retinal disparity that would occur during real-world viewing. An approximation made by almost all existing systems is that the no-parallax point, or center of projection, of the human eye is co-located with the center of rotation. However, recent work suggests that this is not in fact the case, and that the no-parallax point is instead slightly forward offset from the center of rotation [18]. As illustrated in Figure 6.1, when the user verges far, this finding has negligible effect, as the distance between the pupils (or IPD) is also the distance between the no-parallax points. But as the user verges closer, the true IPD is actually gaze-contingent and the geometry of the disparity calculation should change depending on where you are looking.

In this chapter, we analyze how not accounting for this ocular motion causes disparity distortion in stereo viewing conditions (see Figure 6.2 for an example), and propose and evaluate a new gaze-contingent stereo rendering approach for VR and AR. Moreover we design and conduct a number of user experiments that allow us to make important insights and improvements to existing stereo rendering techniques. First, we experimentally determine the location of the no-parallax point with a small group of subjects and verify that this is well approximated by recently employed model eyes. Second, we experimentally demonstrate that our approach significantly improves disparity distortion in VR settings and perceived alignment of digital and physical objects in AR.

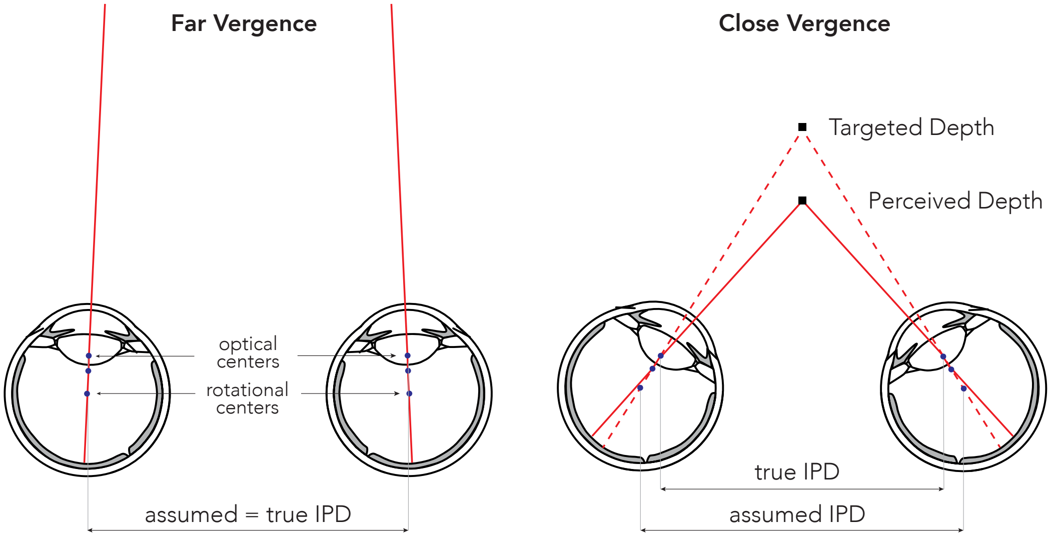

Figure 6.1: Schematic showing how true IPD changes with gaze position because the no-parallax point, or optical center of the eye, is slightly forward offset from the center of rotation. When a user verges far, this finding has no effect, as the distance between the pupils (IPD) is also the distance between the no-parallax points. But as the user verges closer, the points shift inwards, meaning that true IPD is actually gaze-contingent and the geometry of the disparity calculation should change depending on where you are looking.

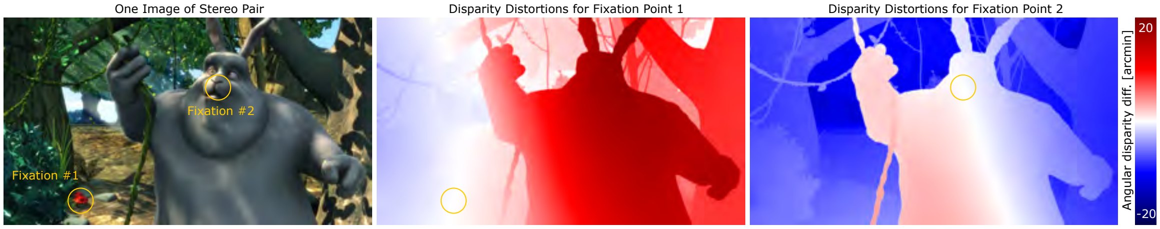

Figure 6.2: Ocular motion associated with changes of fixation alters the positions of the no-parallax points in both eyes. Rendering models that do not account for this motion can create distortions of binocular disparity, as seen in this example. The color-coded error maps illustrate the magnitude of this effect as the difference between angular disparities resulting from conventional and our gaze-contingent stereo rendering for two different fixation points. Both shortening (red) and stretching (blue) of disparity gradients can be observed.

Specifically, in this chapter we: (1) introduce a gaze-contingent stereo rendering algorithm that includes a more accurate model eye than previous work, design and conduct user experiments that demonstrate significant improvements in (2) disparity distortion and enhanced depth perception in VR, and (3) perceptual realism and alignment of digital and physical objects with optical see-through AR.

6.1 The No-parallax Point of the Eye

6.1.1 Background

The no-parallax point of an optical system, such as the human eye, represents the location around which the entire system can be rotated without observing parallax. While exact eye anatomy varies from person to person, several cardinal points and axes are used to describe optical properties common across the population. For the purpose of describing our gaze-contingent rendering approach, we refer to six of these cardinal points and two axes, however additional definitions are included in Appendix C.1. The first axis is referred to as the optical axis, and is anatomically defined by passing through the anterior vertex of the cornea (V) and the center of rotation (C) whereas the visual axis registers the fixated object (P) with the fovea (F) while passing through front (N) and rear (N') nodal points [217] (illustrated in Figure 6.3). From *ex vivo* examinations, the front nodal point has been estimated to lie 7–8 mm in front of the center of rotation [218]. Note that the location of this point changes with the accommodation state but such variance is relatively minor (less than 0.5 mm [218]), which is particularly true for current fixed-focus VR and AR systems. While it has been postulated that the front nodal point is in fact the no-parallax point of the eye, to our knowledge the only study trying to verify this could not give a confirmation as the larger measured distance hinted at the no-parallax point being located even further forward [219].

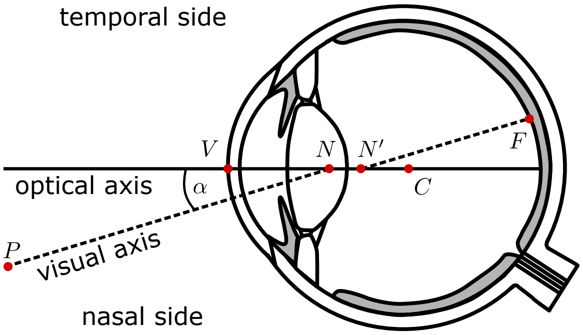

Figure 6.3: Illustration of the optical and visual axes and relevant points in the right eye (top view). The optical axis connects the anterior vertex of the cornea (V) and the center of rotation (C). The visual axis connects the point of fixation (P) with the front nodal point (N), which extends through the rear nodal point (N'), to intersect with the fovea (F). The angle α offsets the visual axis on average by 5° in the nasal and 3° in the inferior directions.

6.1.2 User study

In this work, we conduct our own psychophysical study to experimentally determine the position of the no-parallax point for several users. For this purpose, we adapt the general setup proposed by Bingham [219]. As shown in Figure 6.4a, two surfaces are separated in depth and aligned such that when the users fixate at the rear one, they cannot see its red half. However, when instructed to fixate at the gaze target with an angular displacement θ (with the head fixated), the rotation of the eye shifts the no-parallax point of the eye towards the left. This reveals an extent E of the rear surface and the red half becomes visible in the periphery. Note that this is only possible if the center of perspective of the eye, and equivalently the no-parallax point, is located in front of the center of rotation.

The distance between no-parallax point and center of rotation, NC, can then be calculated using the target distances L_1 and L_2 as:

$$NC = \frac{EL_1}{(L_2 - L_1)\sin\theta + E\cos\theta} \tag{6.1}$$

A more detailed derivation of this equation is included in Appendix C.2. We construct this setup with $L_1 = 0.5\,\mathrm{m}$, $L_2 = 1\,\mathrm{m}$ and $\theta = 30°$ in a controlled experiment to determine the largest extent, E, a user can detect. This is equivalent to determining the number of pixels the red region can be shifted towards the right before the user can no longer detect it. The configuration also means that scattering within the ocular media should not increase a user's ability to detect the red region due to the light not being able to enter the eye at or beyond this threshold. After converting pixels to meters, Equation 6.1 can be used to calculate NC for the user.

Stimuli. We use a 6" Topfoison liquid crystal display (LCD) with a resolution of 1920 × 1080 and an edge-lit light-emitting diode backlight as the far target, as shown in Figure 6.4b. This enables us to easily control and change the displayed extent using an attached laptop, to a precision of 0.069 mm (the pixel pitch of the display). Without eye rotation, the front half surface would completely occlude the red stimulus. Each trial contained both an extent stimulus with red and white regions and a control stimulus, where the full LCD displayed only white. The brightness was defined in the RGB space of the display. The white pixels were rendered at 80% brightness across red, green and blue channels to reduce the apparent brightness to approximately that of the red extent with 100% red alone.

Conditions. All stimuli were presented monocularly to the right eye while the user wears an eyepatch on the left eye. With the user accommodated at the L_2 distance, the eye's limited depth of field can cause the edge of the half surface to appear blurred. As Bingham et al. [219] described this as a confounding factor, we try to mitigate it using a lamp to illuminate the gaze target to stop down the pupil aperture and maximize the depth of field.

Subjects. Eight adults participated (age range 21–29, 3 female). Due to the demanding nature of our psychophysical experiment, only a few subjects were recruited, which is common for low-level psychophysics (see e.g. [13]). All subjects in this and all following experiments had normal

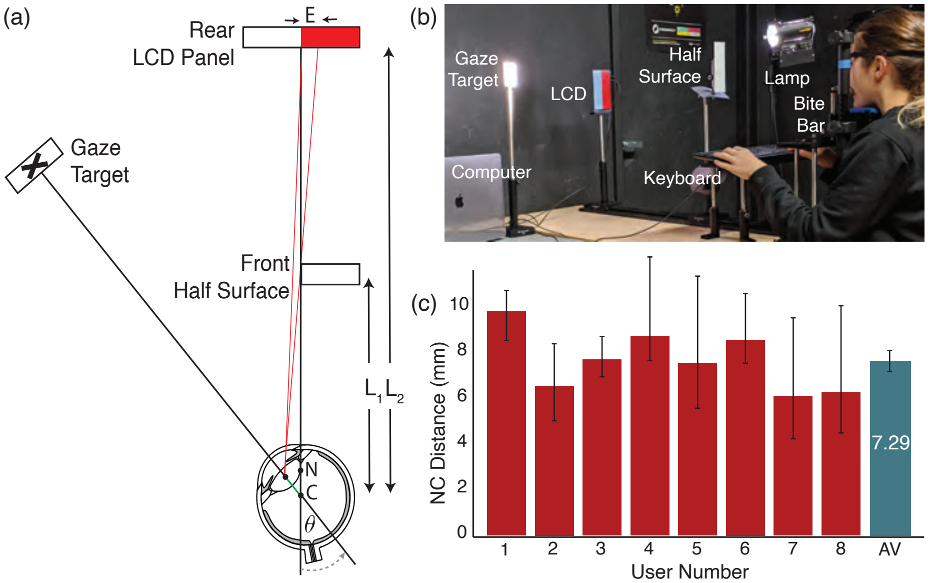

Figure 6.4: Psychophysical experiment to measure the position of the no-parallax point. (a) A diagram of the experimental setup. A half surface and an LCD panel are set up at distances L_1 and L_2, respectively. As the eye rotates counterclockwise about its center by an angle, θ, the no-parallax point is translated to the left revealing an extent, E, along the rear surface. The distance NC, shown in green, corresponds to the largest distance the red extent can be shifted towards the right before the user can no longer identify it. (b) Photograph of a user conducting the experiment. The head is kept stationary with a headrest and bite bar. The computer is used to change the stimulus on the LCD panel and record the user's response given by the keyboard. The lamp is used to illuminate the gaze target to reduce depth of field blurring of the edge of the half surface. (c) Results of psychophysical user experiment. The NC distances measured for each of the 8 study participants are shown with 95% confidence interval represented as an error bar. The final column (AV) represents the mean of all participants, 7.29 mm, with error bar showing the standard deviation of 1.25 mm.

or corrected to normal vision, no history of visual deficiency, and no color blindness. All subjects gave informed consent. The research protocol was approved by the Institutional Review Board at Stanford University.

Procedure. To start the session, each subject was instructed to use the left and right arrow keys to shift the red portion on the screen such that they could just see it when looking down the center of the targets. After subtracting one pixel, this was used as $E = 0$ for all subsequent trials.

Each trial constituted a two-alternative forced choice (2AFC) test, and subjects were asked to use the keyboard to choose which of the two displays contained the red extent. The keyboard could also be used to toggle between the two displays as the users desired. Most users did so less than 10 times. However once a selection was made, this concluded the trial. No feedback was provided. Subjects were instructed to fixate only as far as the cross target, but were free to look back to the far display if desired.

Subjects completed 60 trials, consisting of 12 displayed extent, E, configurations, each tested 5 times. The experiment took about 20 minutes per subject to complete, including instruction and centerline calibration.

For the first block of 30 trials, E for a given trial was randomly chosen from 6 evenly spaced values between 20 and 80 pixels (1.38 and 5.52 mm), covering the range of expected values. For the second block of 30, E for a given trial was randomly chosen from 6 evenly spaced values between the values of E from the previous trial block where the user was getting less than 90% and more than 60% correct. This paradigm was chosen to maximize sampling around the threshold value, without causing observable fatigue.

Analysis. For each E displayed, we compute the proportion of correct responses. Using the *psignifit* Python package [220], we fit a psychometric function to each subject's data using Bayesian inference. Each psychometric function gives us a detection threshold, measured as pixel shifts from the initial set position. The thresholds represent where the psychometric function exceeded a 75% chance for a correct response. Individual psychometric fits are included in Appendix C.4. This is converted to meters using the pixel pitch of the display (0.069 mm) and NC is then calculated using Equation 6.1.

Results. The results of this experiment are shown in Figure 6.4c, giving a mean of 7.29 mm for NC, which is within the originally expected range of values (7–8 mm). Surprisingly, we observe a variation of about 3.54 mm among our subjects, indicating that there may be value in measuring and accounting for individual variation. However, the difficulty of measuring a person's NC distance makes such an approach impractical at the moment, so we continue to model an "average observer" as having $NC = 7.29$ mm for the remainder of this chapter.

6.2 Implications for Stereo Rendering

In this section, we study the effects of the no-parallax point on binocular vision and the horopter, leading to the prediction of a surprisingly high degree of disparity distortions with conventional stereo displays. We outline a gaze-contingent (GC) stereo rendering pipeline that takes the no-parallax point into account for precise disparity rendering.

6.2.1 Binocular Vision and the Horopter

The binocular horopter refers to the set of points in space that give rise to the same disparity on the retina [221]. Thus, the horopter provides a useful tool for analyzing and comparing different models for binocular vision. It is geometrically modeled as an arc on a Vieth-Müller circle formed by the two no-parallax points and the fixation point [221]. The choice of no-parallax point determines the

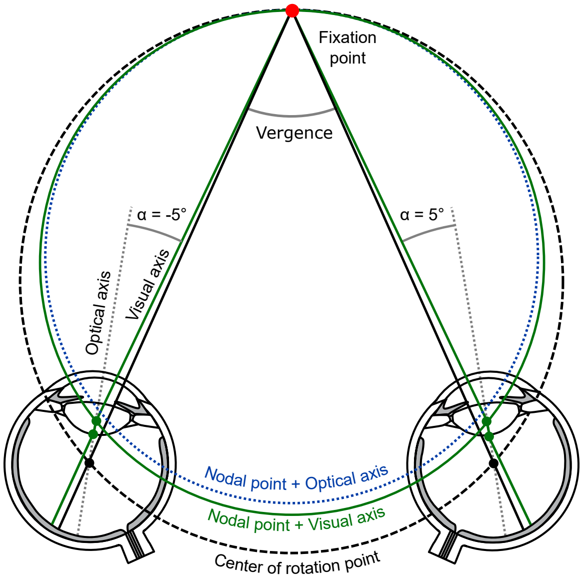

Figure 6.5: The horopters predicted for various eye models: the center of rotation model (black dashed), the front nodal point with the gaze vector being the optical axis (blue dotted) and the visual axis (green solid). Note that fixation leads to a different eye rotation for each axis (only visual axis-aligned eyes are shown here). The angle α is exaggerated for clarity.

specific shape of the horopter [222]. In Figure 6.5 we show geometrical horopters for the no-parallax point in the center of rotation, as commonly used in computer graphics (black), and in the front nodal point, as used here. The shape is further differentiated by the choice of gaze vector where the angular offset of the visual axis used in our model (green) yields a different horopter than the approximation using the optical axis as used by Konrad et al. [18] (blue). In the following, we outline an adequate rendering pipeline for our model and then analyze the expected disparity distortions when using other models.

6.2.2 Stereo Rendering

Traditional stereo rendering models the projection of 3D points into eye coordinates using a matrix–vector multiplication of the matrix $\mathbf{P}_{L/R} \cdot \mathbf{E}_{L/R} \cdot \mathbf{V} \cdot \mathbf{M}$ with a vertex specified in object coordinates. Here, $\mathbf{M}$ is the model matrix, $\mathbf{V}$ is the view matrix, $\mathbf{E}_{L/R}$ is the eye matrix and $\mathbf{P}_{L/R}$ the projection matrix for left and right eye, respectively. Accounting for the no-parallax point requires changes to the eye and projection matrices, which we describe in the following.

Eye matrix Assume that the centers of rotation of the eyes are $\mathbf{C}_{L/R} = (\mp\frac{\text{IPD}}{2}, 0, 0)$, where IPD is the IPD. Conventionally, the eye matrices are defined as translations into the eye centers, i.e. $\mathbf{E}_{L/R} = \mathbf{T}(-\mathbf{C}_{L/R})$, using the translation matrix $\mathbf{T}$. To account for the distance between centers of

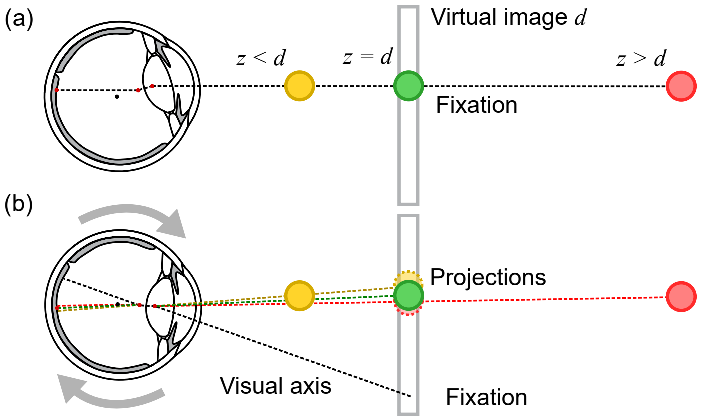

Figure 6.6: Ocular parallax for a virtual image at distance d and objects at distances $z < d$ (yellow), $z = d$ (green) and $z > d$ (red). (a) All points project to the same screen coordinates and a single spot on the fovea when the green point is fixated. (b) After a saccade the display projection of the near point moves up, the projection of the far point moves down and the projection of the middle point remains the same. The retinal image changes accordingly.

projection and rotation, NC, we calculate the gaze-dependent location of the no-parallax point with respect to $\mathbf{C}_{L/R}$

$$\mathrm{N}_{L/R} = \mathbf{R}\left(-\theta^{(v)}_{L/R}, -\theta^{(h)}_{L/R}, 0\right) \cdot \mathbf{R}\left(-\alpha_{L/R}\right) \cdot \begin{pmatrix} 0 \\ 0 \\ -\mathrm{NC} \end{pmatrix} \tag{6.2}$$

Here, $\alpha_{L/R} = (-3°, \mp 5°, 0)$ is the offset, in eccentricity angle, between optical and visual axis for the two eyes [218], $\theta^{(h,v)}$ represent the horizontal and vertical gaze angle for each eye, $\mathbf{R}$ is a 3×3 rotation matrix using Euler angles, and we use $\mathrm{NC} = 7.29$ mm from our earlier experiment. The eye matrices then become $\mathbf{E}_{L/R} = \mathbf{T}(-\mathrm{N}_{L/R}) \cdot \mathbf{T}(-\mathbf{C}_{L/R})$. Note that this notation uses a right-handed coordinate system, such as that used by OpenGL.

Projection matrix We use standard asymmetric off-axis perspective projection matrix $\mathsf{P}_{L/R}$ defined for a magnified virtual image of the microdisplay at distance d [18]. Note that the projection matrix depends on the gaze-dependent position of the no-parallax point (Equation 6.2). As illustrated in Figure 6.6, for 3D points located at distance d no ocular parallax is observed, i.e., this is the zero-parallax plane. Setting the parameter d to match the virtual image distance of a near-eye display is critical for correct reproduction of disparity with gaze-contingent stereo rendering.

6.2.3 Disparity Distortion

We model the image formed on the retina using the considered binocular projection model to quantify the magnitude of the expected disparity distortion. In Figure 6.7, we present the differences of the vergence angles predicted for the same fixation points. We further use the disparity perception model by Didyk et al. [223] to compute just-noticeable differences (JND). Values above 1 JND predict visibility of the predicted distortions. Note, that verging at each fixation point requires eye rotation to achieve required gaze eccentricity. We mark the normal range of horizontal eye rotation (±45° [26]) by the red lines.

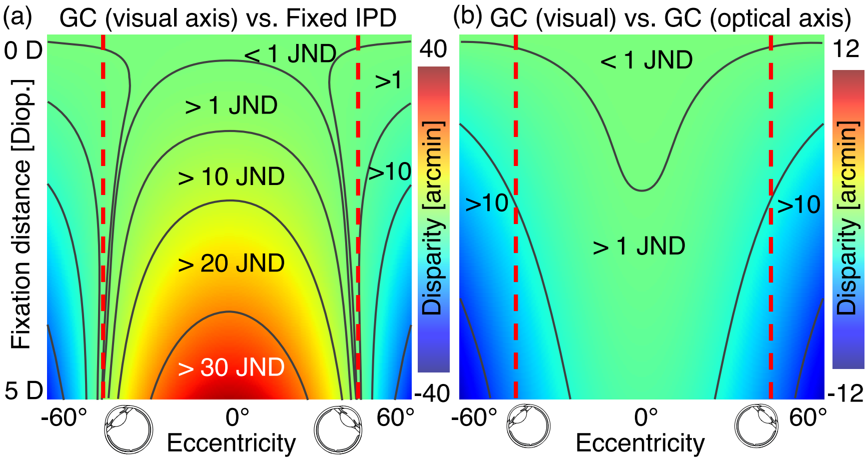

Figure 6.7: The difference of eye vergence angles predicted for different fixation points by the gaze-contingent (GC) model with visual axis and either the standard model using center of rotation (a) or the GC model with optical axis (b). The fixations are expressed relative to the midpoint between both no-parallax points in relaxed state. The initial IPD = 64 mm was defined for the eyes looking straight ahead. The isolines mark levels of stereoacuity JNDs from Didyk et al.'s [223] model for the optimal spatial frequency of 0.4 cpd. Values of JND of 1 and larger predict visibility of a difference in a direct comparison for an average observer. The range of eccentricities covers the full binocular FOV achieved through combination of gaze and retinal eccentricities [217]. Note, that normal range of eye rotation needed for gaze fixation is delimited by the red bars [26].

The left panel compares our full model and the model based on the center of rotation. We observe that even for fixations as far as 2.5 m the difference of vergence angles yields visible disparity differences. This difference further grows with decreasing distance. Additionally, the horizontal eccentricity affects the shape of the distortion field. While the model predicts larger perceived distances in our model for the central visual field, this trend is reversed for eccentricities above ≈ 40°.

This effect can be practically demonstrated in AR where real objects are superimposed with virtual objects. In Section 6.4, we show that traditional stereo rendering causes visible misalignment of physical and digitally rendered objects and we demonstrate how gaze-contingent stereo rendering

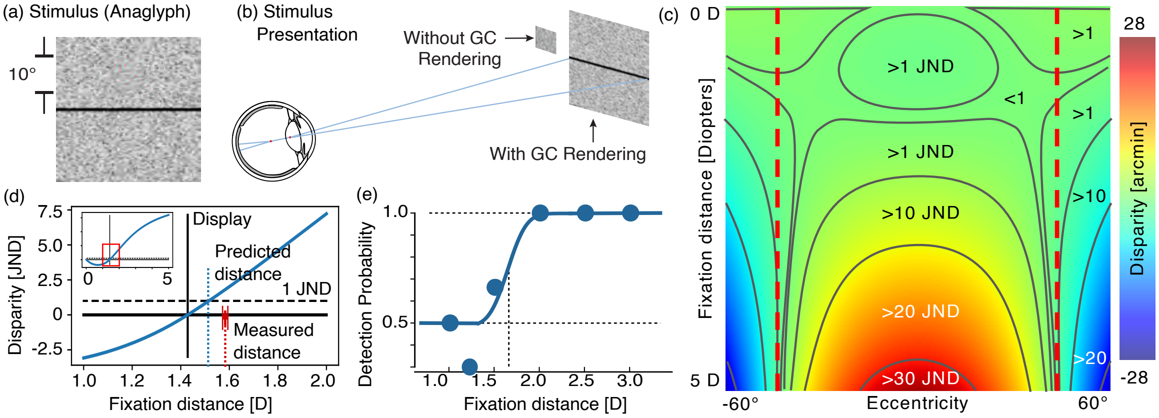

Figure 6.8: Verifying the detection threshold for gaze-contingent rendering. (a) The stereoscopic stimulus visualized in anaglyph. (b) Conceptual side view (schematic) of the stimulus presentation. The stimulus rendered without gaze-contingent (GC) rendering appears to pop out from the background, unlike the other stimulus (unseen under the black line), which are both rendered with GC rendering. (c) The predicted disparity differences between models with and without ocular parallax for a VR display with a display distance $d = 0.7\,\mathrm{m}$. The red bars delimit normal range of horizontal eye rotation which restricts the range of gaze fixation eccentricities [26]. (d) The JNDs for different fixation distances of the central vision around the display (the inset shows larger distance range). The red interval marks the detection threshold and SE interval measured in our experiment. (e) An example of psychometric function fit for one user.

reduces this issue significantly.

The right panel of Figure 6.7 further compares our model with a variant that assumes the optical axis, instead of the visual axis, to be a good approximation of the gaze direction (Figure 6.5). While the differences for the central visual field are relatively small, the role of the axis is notable for larger viewing angles.

Finally, we also analyze the effect in terms of no-parallax point separation and its change with respect to the initial IPD. For an average IPD of 64 mm and a fixation distance of 30 cm we observe an effective decrease of viewpoint separation to as low as 62.5 mm. That corresponds to a shift from 68 to 52-percentile of the IPD distribution in the female population [224] which is a deviation that typically requires user adjustment in common VR systems. To illustrate this phenomenon, we explore the produced shape distortions in a VR scenario in Section 6.3.

6.2.4 Verifying the Model

Figure 6.7 predicts the visibility of the difference between both models for an average observer and a theoretical display with virtual image distance, $d \to \infty$. To validate our model we tested this hypothesis with a user study using a VR platform. Using the same equipment as in Section 6.3, we conducted a detection study using a random dot stereogram (RDS) stimulus at varying depths. See Appendix C.3 for more detail on RDS stimuli. The users were presented with two RDS stimuli, rendered at the same depth, but placed on top of each other (i.e., vertically) in the center of the visual field (Figure 6.8a). Each stimulus was a square 10° in width. In random order, one stimulus was rendered with and one without gaze-contingent stereo rendering. As a result, one of the stimuli had a different disparity than the background, which is itself an RDS stimuli rendered at the same depth with our gaze-contingent mode (illustrated in Figure 6.8b). Users were tasked with a 2AFC and used a keyboard to report whether the upper or lower segment contains the patch that protrudes from the background. All stimuli were rendered at a distance of 1, 1.33, 1.5, 1.75, 2, 2.5 or 3 D (diopters, inverse meters), with 6 trials at each distance in a randomly shuffled order. A black screen was shown for 3 s between trials to assist in eye adjustment. For each of the 7 distance configurations, we computed the proportion of correct responses. Using Bayesian inference methods [220,225], we fit a psychometric function to each subject's responses, finding the fixation distance with 75% detection threshold (shown in Figure 6.8e). 11 subjects (6 male, 5 female, aged 18–54) took part in the study. Individual psychometric fits are included in Appendix C.4.

Using a digital single-lens reflex (DSLR) camera, we measured the display distance d of our HTC Vive Pro to be $\approx$ 70 cm. This changes the distribution of depth distortion in Figure 6.7 such that the 1 JND occurs at a distance of 66 cm (see Figure 6.8c). We found that the depth distortion was detectable, on average, at a distance of 62.8 ± 1.3 cm or 1.59 ± 0.033 D (Standard Error, SE). This confirms the importance of taking the no-parallax point into account for accurate stereo rendering. While we chose not to additionally burden users with measuring their stereoacuity, the similarity of the mean measured detection distance to the model-predicted expected distance of 1.52 D confirms our model's ability to predict observable disparity distortions of different rendering models (Figure 6.8d).

6.3 Depth Distortion in VR

The analysis and experiments in Section 6.2.3 predict visibility of disparity distortion for rendering that ignores the gaze-contingent shift of the no-parallax point. Here, we explore this issue further and experimentally test a hypothesis that a shape of a 3D object rendered using the traditional stereo rendering will appear distorted as a function of fixation distance. Further, we validate that our gaze-contingent rendering reduces this distortion significantly.

Hardware and Software. We used an HTC Vive Pro VR system, which has a diagonal FOV of 145°, a refresh rate of 90 Hz and a 1440×1600 pixel organic light-emitting diode display per eye, resulting in a theoretical central resolution of 4.58 arcmin/pixel. The HTC Vive Pro supports built-in IPD adjustment. Unity was used as the rendering engine for all rendering modes and user experiments.

Stimuli and Conditions. For this experiment, we require a stimulus whose apparent shape does not rely on metric structure, but only on ratios of its dimensions. For this purpose, we emulated the triangle wave experiment performed by Glennester et al. [226] for measuring stereoscopic depth constancy. As illustrated by the schematic in Figure 6.9a, this stimulus is a triangle wave formed by an RDS pattern. It is rendered such that the amplitude is half the period of the peaks. Thus, if the depth rendering is physiologically correct, the dihedral angle of the peaks and troughs should be at 90° (shown in dark green). However, if the depth space is perceived as stretched, as we predict is the case without gaze-contingent rendering (see Section 6.2), then the angles should appear smaller than 90°. Similarly, a perceived compression of the depth space would increase the angles. A scaled crop of an illustrative anaglyph reproduction can be seen in Figure 6.9b.

During the user study two identical patterns were shown horizontally side-by-side, one with and one without GC rendering, at a depth of either 0.3, 0.5 or 0.7 m. These depths were chosen such as to increment to the measured display distance of the HTC Vive Pro (0.7 m), where no disparity distortion should be observed. We refer to the rendering without GC as fine-tuned (FT) rendering, since we first set the subject's IPD using the physical knob provided on the device, and then allow the user to further tune the horizontal separation of both virtual images in projection matrices. This was done by rendering a single pattern at a fixed far distance and instructing the users to tweak the separation until the stimulus exhibited 90° angles. The distance of 2 m was chosen as a compromise, where the effect of GC rendering diminishes yet binocular disparity is still a relevant depth cue. The GC mode was identical to the FT mode, but with the modifications described in Section 6.2.2. Finally, to ensure fair comparison of these rendering modes, we shift the center of projection in the FT mode from the center of rotation to the position of the GC no-parallax point of a user looking towards optical infinity. This ensures that the only difference between the two modes comes from the gaze-contingent movement of the no-parallax point and not from an arbitrary initial position bias.

Procedure. Before starting the trials, each of the nine adult subjects (age range 18–54, 4 female) completed the calibration procedure described above to set-up the FT rendering mode. Each trial then constituted a 2AFC, where one of the three tested depths was randomly chosen for rendering, and subjects were asked to choose which of the two randomly ordered patterns (left or right) exhibited angles closer to 90°. A total of 24 trials were conducted, taking each user approximately 10 minutes to complete the study.

Results. The results of the comparisons averaged across users and trials are plotted in Figure 6.9c. At 0.3 and 0.5 m, the GC rendering was chosen as closer to the target of 90° in 73.6% and 62.5% of trials, respectively. This is significantly more than FT ($p < 0.001$, respective $p < 0.05$, one-tailed binomial test). The visibility of the difference decreases towards the display distance $d = 0.7\,\text{m}$ where GC was only preferred at near chance level of 51.4%.

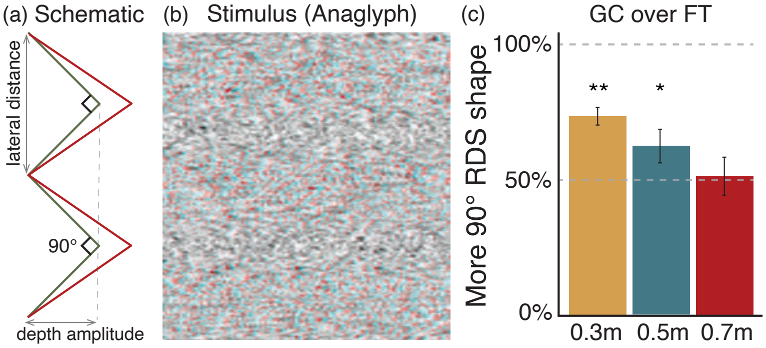

Figure 6.9: Evaluating shape distortion of virtual content. Subjects simultaneously viewed two identical triangle wave random dot stereogram (RDS) stimuli, one rendered with fine-tuned (FT) IPD and the other with gaze-contingent (GC) rendering. (a) A schematic of a cross-section of the stimulus. Designed to evaluate shape distortion caused by incorrect depth scaling, the dimensions of the RDS triangles are calculated such that the amplitude of the peaks (in depth) is twice the lateral distance (period of the pattern). If the depth space is correct, the dihedral angle of the peaks should be at 90° (green), but if the depth space is stretched (as it is without gaze-contingent rendering), the angles should appear smaller (red). (b) An illustrative anaglyph rendering of the stimulus (not to scale). Both stimuli were rendered at a target depth of either 0.3, 0.5 or 0.7 m and we asked subjects to indicate which of the two contained angles is closer to 90°. (c) The percentage of times that the gaze-contingent mode was chosen as more accurate per distance. Despite the seemingly small effect size, shape distortion is detectable, in particular for closer distances. Error bars represent Standard Error (SE) and significance is indicated at the $p < 0.05$ and 0.001 levels with * and ** respectively.

These results suggest that accounting for the gaze-contingent no-parallax point is important for correct depth scaling needed to properly convey relative distance and shape of objects within a scene, particularly when a user is verging to a close object or familiar shape, such as a cube. Judging the angle at which two planes meet requires higher-level reasoning and combination of both absolute and relative depth cues. We expect that the distortion can be even easier to detect in tasks where the relative displacement of two surfaces alone is a sufficient cue. We explore this hypothesis in the following AR alignment study.

6.4 Alignment Inaccuracy in AR

Many applications in AR desire accurate alignment of digital and physical objects. For example, a surgeon aligning medical data to a patient will want to rely on it being displayed in the correct place. As such, accurate depth rendering is critical. Section 6.2.3 predicts displacements of virtual objects when the position of the no-parallax point is not taken into account. Here, we experimentally verify visibility of this effect in an AR environment. We further test a hypothesis that our gaze-dependent rendering can noticeably improve the accuracy of alignment between the virtual and real objects.

Hardware and Software. We used a Microsoft HoloLens 1 optical see-through AR headset, which has a diagonal FOV of 34°, a refresh rate of 60 Hz and a 1280×720 waveguide display per eye, resulting in a theoretical central resolution of 1.39 arcmin/pixel. As with the VR user experiment, we again used Unity to render all modes and control the user experiment.

Stimuli and Conditions. The stimuli consisted of a single 8 cm tall flat surface, textured with a playing card image (see Figure 6.10a), displayed at target fixation distance of either 0.5, 1.0, 1.5 or 2.0 m. Again, these distances were chosen such as to increment to the display distance of the Microsoft HoloLens (2 m), where no disparity distortion should be observed. A physical target was placed at the same distance from the user (as measured from the user's eyes in the physical world) but with a small lateral displacement, such that the virtual and physical objects would appear side by side (see Figure 6.10b). In the experiment, subjects viewed the rendered stimulus in three rendering conditions: conventional (HL, i.e., as provided by the Windows Mixed Reality SDK in Unity), fine-tuned (FT), and gaze-contingent (GC). For the HL rendering mode, we implemented the online instructions provided by Microsoft for rendering to the HoloLens with Unity. We let the rendering be set up by the supplied Windows Mixed Reality SDK and only adjusted the IPD setting for each user through their Developer Portal interface. For the FT rendering mode, we followed a similar procedure for adjusting virtual image separation as in Section 6.3. In this case, the manufacturer-provided calibration was fine-tuned for each subject by aligning the card stimulus at a calibration distance of 2 m. Finally, the GC rendering mode was identical to the FT mode, but with the same modifications as in Section 6.3. This was again motivated by the desire to show that even a more accurate calibration of the IPD is insufficient to remove the misalignment observed at closer distances if the position of the no-parallax point is not taken into account. While wearing the headset, an SR Research head rest was used to keep the subject's head fixed with respect to the physical targets throughout the study.

Procedure. Each set of trials began with the IPD fine-tuning task required to set up the FT rendering for each of the thirteen participants (age range 18–54, 7 female). Each trial constituted a 2AFC, where one of the three target depths was randomly chosen, and subjects were asked to choose which of the two selected modes provided the best alignment in depth with the physical target, which was placed by the researcher before the stimulus was shown. Subjects had the ability to freely switch between the modes using a keyboard key before making a selection, though most users only made a single switch per trial. A total of 12 trials were conducted comparing FT and HL rendering, followed by a short rest break. After which, the calibration was repeated, and another 12 trials were conducted comparing FT to GC rendering.

Results. The results of the comparisons averaged across users and trials are plotted in Figure 6.10c and d. At all measured distances the FT rendering achieves significantly better alignment of the rendered and physical stimulus than the HL rendering (100%, 100%, 94.2% and 86.5% of trials, $p < 0.001$, one-tailed binomial test). Some users found it harder to judge the difference between the two modes as the planes moved further away, but overall, it can be seen that fine-tuning the user's IPD measurement by calibration almost consistently improved alignment compared to the conventional approach.

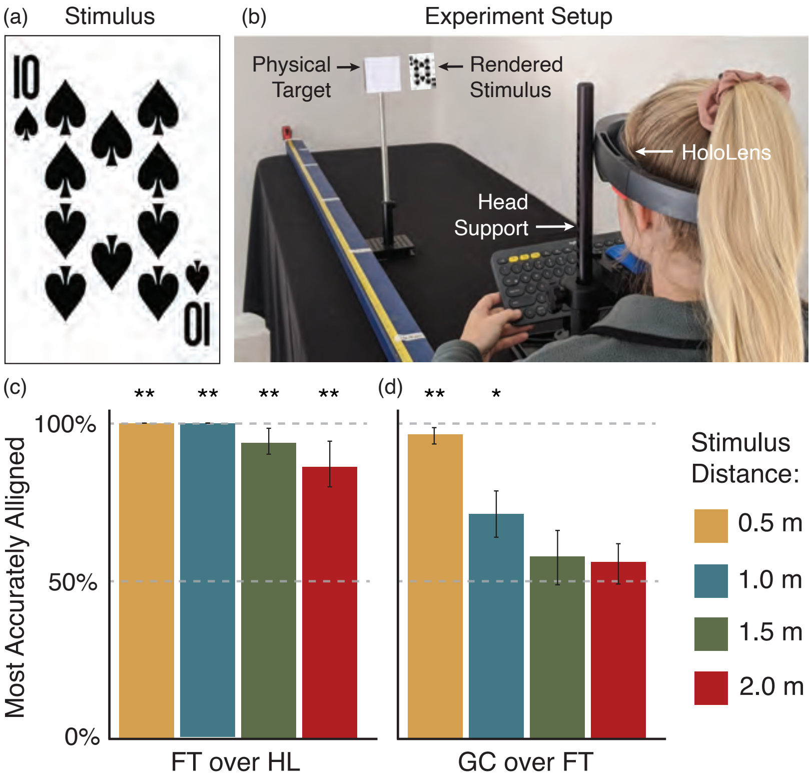

Figure 6.10: Evaluating alignment of real and virtual content. Subjects viewed a playing card (a) rendered at a target depth of either 0.5, 1.0, 1.5 or 2.0 m next to a physical reference. In the first set of trials (c), this stimulus was presented with either native HoloLens (HL) or fine-tuned (FT) rendering and we asked subjects to indicate which rendering mode provided the most accurate alignment with the physical target. A photograph of the experiment set-up is shown in (b) (the card is added for illustrative purposes). In the second set of trials (d), subjects were asked to compare fine-tuned and gaze-contingent (GC) rendering. Results of these comparisons show the percentage of times the first member of the pair was chosen over the second. It can be seen that using an initial calibration procedure to accurately measure the subject's IPD significantly improved alignment compared to the standard HoloLens approach for all distances. Furthermore, GC rendering was able to further improve alignment at closer distances indicating that it is most critical for arm's reach viewing. Error bars represent Standard Error (SE) and significance is indicated at the $p < 0.05$ and 0.001 levels with * and ** respectively.

Moreover, additional improvement of alignment was observed in the GC rendering mode which achieved a significant preference over the FT for distances of 0.5 m (96.2%, $p < 0.001$) and 1.0 m (71.2%, $p < 0.05$). While it was more difficult to detect differences for larger distances (57.7% at

1.0 m) the results of the experiment confirm our hypothesis that accounting for the gaze-contingent shift of the no-parallax point is crucial for accurate reproduction of stereoscopic disparity. Although fine tuning of the IPD proved helpful, the gaze-contingent rendering was required to ensure good alignment of virtual and physical objects across distances in AR. While the shift may become indistinguishable for far away objects, gaze-contingent stereo rendering could be critical in several near AR tasks, including AR-assisted surgery, maintenance, and training.

6.5 Discussion

In summary, we study the disparity distortion induced by ignoring the gaze-contingent location of the no-parallax point in the human eye. Using several user studies, we experimentally validate the location of the no-parallax point and demonstrate that modeling it accurately during stereo rendering significantly reduces disparity and shape distortion in a VR setting and significantly improves consistent alignment of physical and digital objects in an optical see-through AR setting.

The results of our experiments show that disparity distortions are easier to detect in the AR alignment task (Section 6.4) than in the VR shape matching task (Section 6.3). This is expected as the human visual system is sensitive to even small disparity changes between a physical reference and a digitally rendered object [223]. On the other hand, the shape judgment task required subjects to interpret the relative disparity in the context of estimated object distance. Without a real-world reference, the relatively poor absolute depth cue of eye convergence likely increased the difficulty of the task [227].

Limitations and Future Work Gaze-contingent stereo rendering relies on robust gaze tracking, however, since the magnitude of parallax changes gradually with eye rotation, we do not require extreme accuracy in gaze prediction. For the model situation in Figure 6.8 and a central vision fixation distance of 1 m, a 1° differential tracking error between gaze angles of both eyes results in a disparity rendering error of 12 ” (arcseconds) which is a difference on the limit of human stereoacuity in ideal conditions [223]. The technique is more sensitive to latency as a delayed response could produce visible jumps of disparity. Thus without implementing significant temporal smoothing, eye tracking therefore remains a challenging problem. Furthermore, any stereo rendering approach, including ours, is only as good as the optics and calibration of the headset and the accuracy of user-specific parameters. Variation in lens distortion as the eye rotates off axis and across the lens (commonly referred to as pupil swim) can cause its own disparity distortion. This is not something our approach inherently corrects for, but it could be used in combination with existing pupil swim correction approaches (see Section 2.2.3). Similarly, in practical use cases there is likely to be per-user variation in the parameters of our eye model or even inaccuracies when measured on a per-user basis. Even so, our model generally pushes the disparity in the empirically correct direction.

While our studies demonstrate statistically significant effects, they all use task-specific stimuli. Studies in more complex environments where the user's cognitive load is higher may provide an interesting setting for future user experiments. Moreover, it would be interesting to explore the adaption of this technique for varifocal and multifocal displays, and the interaction of gaze-contingent stereo rendering with other depth cues.

6.6 Summary

As VR and AR display systems strive to pass the visual Turing test, being able to render dynamic gaze effects is important for their success. While gaze-contingent display techniques to improve focus cues and visual comfort have received a lot of attention recently, with this work, we hope to demonstrate that eye tracking can also improve stereo rendering. We demonstrate that accounting for the dynamic shifts in the optical center of the eye can significantly improve disparity and depth rendering; critical for achieving perceptual realism in emerging wearable computing systems.

Chapter 7

Concluding Remarks

In this dissertation, we introduce and evaluate several approaches that aim to create VR and AR display systems capable of passing the visual Turing test using the limited bandwidth available in current-generation systems. In particular, we build the foundations for new gaze-contingent display techniques with higher bandwidth saving potential by exploiting additional perceptual limitations of the HVS, including variations in spatio-temporal sensitivity across the retina (Chapter 3) and attentional affects (Chapter 4), or by reducing system latency (Chapter 5). Additionally, we affirm how gaze-contingent display is also powerful for rendering dynamic effects that improve perceptual realism, namely by accounting for dynamic shifts of the optical center of the human eye (Chapter 6).

Here, we summarize key lessons learned and how the work described in this dissertation advances the related fields of study. We also discuss limitations of the proposed approaches, new research questions, and open problems which merit attention in future work.

7.1 Lessons Learned

Towards Spatio-temporal Rendering. The perceptual model for eccentricity-dependent spatio-temporal flicker fusion that we describe in Chapter 3 aims to predict the temporal thresholds at which a visual stimulus becomes (in)visible, for a given spatial frequency, eccentricity and luminance. As such, this work forms the enabling foundation for new temporally foveated graphics techniques, with the potential for bandwidth savings up to seven times higher than those afforded by current spatial-only foveated models.

Inspired by a casual observation whereby a streetlight appears to irritatingly flicker in the periphery, but appears stable upon direct gaze, here we set out to measure how human temporal sensitivity is “antifoveated” in that it peaks in the mid-periphery. With foveated graphics focusing so heavily on variations in spatial sensitivity, we had to wonder why we could not find discussions of the temporal domain in the literature and set out to provide a model that could change this. This work has already inspired more interest in the field, with subsequent work demonstrating a comprehensive model of CSF [56], which can also model relative sensitivity. However, it is still unclear how such a temporally foveated display would be realized and further innovation is needed in graphics and display technology to see this through.

Towards Attention-aware Rendering. The attention-aware CSF model that we describe in Chapter 4 illustrates how our spatial perception changes depending on how we allocate our attention across our visual field. Conventional methods for measuring CSF in the periphery encourage users to keep their eyes fixated centrally but focus their perception in the periphery – a configuration that likely overestimates sensitivity for most natural viewing conditions. We show that drawing attention to the gaze position with a visual-based task does in fact decrease contrast sensitivity, with an increasing effect with eccentricity. Consequently we demonstrate that tolerance for foveation in the periphery is significantly higher when the user is concentrating on a task in the fovea, and estimate potential bandwidth savings up to seven times higher that those afforded when this effect is not considered.

Inspired by a recent explosion of wearable devices for measuring brain activity (e.g. EEG and NIRS), we wondered whether cognitive state would affect perception. After diving into the literature, we finally stumbled upon the concept of visual attention. While the concept of of "brain bandwidth" intuitively made sense, it was not until we tried the foveation user study for the first time and could actually feel ourselves tunnel-visioning, that we were convinced. But of course, the biggest hurdle to adoption is the lack of a method for measuring this cognitive state. However, as we see extra sensing capabilities being adopted in VR and AR devices e.g. outward facing cameras, EEG and heart rate measurement, it's almost inevitable that we start seeing new techniques for monitoring cognitive state to improve user experience.

Low-latency Foveated Display. The latency, or the time between a change in the viewer's gaze and the resulting change in the display's pixels, of a gaze-contingent display system is not just important for managing user discomfort, but is critical to the bandwidth savings achievable by foveated graphics techniques. In Chapter 5, present a custom, low-latency foveated compression system and demonstrate that up to double the bandwith savings could be achieved, just by lowering the system latency.

Despite being key to their success, the system latency of gaze-contingent displays is rarely discussed with any kind of vigor. Several works quote poorly justified thresholds or acceptable ranges, but few truly dive into their impact. With this work we hope to frame system latency as not just a systems engineering challenge, but as axis for improving the performance of every algorithm that makes up the suite of foveated graphics. If we were able to demonstrate a doubling in performance using our simple approach, we are excited to see how other authors use this axis to improve their own works.

Gaze-contingent Stereo Rendering. Humans have very high stereoscopic acuity, and so even small inaccuracies in stereoscopic rendering can lead to objects being perceived at a different depth that intended. Gaze-contingent stereo rendering, described in Chapter 6, improves the accuracy of disparity and depth rendering by accounting for the dynamic shifts of the optical center of the human eye. Our findings demonstrate significant improvements in shape distortion in a VR setting, and consistent alignment of physical and digitally rendered objects in AR.

We've gotten away with conventional stereo rendering for while now, since the vergence–accommodation conflict has prevented the rendering of content close to the eyes. But with applications such as surgical guidance, begging for solutions, and the recent of emergence of varifocal and multifocal display paradigms, it's likely that the VR and AR displays of the future will quickly reveal the importance of gaze-contingent stereo rendering. With disparity being the most significant depth cue for objects within 1 m from the eyes [108], we would hope that even millimeters of error would be deemed unacceptable before we would allow surgeons to use this technology to decide where to cut into patients.

7.2 Future Directions

As VR and AR display systems strive to pass the visual Turing test, gaze-contingent rendering and display paradigms are likely to play a key role. However, these solutions do not come without challenges. In fact the very technology that enables it, eye tracking, may also present its biggest challenge. All of the techniques described in this dissertation are heavily dependent on real-time, accurate gaze position information. Eye tracking technology has advanced significantly since its modern inception in the early 20th century, but still, can hardly be called a "solved problem". The accuracies and latencies of the commercial devices listed in Chapter 2, are recorded until the most ideal of circumstances, and in practice are often much less ideal [119]. Such accuracies also often rely on per-use calibration procedures which can be disruptive to a smooth user experience. Furthermore, 3D gaze tracking is almost always significantly worse and still an area of active research [9]. While an outward facing "scene" camera has been a popular approach to help disambiguate sharp changes in depth fixation, this adds to the privacy concerns already on the minds of consumers [228]. Finally, power consumption and heat generation is also a factor to consider as VR and AR manufacturers try to optimize battery life and comfort of their devices.

In working towards the goal of achieving perfect perceptual realism, much of the future work in gaze-contingent displays will focus on accounting for individual differences between users. For example, accounting for the spatio-temporal or attentional sensitivity of a particular user. For gaze-contingent disparity rendering the nodal point, which we assumed to be the same for the entire population, could be measured and accounted for in the rendering for each individual. Furthermore, it's likely that these techniques need to be adjusted and integrated with a user's individual prescription corrective optics. Finally, combining several gaze-contingent display techniques together, for example, attention-aware with gaze-contingent stereo rendering, in perceptually correct manner will be a big undertaking for the VR and AR communities moving forward.

The ultimate goal for VR and AR displays is the ability to make a user question whether what they are seeing is real or virtual. While today, VR can provide amazingly immersive experiences and AR can render impressive world-locked content, the percept is still not fully convincing. It struggles to match the spatio-temporal and depth sensing capabilities of the human vision under the limited compute budgets, hardware, and transmission bandwidths of wearable computing systems. In this dissertation, we hope to spur progress in improving the perceptual realism of these displays by demonstrating the powerful potential of gaze-contingent display.

Appendix A

Additional Material for Towards Spatio-temporal Rendering

A.1 Gabor wavelets in display space

In Section 3.1 of the main text, we describe the general form of a Gabor wavelet as being a complex sinusoid modulated by a Gaussian envelope, defined as:

$$g(\mathbf{x}, \mathbf{x_0}, \theta, \sigma, f_s) = \exp\left(\frac{-\mathbf{x} - \mathbf{x_0}^2}{2\sigma^2}\right) \cos\left(2\pi f_s \mathbf{x} \cdot [\cos\theta, \sin\theta]\right), \tag{A.1}$$

where $\mathbf{x}$ denotes the spatial location on the display, $\mathbf{x_0}$ is the center of the wavelet, σ is the standard deviation of the Gaussian in visual degrees, and f_s and θ are the spatial frequency in cpd and angular orientation in degrees for the sinusoidal grating function.

In our application, it is more convenient to describe spatial position in terms of eccentricity, measured in degrees of visual angle. Such transformation requires information about physical size of a pixel, p, and the dimensions and pixel resolution of the display. Using Figure A.1, we can see that:

$$\tan(e) = \frac{p(\mathbf{x} - \mathbf{x_C})}{d/M}, \tag{A.2}$$

where e is eccentricity, $\mathbf{x_C}$ is the location of the pixel directly in front of the eye, d is the distance to the virtual image and M is the magnification factor of the lenses. Finally, Equation A.2 can be re-arranged and substituted into Equation A.1 to re-define the Gabor wavelet equation in terms of eccentricity.

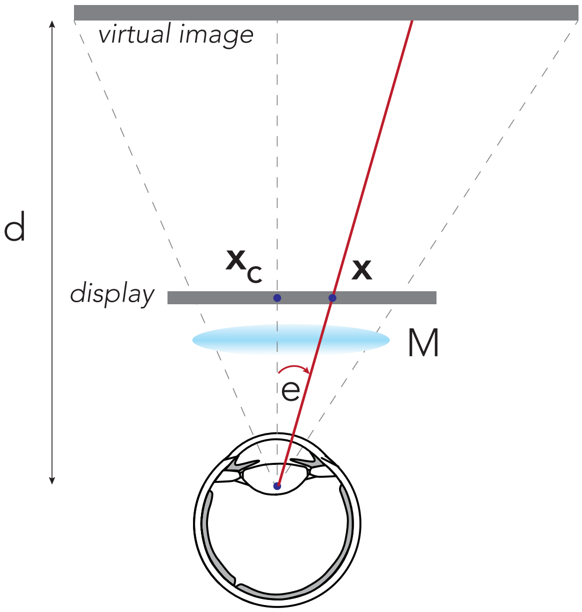

Figure A.1: Schematic illustrating the geometrical relationship used to convert spatial location in pixels to eccentricity in degrees of visual angle.

A.2 Fitting the Model

The model fitting in Section 3.2.1 of the main text introduces traditional regression fitting criteria as well as a set of constraints that each variant of the model should adhere to. We also account for the localization uncertainty in our measured sample point set $\mathbf{X} = (e, f_s, f_t)$ of eccentricities e, spatial frequencies f_s and detected CFF frequencies f_t. To this goal we expand $\mathbf{X}$ by linearly interpolating e in the entire extent of each stimulus $\mathbf{m} = [e \pm u]$ deg. This yields a new set $\mathbf{\Omega}$ where each measured CFF value is repeated 100 times with varying e covering stimuli radii. Further we split this set to perceptible stimuli $\mathbf{\Omega}_1$ with $f_t > 0$ and remaining imperceptible stimuli $\mathbf{\Omega}_0$ where CFF values could not be measured. $\mathbf{\Omega}_0$ is further expanded by acuity samples for fitting of the full model. These are points in form $(e, A(e), 0)$ sampled using the acuity model of Geisler and Perry [7] introduced in the main text. The e for is sampled for 100 points between 0 and 25 cpd which represents the measurement range of the original acuity data.

We express all our goals as loss expressions and therefore soften the originally hard constraints to find parameters. We are looking for a set of parameters $\mathbf{p}$ that minimizes the total loss for our CFF predictor $\Psi(e, f_s; \mathbf{p})$. First, an L2 loss is used to minimize the fitting error of the model as:

$$\mathcal{L}_r = \frac{1}{|\mathbf{\Omega}_1|} \sum_{(e,f_s,f_t)\in\mathbf{\Omega}_1} \|\Psi(e, f_s; \mathbf{p}) - f_t\|_2^2. \tag{A.3}$$

Second, we enforce the one-sided conservative and relaxed constraints using a thresholded linear loss as:

$$\mathcal{L}_c = \frac{1}{|\mathbf{\Omega}_1|} \sum_{(e,f_s,f_t)\in\mathbf{\Omega}_1} |\max\left(\beta(f_t - \Psi(e, f_s; \mathbf{p})), 0\right)|, \tag{A.4}$$

with $\beta = 1$ for the conservative model and $\beta = -1$ for the relaxed model fit. Finally, a similar approach is used to enforce zero predictions for imperceptible stimuli points as:

$$\mathcal{L}_a = \frac{1}{|\mathbf{\Omega}_0|} \sum_{(e,f_s,f_t)\in\mathbf{\Omega}_0} |\Psi(e, f_s; \mathbf{p})|. \tag{A.5}$$

Together, our loss is $\mathcal{L} = \mathcal{L}_r + \alpha(\mathcal{L}_c + \mathcal{L}_a)$ where α is 40 for the conservative model and 1 for the relaxed and full models. We optimized $\mathbf{p}$ using the Adam solver in PyTorch initialized by the Levenberg–Marquardt algorithm in Python's scipy package.

A.3 Estimation Study Results

Table A.1 shows the CFF values measured for each test Gabor wavelet. L1 refers to the original 380 cd/m^2 luminance, averaged across the 9 participants, while L2 and L3 are the stimuli re-tested at 23.9 cd/m^2 (L2) and 3.0 cd/m^2 the adaption luminance study (described in Section 3.2 of the main text), averaged across the 4 participants.

A.4 Additional Details on Validation Study

A.4.1 Stimuli

We use our custom high-speed VR display to conduct a validation study, presenting all videos monocularly to the right eye with the DLP in the 360 Hz 8-bit monochromatic mode (using only the green LED), as previously. The videos consist of a single image frame perturbed by Gabor wavelet(s) with carefully chosen parameters. Tables A.2 and A.3 list the exact list of chosen parameters for the 5 groups of conditions outlined in the main text.

Three different images were chosen for variety (shown in Figure A.2), extracted (with permission) from the stereo panoramas used in the work of Sitzmann et al. [229]. The same image pairings and stimuli were used for every user. To combine the background image and Gabor wavelet(s) the contrast of each was re-scaled to $[0.25, 0.75]$ and $[-0.25, 0.25]$, respectively, such that addition was not clamped.

A.4.2 Participants

Eighteen adults participated in the study (age range 18-53, 8 female). All participants have normal or corrected to normal vision and no history of visual deficits. The research protocol was approved

Table A.1: Measured CFFs for each test Gabor wavelet averaged across participants. As described in the main text, we define 6 orders by spatial frequency (and radius) of stimuli. The number and eccentricity locations per order were chosen based on radius to uniformly sample the available eccentricity range. f_s: spatial frequency, σ: wavelet standard deviation, e: eccentricity. L1 refers to the original 380 cd/m^2 luminance study (9 participants) while L1 and L2 are the 23.9 cd/m^2 (L2) and 3.0 cd/m^2 luminances re-tested for the adaption luminance study (4 participants) described in Section 3.2.1 of the main text. (*) Note that in practice the extent was limited by our display FOV and $f_s = 0.0055$ cpd is used for analysis.

Order	f_s (cpd)	σ (°)	e (°)	Av. CFF (Hz)		
				L1	L2	L3
0	0.000(*)	inf(*)	0.0	94.41	-	-
1	0.011	63.0	0.0	92.80	78.97	69.41
2	0.041	17.2	0.0	85.99	68.11	62.33
			19.2	89.63	-	-
3	0.154	4.6	0.0	70.29	55.78	50.73
			24.5	87.54	-	-
			48.2	76.90	-	-
4	0.571	1.2	0.0	56.92	48.43	39.79
			14.8	71.15	61.94	55.22
			29.2	71.58	59.59	47.51
			42.7	65.41	53.23	50.30
			55.0	51.91	39.64	23.01
5	2.000	0.5	0.0	40.74	-	-
			12.3	49.15		
			24.2	36.70		
			35.7	32.43		
			46.5	10.12		
			56.5	NA		

by the Institutional Review Board at Stanford University. Before each experiment, each subject was individually briefed about the goal of the experiment and viewed an example of each image unperturbed to orient them towards the maximum image quality achievable with our custom VR display. An example of an image with a small flickering Gabor wavelet in the periphery was also shown to demonstrate to the user that they should remain looking at the fixation cross for the duration of the trial, even if such artifacts appear, and only observe the full FOV to the extent possible with their peripheral vision.

A.4.3 Procedure

After the briefing, each subject was instructed to position their chin on the headrest as in the previous user study. The start of each trial began with the subject being shown a small (1°) white cross for 1.5 s to indicate where they should fixate, after which the image would appear underneath for 3 seconds. The subject then selected a categorical ranking from 1 (“bad“), 2 (“poor”), 3 (“fair“),

Table A.2: Gabor wavelet parameters used to perturb the image in groups 1–3 of the validation study.

Group	σ (°)	e (°)	f_s (cpd)	f_t (Hz)	Predicted CFF (Hz)
1	0.8	30.10	1.000	35 45 59 73 83	58.98
2	0.8	18.80	1.600 1.330 0.800 0.530	58	47.69 53.13 65.03 72.05
3	1.875	-27.46 -4.90 0.00 4.90 27.46	0.267	75	81.67 70.79 65.77 70.97 81.67

Table A.3: Gabor wavelet parameters used to perturb the image in groups 4 and 5 of the validation study. Unlike groups 1–3, each row within a group is cumulative, and those rows with multiple Gabors, the eccentricities are given in terms of a magnitude, the angular position of the first and the separation angle between them.

Group	No. Gabors	σ (°)	e (°)	f_s (cpd)	f_t (Hz)	Predicted CFF (Hz)
4	2 8 16	1.875	$30.1, [\angle 0°, \Delta 180°]$ $30.1, [\angle 0°, \Delta 45°]$ $30.1, [\angle 0°, \Delta 30°]$	1.000	75	58.98
5	4	2.5	$24.9, [\angle 0°, \Delta 90°]$	0.800	61 72	65.55
	8	5	$30.0, [\angle 45°, \Delta 90°]$	0.267	86	81.50
	9	20	0.0	0.100	82	75.17

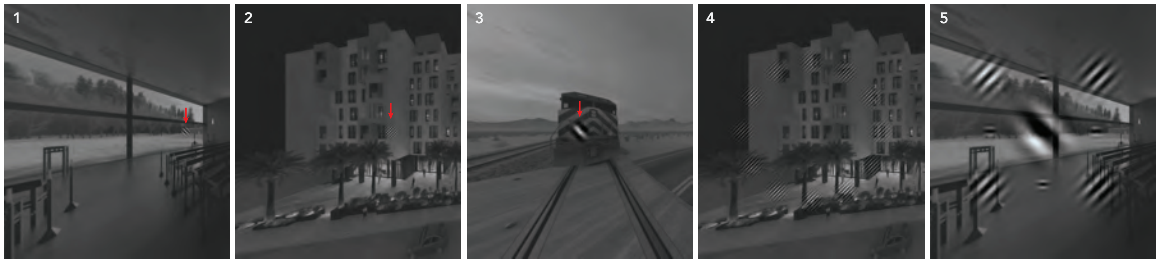

Figure A.2: A single frame from each of the validation study stimuli groups. A red arrow has been added to groups 1-3 to highlight the position of the single Gabor wavelet. Groups 4 and 5 show the final condition where all Gabor stimuli are present.

4 (“good”) or 5 (“excellent”) using a provided keyboard. The subject was allowed to take as much time as needed to enter the score, and replay the video up to one time. However, the subject could not change the score once entered, after which the next video was displayed. Please see the main text for the results.

Appendix B

Additional Material for Attention-aware Rendering

B.1 Cortical Magnification

In Section 4.1 of the main text, we describe using the *cortical magnification factor* to scale the spatial frequencies and diameters at the other retinal positions such that the discrimination thresholds should be approximately the same. It has been argued that visual performance depends importantly on the amount of cortical tissue devoted to the task. Thus, changes in detection and discrimination thresholds across eccentricity can be explained by the concept of *cortical magnification* [62,201]. This model describes how many neurons in the visual cortex are responsible for processing a particular part of the visual field. The central, foveal, region is processed by many more neurons (per degree of visual angle) than the periphery (illustrated in Figure B.1a).

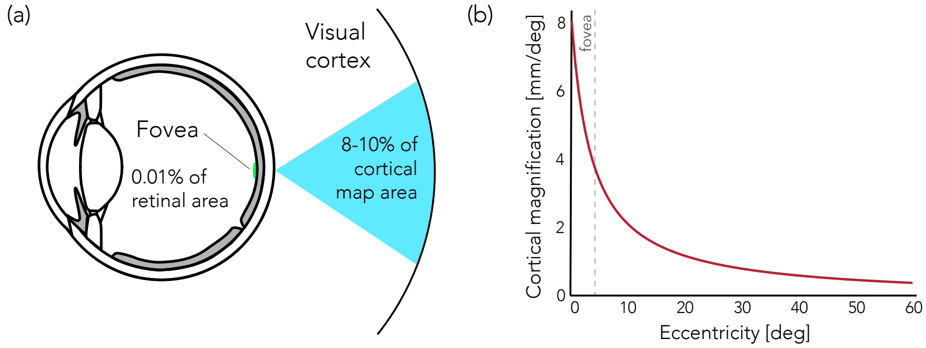

Figure B.1: Illustration of cortical magnification. (a) The cortical magnification maps the small area of the fovea to a much larger area on the visual cortex. (b) The magnitude of cortical magnification according to Dougherty et al. [230].

For quantitative purposes, the cortical magnification factor is normally expressed in millimeters of cortical surface per degree of visual angle. Several models have been proposed in the literature, but for this work we use the model by Dougherty et al. [230] (shown in Figure B.1b) which was fitted to fMRI measurements of V1. It is modeled as:

$$M(e) = \frac{a_0}{e + e_2} \tag{B.1}$$

where e is eccentricity in visual degrees and the fitted parameters are $a_0 = 29.2\,\text{mm}$ and $e_2 = 3.67°$. It can be seen that the magnification factor M is largest for those areas corresponding to the fovea and decreases with eccentricity for peripheral areas.

Virsu and Rovamo [62, 201] showed that the differences in detection of sinusoidal patterns and also discrimination of their orientation or direction of movement, can be compensated by increasing the size of the stimuli in the peripheral vision and the size increase is consistent with the inverse of cortical magnification. This agrees with the contrast thresholds we measure for the "low" foveal attention condition in Section 4.1.

B.2 Study Results

In Section 4.1 of the main text, we measure contrast thresholds under "low" , "medium" and "high" foveal attention conditions. Here, we provide additional details for the results of our experiments. Figure B.2 shows trends from the main study plotted for individual users (averaged across the two repetitions).

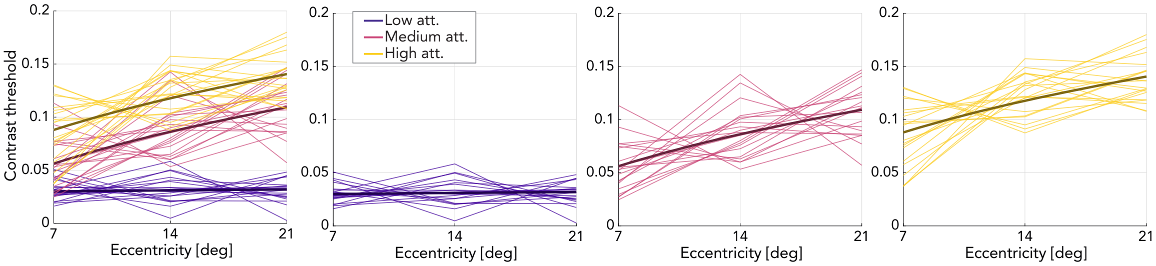

Figure B.2: Contrast thresholds measured for individual subjects (thin lines) in our main study that were used to fit our model (thick lines). For clarity, the attention levels are plotted together in the first panel and separately in the other panels. Mean thresholds for each plot line were re-scaled to match the respective global attention level means in order to remove subject-specific variation of the base sensitivity and highlight the variation among attention levels and eccentricities.

In Table B.1 we list these contrast thresholds average across the 10 and 6 participants in each of the main (No. 1-3) and validation (No.4-5) studies, respectively.

Table B.1: Mean measured contrast thresholds for the stimuli in our main and validation studies for different attention conditions (Section 4.1).

No.	**Contrast threshold**		
	Low	**Medium**	**High**
1	0.0297	0.0561	0.0851
2	0.0317	0.0864	0.1242
3	0.0314	0.1091	0.1368
4	0.0325	0.0607	0.0905
5	0.0452	0.1304	0.1806
6	0.0573	0.1415	0.1926
7	0.0508	0.1059	0.1832

In Section 4.2 of the main text we calibrate the perceptible foveation intensity for each of the "low" , "medium" and "high" foveal attention conditions. In Table B.2 we list the MAR slopes averaged across the 15 participants for each of the two test images ("Tulips" and "City").

Table B.2: Mean measured MAR slopes from the foveation study for different attention conditions (Section 4.2 of the main paper).

Image	**MAR slope**		
	Low	**Medium**	**High**
Tulips	0.0222	0.0499	0.0651
City	0.0153	0.0449	0.0623
Mountain	0.0221	0.0369	0.0581
Forest	0.0197	0.0361	0.0531

Appendix C

Additional Material for Gaze-contingent Stereo Rendering

C.1 Modeling the Human Eye

In Section 6.1 of the main text we refer to six cardinal points and two axes of the human eye, however additional definitions are prominent in the literature and for completeness, are explained here.

The lack of symmetry within the the human eye means that several different axes are required to fully describe its optical properties [231]. The direction of these axes are often defined relative to several cardinal points, as shown in Figure C.1. As we describe in the main text, the *optical axis* connects the anterior vertex of the cornea (V) and the center of rotation (C). When they eye is looking at a fixation point (P), the *visual axis* connects the point of fixation with the front nodal point (N), which extends through the rear nodal point (N'), to intersect with the fovea (F). This axis is not a straight line, since the nodal points are not coincident. Atchinson [217] also describes two additional axes defined in terms of the entrance pupil of the eye (E), i.e., the optical image of the pupil aperture as seen from the front of the eye. The first, the *line-of-sight* axis, connects the center of the entrance pupil to the fixation point, while the *pupillary axis* connects it to the point perpendicular to the surface of the cornea. If the eye was a centered system and the pupil was also centered, the pupillary axis would lie along the optical axis, however, this is rarely the case.

The directions of these axes are also commonly described relative to each other, in terms of the angles between them. As we describe in the main text, the angle between the optical and visual axes is most often referred to as α. However also measured is λ and κ, the pupillary axis to line-of-sight and visual axis angles, respectively. Since the pupillary and line-of-sight axes are defined with respect to the center of the pupil, which changes with pupil diameter [232], the direction of these axes and λ and κ vary with viewing conditions.

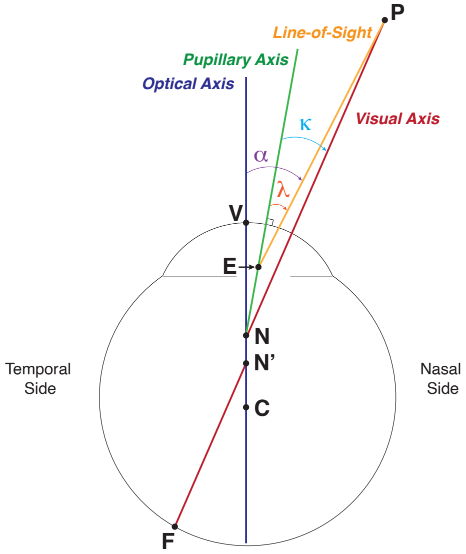

Figure C.1: Schematic illustrating the axes and angles referred to in the literature. The fixation point has been shown extremely close to the eye, thus exaggerating the angular differences between the visual, line-of-sight and fixation axes. C: center of rotation, E: entrance pupil center, F: fovea, N: front nodal point, N': rear nodal point, P: fixation point, V: anterior vertex of the cornea.

C.2 Psychophysical Experiment Equation Derivation

In Section 6.1 of the main text, we present and use an equation first derived by Bingham [219] to calculate the distance between the no-parallax point and the center of rotation, NC, using the target distances L_1 and L_2, the rotation of the eye θ and the largest extent that a user can detect E:

$$NC = \frac{EL_1}{(L_2 - L_1)\sin\theta + E\cos\theta} \tag{C.1}$$

As illustrated in Figure C.2, this equation can be derived from the geometry of the experiment. Namely, draw a line from the no-parallax point N to intersect the rear surface at 90° and call this point X. Then call the points at the left-most edge of the front surface and center of the rear surface W and Y, respectively, and the point at left-most edge of E, Z.

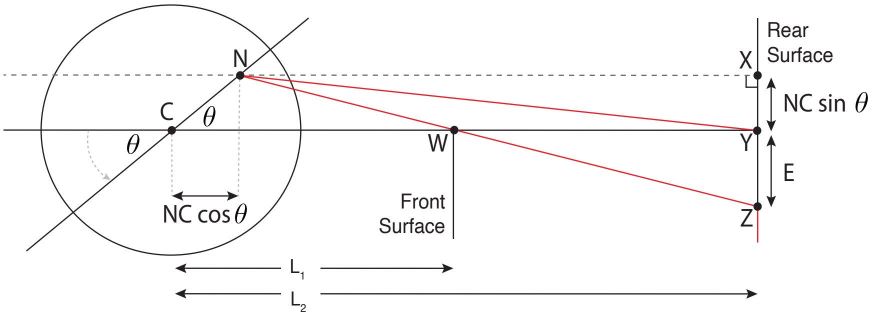

Figure C.2: Schematic illustrating the derivation of the equation used to calculate the distance between the no-parallax point and the center of rotation. Shown from above, front and rear surfaces are set up at distances L_1 and L_2, respectively. As the eye rotates counterclockwise about its center C through an angle, θ , the no-parallax point N is translated to the left revealing an extent, E, along the rear surface. W is the point at the leftmost edge of the front surface. Y is the point at the center of the rear surface. Z is the point at the left-edge of E. X is the point at which a ray drawn from N intersects the rear surface at 90°.

Then $\triangle NXZ$ and $\triangle WYZ$ are similar and thus we have:

$$\frac{E}{E + NC\sin\theta} = \frac{L_1 - L_2}{L_2 - NC\cos\theta} \tag{C.2}$$

which can be re-arranged to give Equation C.1.

C.3 Random Dot Stereograms

In Sections 6.2 and 6.3 of the main text, we describe the use of random dot stereograms (RDS) for rendering, and measuring disparity and depth perception in VR. Here we provide more description on how they work. An RDS stimuli consists of a pair of related random dot patterns. Each image seen separately appears to be a random collection of black and white dots. Yet, when the two images are presented separately to the two eyes, the relationship between the two collections of dots is detected by the visual pathways and the observer can perceive the surface of the object in depth [6].

Figure C.3 shows how to create the two images comprising a RDS. First, create a sampling grid and randomly assign a black or white intensity to each position in the grid. This random image of black and white dots will be one image in the stereo pair. Next, select a region of the first image. Displace this region horizontally, over-writing the original dots. Displacing this region of dots leaves some unspecified positions; fill in these unspecified positions randomly with new black and white dots.

RDS are a fascinating tool for vision science because the patterns we see in these stereo pairs are computed by signals that are carried separately by the two eyes. First, they demonstrate the simple but important point that even though we cannot see any monocular edge or surface information of the object, we can still see the object based on the disparity cue [233].

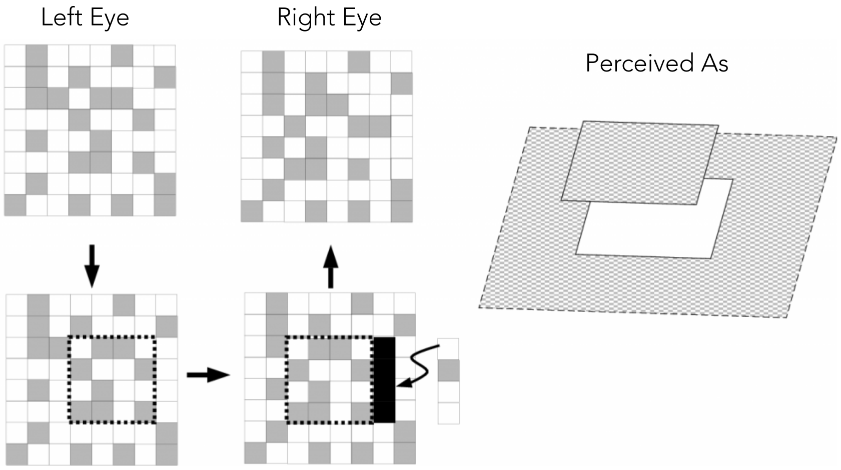

Figure C.3: Construction of a random dot stereogram. First, a random dot pattern is created to present to, say, the left eye. The stimulus for the right eye is created by copying the first image, displacing a region horizontally, and then filling the gap with a random sample of dots. When the right and left images are viewed simultaneously, the shifted region appears to be in a different depth plane from the other dots. Image adapted from Wandell [233].

C.4 Psychometric Functions

We present the full set of psychometric functions from our no-parallax point measurement (Section 6.1) and detection threshold (Section 6.2) experiments in Figures C.4 and C.5. The *psignifit* Python package [220], was used for fitting psychometric functions to the data using Bayesian inference.

C.4.1 No-Parallax Point Position

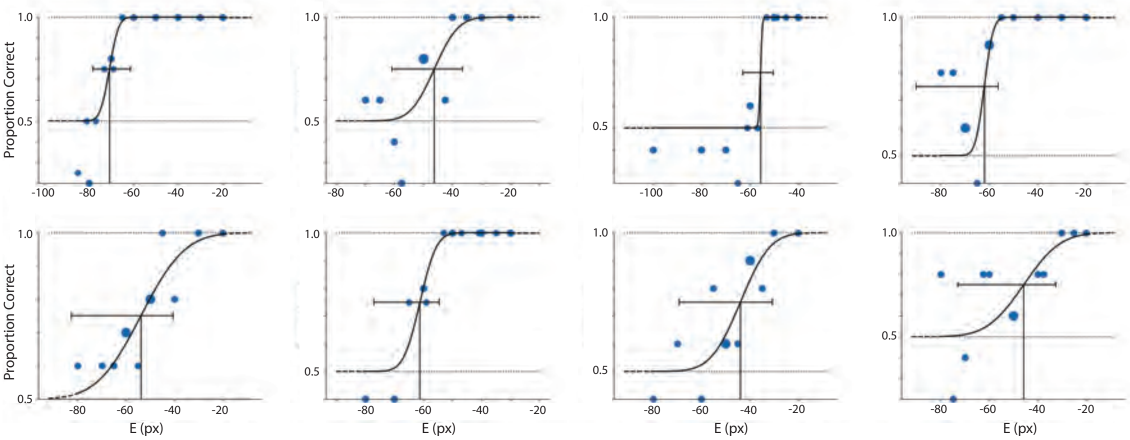

Figure C.4: Psychometric functions from the no-parallax point measurement experiment. Each plot represents the fit for a single subject. The vertical lines intersecting the psychometric fit indicate the threshold corresponding to the 75% correct response rate, and the horizontal line abutting it represents the 95% confidence interval. More negative E values correspond to subjects having a larger NC distance.

C.4.2 Detection Threshold in VR

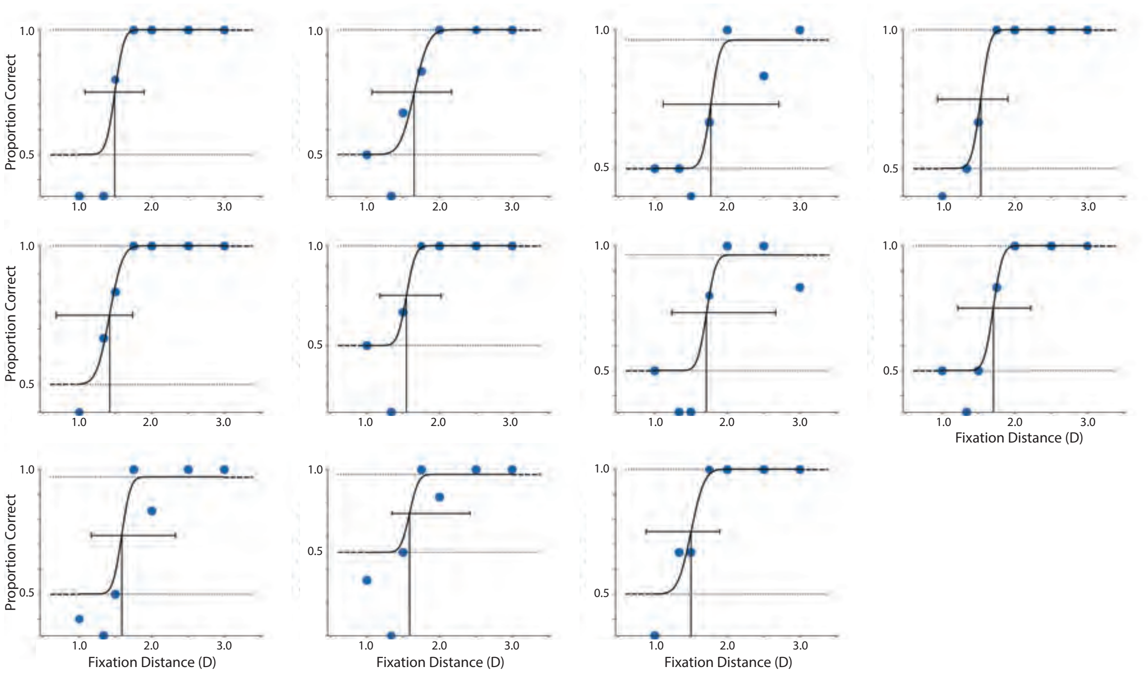

Figure C.5: Psychometric functions from the detection threshold experiment. Each plot represents the fit for a single subject. The vertical lines intersecting the psychometric fit indicate the threshold corresponding to the 75% correct response rate, and the horizontal line abutting it represents the 95% confidence interval. Lower fixation distances (in D) correspond to subjects being more sensitive to the shift in depth caused by gaze-contingent rendering.

Bibliography

[1] Lauren Mason and Donna Chrobot-Mason. Immersive inclusion: Diversity and inclusion training using virtual reality. In *Dismantling Bias Conference Series*, volume 3, page 4.

[2] Anthony G Gallagher, E Matt Ritter, Howard Champion, Gerald Higgins, Marvin P Fried, Gerald Moses, C Daniel Smith, and Richard M Satava. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. *Annals of surgery*, 241(2):364, 2005.

[3] Georg Eggers, Tobias Salb, Harald Hoppe, Luder Kahrs, Sassan Ghanai, Gunther Sudra, Jorg Raczkowsky, Rudiger Dillmann, Heinz Wörn, Stefan Hassfeld, et al. Intraoperative augmented reality: the surgeons view. *Studies in Health Technology and Informatics*, 111:123–125, 2005.

[4] Florian Heinrich, Gerd Schmidt, Florian Jungmann, and Christian Hansen. Augmented reality visualisation concepts to support intraoperative distance estimation. In *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology*, pages 1–2, 2019.

[5] Gordon Wetzstein and Douglas Lanman. Factored displays: improving resolution, dynamic range, color reproduction, and light field characteristics with advanced signal processing. *IEEE Signal Processing Magazine*, 33(5):119–129, 2016.

[6] Stephen E Palmer. *Vision science: Photons to phenomenology.* MIT press, 1999.

[7] Wilson S Geisler and Jeffrey S Perry. Real-time foveated multiresolution system for low-bandwidth video communication. In *Human vision and electronic imaging III*, volume 3299, pages 294–305. SPIE, 1998.

[8] J. Gordon Robson and Norma Graham. Probability summation and regional variation in contrast sensitivity across the visual field. *Vision research*, 21(3):409–418, 1981.

[9] Pavneet Bhatia, Arun Khosla, and Gajendra Singh. A review on eye tracking technology. *Interdisciplinary Approaches to Altering Neurodevelopmental Disorders*, pages 107–130, 2020.

[10] Andrew T Duchowski, Nathan Cournia, and Hunter Murphy. Gaze-contingent displays: A review. *Cyberpsychology & behavior*, 7(6):621–634, 2004.

[11] George Alex Koulieris, Kaan Akşit, Michael Stengel, Rafał K Mantiuk, Katerina Mania, and Christian Richardt. Near-eye display and tracking technologies for virtual and augmented reality. In *Computer Graphics Forum*, volume 38, pages 493–519. Wiley Online Library, 2019.

[12] Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. Foveated 3d graphics. *ACM transactions on Graphics (tOG)*, 31(6):1–10, 2012.

[13] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics (TOG)*, 35(6):1–12, 2016.

[14] Sanghoon Lee, Marios S Pattichis, and Alan C Bovik. Foveated video compression with optimal rate control. *IEEE Transactions on Image Processing*, 10(7):977–992, 2001.

[15] Gazi Karam Illahi, Thomas Van Gemert, Matti Siekkinen, Enrico Masala, Antti Oulasvirta, and Antti Ylä-Jääski. Cloud gaming with foveated video encoding. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(1):1–24, 2020.

[16] Jannick P Rolland, Akitoshi Yoshida, Larry D Davis, and John H Reif. High-resolution inset head-mounted display. *Applied optics*, 37(19):4183–4193, 1998.

[17] Jonghyun Kim, Youngmo Jeong, Michael Stengel, Kaan Akşit, Rachel Albert, Ben Boudaoud, Trey Greer, Joohwan Kim, Ward Lopes, Zander Majercik, et al. Foveated ar: dynamically-foveated augmented reality display. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019.

[18] Robert Konrad, Anastasios Angelopoulos, and Gordon Wetzstein. Gaze-contingent ocular parallax rendering for virtual reality. *ACM Transactions on Graphics (TOG)*, 39(2):1–12, 2020.

[19] Jonathan Martschinke, Jana Martschinke, Marc Stamminger, and Frank Bauer. Gaze-dependent distortion correction for thick lenses in hmds. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 1848–1851. IEEE, 2019.

[20] Ungsoo Samuel Kim, Omar A Mahroo, John D Mollon, and Patrick Yu-Wai-Man. Retinal ganglion cells—diversity of cell types and clinical relevance. *Frontiers in neurology*, 12:661938, 2021.

[21] Roy Taylor and Deborah Batey. *Handbook of retinal screening in diabetes: diagnosis and management*. John Wiley & Sons, 2012.

[22] Kristin Koch, Judith McLean, Ronen Segev, Michael A Freed, Michael J Berry II, Vijay Balasubramanian, and Peter Sterling. How much the eye tells the brain. *Current biology*, 16(14):1428–1434, 2006.

[23] Lester C Loschky, Sebastien Szaffarczyk, Clement Beugnet, Michael E Young, and Muriel Boucart. The contributions of central and peripheral vision to scene-gist recognition with a 180 visual field. *Journal of Vision*, 19(5):15–15, 2019.

[24] Roger H.S. Carpenter. *Movements of the Eyes, 2nd Rev.* Pion Limited, 1988.

[25] John Ross, M Concetta Morrone, Michael E Goldberg, and David C Burr. Changes in visual perception at the time of saccades. *Trends in neurosciences*, 24(2):113–121, 2001.

[26] Y Shin, HW Lim, MH Kang, M Seong, H Cho, and JH Kim. Normal range of eye movement and its relationship to age. *Acta Ophthalmologica*, 94, 2016.

[27] Do A Robinson. The mechanics of human smooth pursuit eye movement. *The Journal of Physiology*, 180(3):569, 1965.

[28] Eileen Kowler. Eye movements: The past 25 years. *Vision research*, 51(13):1457–1483, 2011.

[29] Theodore C Ruch and John F Fulton. Medical physiology and biophysics. *Academic Medicine*, 35(11):1067, 1960.

[30] Arnold Knapp. An introduction to clinical perimetry. *Archives of Ophthalmology*, 20(6):1116–1117, 1938.

[31] Xiaoxu Meng, Ruofei Du, Matthias Zwicker, and Amitabh Varshney. Kernel foveated rendering. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 1(1):1–20, 2018.

[32] Gerald Westheimer. Visual acuity. *Annual Review of Psychology*, 16(1):359–380, 1965.

[33] Michael F Deering. The limits of human vision. In *2nd International Immersive Projection Technology Workshop*, volume 2, page 1, 1998.

[34] Yiyi Wang, Nicolas Bensaid, Pavan Tiruveedhula, Jianqiang Ma, Sowmya Ravikumar, and Austin Roorda. Human foveal cone photoreceptor topography and its dependence on eye length. *Elife*, 8:e47148, 2019.

[35] LN Thibos, DJ Walsh, and FE Cheney. Vision beyond the resolution limit: aliasing in the periphery. *Vision Research*, 27(12):2193–2197, 1987.

[36] Bipul Mohanto, ABM Tariqul Islam, Enrico Gobbetti, and Oliver Staadt. An integrative view of foveated rendering. *Computers & Graphics*, 102:474–501, 2022.

[37] David Luebke, Benjamin Hallen, Dale Newfield, and Benjamin Watson. Perceptually driven simplification using gaze-directed rendering. Technical report, Tech. Rep. CS-2000-04, Department of Computer Science, University of . . . , 2000.

[38] Hunter A Murphy and Andrew T Duchowski. Gaze-contingent level of detail rendering. In *Eurographics (short presentations)*, 2001.

[39] Christine A Curcio and Kimberly A Allen. Topography of ganglion cells in human retina. *Journal of comparative Neurology*, 300(1):5–25, 1990.

[40] Jie Shen, Christopher A Clark, P Sarita Soni, and Larry N Thibos. Peripheral refraction with and without contact lens correction. *Optometry and vision science: official publication of the American Academy of Optometry*, 87(9):642, 2010.

[41] MH Pirenne. Rods and cones. In *The Visual Process*, pages 13–29. Elsevier, 1962.

[42] Yong He, Yan Gu, and Kayvon Fatahalian. Extending the graphics pipeline with adaptive, multi-rate shading. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014.

[43] Michael Stengel, Steve Grogorick, Martin Eisemann, and Marcus Magnor. Adaptive image-space sampling for gaze-contingent real-time rendering. In *Computer Graphics Forum*, volume 35, pages 129–139. Wiley Online Library, 2016.

[44] Karthik Vaidyanathan, Marco Salvi, Robert Toth, Theresa Foley, Tomas Akenine-Möller, Jim Nilsson, Jacob Munkberg, Jon Hasselgren, Masamichi Sugihara, Petrik Clarberg, et al. Coarse pixel shading. In *Proceedings of High Performance Graphics*, pages 9–18. 2014.

[45] Frank W Weymouth. Visual sensory units and the minimal angle of resolution. *American journal of ophthalmology*, 1958.

[46] Amritha Stalin and Kristine Dalton. Relationship of contrast sensitivity measured using quick contrast sensitivity function with other visual functions in a low vision population. *Investigative ophthalmology & visual science*, 61(6):21–21, 2020.

[47] Otto H Schade. Optical and photoelectric analog of the eye. *JoSA*, 46(9):721–739, 1956.

[48] Fergus W Campbell and John G Robson. Application of fourier analysis to the visibility of gratings. *The Journal of physiology*, 197(3):551, 1968.

[49] Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14, 2011.

[50] Scott J Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III*, volume 1666, pages 2–15. SPIE, 1992.

[51] Pontus Andersson, Jim Nilsson, Tomas Akenine-Möller, Magnus Oskarsson, Kalle Åström, and Mark D Fairchild. Flip: A difference evaluator for alternating images. *Proc. ACM Comput. Graph. Interact. Tech.*, 3(2):15–1, 2020.

[52] Albert J Ahumada Jr and Heidi A Peterson. Luminance-model-based dct quantization for color image compression. In *Human vision, visual processing, and digital display III*, volume 1666, pages 365–374. SPIE, 1992.

[53] Wenjun Zeng, Scott Daly, and Shawmin Lei. An overview of the visual optimization tools in jpeg 2000. *Signal Processing: Image Communication*, 17(1):85–104, 2002.

[54] Rafał Mantiuk, Scott Daly, and Louis Kerofsky. Display adaptive tone mapping. In *ACM SIGGRAPH 2008 papers*, pages 1–10. 2008.

[55] Okan Tarhan Tursun, Elena Arabadzhiyska-Koleva, Marek Wernikowski, Radosław Mantiuk, Hans-Peter Seidel, Karol Myszkowski, and Piotr Didyk. Luminance-contrast-aware foveated rendering. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

[56] Rafał K. Mantiuk, Maliha Ashraf, and Alexandre Chapiro. Stelacsf: A unified model of contrast sensitivity as the function of spatio-temporal frequency, eccentricity, luminance and area. *ACM Transactions on Graphics (TOG)*, 41(4), 2022.

[57] JG Robson. Contrast sensitivity: One hundred years of clinical measurement in proceedings of the retina research foundation symposia, vol. 5 eds r. shapley, d. man-kit lam, 1993.

[58] Juvi Mustonen, Jyrki Rovamo, and Risto Näsänen. The effects of grating area and spatial frequency on contrast sensitivity as a function of light level. *Vision research*, 33(15):2065–2072, 1993.

[59] Shinyoung Yi, Daniel S Jeon, Ana Serrano, Se-Yoon Jeong, Hui-Yong Kim, Diego Gutierrez, and Min H Kim. Modelling surround-aware contrast sensitivity for hdr displays. In *Computer Graphics Forum*, volume 41, pages 350–363. Wiley Online Library, 2022.

[60] Peter GJ Barten. Formula for the contrast sensitivity of the human eye. In *Image Quality and System Performance*, volume 5294, pages 231–238. SPIE, 2003.

[61] Jyrki Rovamo, Olavi Luntinen, and Risto Näsänen. Modelling the dependence of contrast sensitivity on grating area and spatial frequency. *Vision Research*, 33(18):2773–2788, 1993.

[62] V Virsu and J Rovamo. Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental brain research*, 37(3):475–494, 1979.

[63] Sophie Wuerger, Maliha Ashraf, Minjung Kim, Jasna Martinovic, María Pérez-Ortiz, and Rafał K Mantiuk. Spatio-chromatic contrast sensitivity under mesopic and photopic light levels. *Journal of Vision*, 20(4):23–23, 2020.

[64] Rafał K Mantiuk, Minjung Kim, Maliha Ashraf, Qiang Xu, M Ronnier Luo, Jasna Martinovic, and Sophie Wuerger. Practical color contrast sensitivity functions for luminance levels up to 10000 cd/m2. In *Color and Imaging Conference*, volume 2020, pages 1–6. Society for Imaging Science and Technology, 2020.

[65] John G Robson. Spatial and temporal contrast-sensitivity functions of the visual system. *Josa*, 56(8):1141–1142, 1966.

[66] Veijo Virsu, Jyrki Rovamo, Pentti Laurinen, and Risto Näsänen. Temporal contrast sensitivity and cortical magnification. *Vision Research*, 22(9):1211–1217, 1982.

[67] Donald H Kelly. Motion and vision. ii. stabilized spatio-temporal threshold surface. *Josa*, 69(10):1340–1349, 1979.

[68] Scott J. Daly. Engineering observations from spatiovelocity and spatiotemporal visual models. In Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors, *Human Vision and Electronic Imaging III*, volume 3299, pages 180 – 191. International Society for Optics and Photonics, SPIE, 1998.

[69] Andrew B Watson and Albert J Ahumada. The pyramid of visibility. *Electronic Imaging*, 2016(16):1–6, 2016.

[70] Andrew B Watson. The field of view, the field of resolution, and the field of contrast sensitivity. *Journal of Perceptual Imaging*, 1(1):10505–1, 2018.

[71] Brooke Krajancich, Petr Kellnhofer, and Gordon Wetzstein. A perceptual model for eccentricity-dependent spatio-temporal flicker fusion and its applications to foveated graphics. *ACM Transactions on Graphics (TOG)*, 40(4):1–11, 2021.

[72] Elena Arabadzhiyska, Cara Tursun, Hans-Peter Seidel, and Piotr Didyk. Practical saccade prediction for head-mounted displays: Towards a comprehensive model. *ACM Transactions on Applied Perceptions*, 20(1):1–23, 2023.

[73] Rafał K Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)*, 40(4):1–19, 2021.

[74] David Carmel, Nilli Lavie, and Geraint Rees. Conscious awareness of flicker in humans involves frontal and parietal cortex. *Current biology*, 16(9):907–911, 2006.

[75] Denis G Pelli and Peter Bex. Measuring contrast sensitivity. *Vision research*, 90:10–14, 2013.

[76] E Hartmann, B Lachenmayr, and H Brettel. The peripheral critical flicker frequency. *Vision Research*, 19(9):1019–1023, 1979.

[77] Jyrki Rovamo and Antti Raninen. Critical flicker frequency and m-scaling of stimulus size and retinal illuminance. *Vision research*, 24(10):1127–1131, 1984.

[78] Christopher W Tyler. Analysis of visual modulation sensitivity. iii. meridional variations in peripheral flicker sensitivity. *JOSA A*, 4(8):1612–1619, 1987.

[79] Raunak Sinha, Mrinalini Hoon, Jacob Baudin, Haruhisa Okawa, Rachel OL Wong, and Fred Rieke. Cellular and circuit mechanisms shaping the perceptual properties of the primate fovea. *Cell*, 168(3):413–426, 2017.

[80] H de Lange Dzn. Research into the dynamic nature of the human fovea→cortex systems with intermittent and modulated light. i. attenuation characteristics with white and colored light. *Josa*, 48(11):777–784, 1958.

[81] Floris L Van Nes and Maarten A Bouman. Spatial modulation transfer in the human eye. *JOSA*, 57(3):401–406, 1967.

[82] Auria Eisen-Enosh, Nairouz Farah, Zvia Burgansky-Eliash, Uri Polat, and Yossi Mandel. Evaluation of critical flicker-fusion frequency measurement methods for the investigation of visual temporal resolution. *Scientific reports*, 7(1):1–9, 2017.

[83] Jan J Koenderink, Maarten A Bouman, Albert E Bueno de Mesquita, and Sybe Slappendel. Perimetry of contrast detection thresholds of moving spatial sine wave patterns. iv. the influence of the mean retinal illuminance. *JOSA*, 68(6):860–865, 1978.

[84] Jan J Koenderink, Maarten A Bouman, Albert E Bueno de Mesquita, and Sybe Slappendel. Perimetry of contrast detection thresholds of moving spatial sine wave patterns. iii. the target extent as a sensitivity controlling parameter. *JOSA*, 68(6):854–860, 1978.

[85] James Davis, Yi-Hsuan Hsieh, and Hung-Chi Lee. Humans perceive flicker artifacts at 500 hz. *Scientific reports*, 5(1):7861, 2015.

[86] Christopher W Tyler and Russell D Hamer. Analysis of visual modulation sensitivity. iv. validity of the ferry–porter law. *JOSA A*, 7(4):743–758, 1990.

[87] Jyrki Rovamo and Antti Raninen. Critical flicker frequency as a function of stimulus area and luminance at various eccentricities in human cone vision: a revision of granit-harper and ferry-porter laws. *Vision research*, 28(7):785–790, 1988.

[88] Christopher W Tyler and Russell D Hamer. Eccentricity and the ferry–porter law. *JOSA A*, 10(9):2084–2087, 1993.

[89] Marisa Carrasco. Visual attention: The past 25 years. *Vision research*, 51(13):1484–1525, 2011.

[90] Ronald A Rensink. Change blindness: Implications for the nature of visual attention. *Vision and attention*, pages 169–188, 2001.

[91] Arien Mack and Irvin Rock. Inattentional blindness: Perception without attention. *Visual attention*, 8:55–76, 1998.

[92] Charles W Eriksen and James D St. James. Visual attention within and around the field of focal attention: A zoom lens model. *Perception & psychophysics*, 40(4):225–240, 1986.

[93] Edward Awh and Harold Pashler. Evidence for split attentional foci. *Journal of Experimental Psychology: Human Perception and Performance*, 26(2):834, 2000.

[94] Dale K Lee, Christof Koch, and Jochen Braun. Spatial vision thresholds in the near absence of attention. *Vision research*, 37(17):2409–2418, 1997.

[95] George Sperling and Melvin J Melchner. The attention operating characteristic: Examples from visual search. *Science*, 202(4365):315–318, 1978.

[96] Monireh Mahjoob and Andrew J Anderson. Contrast discrimination under task-induced mental load. *Vision Research*, 165:84–89, 2019.

[97] Liqiang Huang and Karen R Dobkins. Attentional effects on contrast discrimination in humans: evidence for both contrast gain and response gain. *Vision research*, 45(9):1201–1212, 2005.

[98] Monireh Mahjoob, Javad Heravian Shandiz, and Andrew J Anderson. The effect of mental load on psychophysical and visual evoked potential visual acuity. *Ophthalmic and physiological optics*, 42(3):586–593, 2022.

[99] Marisa Carrasco, Anna Marie Giordano, and Brian McElree. Attention speeds processing across eccentricity: Feature and conjunction searches. *Vision research*, 46(13):2028–2040, 2006.

[100] Katharina Anton-Erxleben and Marisa Carrasco. Attentional enhancement of spatial resolution: linking behavioural and neurophysiological evidence. *Nature Reviews Neuroscience*, 14(3):188–200, 2013.

[101] Marisa Carrasco. How visual spatial attention alters perception. *Cognitive processing*, 19(1):77–88, 2018.

[102] Marisa Carrasco, Cigdem Penpeci-Talgar, and Miguel Eckstein. Spatial covert attention increases contrast sensitivity across the csf: support for signal enhancement. *Vision research*, 40(10-12):1203–1215, 2000.

[103] E Leslie Cameron, Joanna C Tai, and Marisa Carrasco. Covert attention affects the psychometric function of contrast sensitivity. *Vision research*, 42(8):949–967, 2002.

[104] Maria Concetta Morrone, V Denti, and D Spinelli. Different attentional resources modulate the gain mechanisms for color and luminance contrast. *Vision research*, 44(12):1389–1401, 2004.

[105] Sam Ling and Marisa Carrasco. Sustained and transient covert attention enhance the signal via different contrast response functions. *Vision research*, 46(8-9):1210–1220, 2006.

[106] Barbara Montagna, Franco Pestilli, and Marisa Carrasco. Attention trades off spatial acuity. *Vision research*, 49(7):735–745, 2009.

[107] Clifton Schor, Ivan Wood, and Jane Ogawa. Binocular sensory fusion is limited by spatial resolution. *Vision research*, 24(7):661–665, 1984.

[108] James E Cutting and Peter M Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of space and motion*, pages 69–117. Elsevier, 1995.

[109] Dennis R Proffitt and Corrado Caudek. Depth perception and the perception of events. 2003.

[110] Suzanne P McKee. The spatial requirements for fine stereoacuity. *Vision research*, 23(2):191–198, 1983.

[111] Samuel C Rawlings and T Shipley. Stereoscopic acuity and horizontal angular distance from fixation. *JOSA*, 59(8):991–993, 1969.

[112] John Siderov and Ronald S Harwerth. Stereopsis, spatial frequency and retinal eccentricity. *Vision research*, 35(16):2329–2337, 1995.

[113] Andrew T Duchowski. A breadth-first survey of eye-tracking applications. *Behavior Research Methods Instruments and Computers*, 34(4):455–470, 2002.

[114] Isayas Berhe Adhanom, Paul MacNeilage, and Eelke Folmer. Eye tracking in virtual reality: a broad review of applications and challenges. *Virtual Reality*, pages 1–24, 2023.

[115] William Steptoe, Oyewole Oyekoya, Alessio Murgia, Robin Wolff, John Rae, Estefania Guimaraes, David Roberts, and Anthony Steed. Eye tracking for avatar eye gaze control during object-focused multiparty interaction in immersive collaborative virtual environments. In *2009 IEEE virtual reality conference*, pages 83–90. IEEE, 2009.

[116] Vrishab Krishna, Yi Ding, Aiwen Xu, and Tobias Höllerer. Multimodal biometric authentication for vr/ar using eeg and eye tracking. In *Adjunct of the 2019 International Conference on Multimodal Interaction*, pages 1–5, 2019.

[117] Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design. In *Proceedings of the 2017 Chi conference on human factors in computing systems*, pages 1118–1130, 2017.

[118] Jiahui Liu, Jiannan Chi, Huijie Yang, and Xucheng Yin. In the eye of the beholder: A survey of gaze tracking techniques. *Pattern Recognition*, page 108944, 2022.

[119] Niklas Stein, Diederick C Niehorster, Tamara Watson, Frank Steinicke, Katharina Rifai, Siegfried Wahl, and Markus Lappe. A comparison of eye tracking latencies among several commercial head-mounted displays. *i-Perception*, 12(1):2041669520983338, 2021.

[120] Kohei Miki, Takashi Nagamatsu, and Dan Witzner Hansen. Implicit user calibration for gaze-tracking systems using kernel density estimation. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 249–252, 2016.

[121] Mohsen Mansouryar, Julian Steil, Yusuke Sugano, and Andreas Bulling. 3d gaze estimation from 2d pupil positions on monocular head-mounted eye trackers. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 197–200, 2016.

[122] John D McCarthy, M Angela Sasse, and Dimitrios Miras. Sharp or smooth? comparing the effects of quantization vs. frame rate for streamed video. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 535–542, 2004.

[123] Budmonde Duinkharjav, Kenneth Chen, Abhishek Tyagi, Jiayi He, Yuhao Zhu, and Qi Sun. Color-perception-guided display power reduction for virtual reality. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022.

[124] Toshikazu Ohshima, Hiroyuki Yamamoto, and Hideyuki Tamura. Gaze-directed adaptive rendering for interacting with virtual space. In *Proc. IEEE VR*. IEEE, 1996.

[125] David Luebke and Benjamin Hallen. Perceptually driven simplification for interactive rendering. In *Rendering Techniques 2001: Proceedings of the Eurographics Workshop in London, United Kingdom, June 25–27, 2001 12*, pages 223–234. Springer, 2001.

[126] Radosław Mantiuk and Mateusz Markowski. Gaze-dependent tone mapping. In *Image Analysis and Recognition: 10th International Conference, ICIAR 2013, Póvoa do Varzim, Portugal, June 26-28, 2013. Proceedings 10*, pages 426–433. Springer, 2013.

[127] David E. Jacobs, Orazio Gallo, Emily A. Cooper, Kari Pulli, and Marc Levoy. Simulating the visual experience of very bright and very dark scenes. *ACM Transactions on Graphics (TOG)*, 34(3):1–15, 2015.

[128] Michael Mauderer, David R Flatla, and Miguel A Nacenta. Gaze-contingent manipulation of color perception. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5191–5202, 2016.

[129] Qi Sun, Fu-Chung Huang, Joohwan Kim, Li-Yi Wei, David Luebke, and Arie Kaufman. Perceptually-guided foveation for light field displays. *ACM Transactions on Graphics (TOG)*, 36(6):1–13, 2017.

[130] Nianchen Deng, Zhenyi He, Jiannan Ye, Budmonde Duinkharjav, Praneeth Chakravarthula, Xubo Yang, and Qi Sun. Fov-nerf: Foveated neural radiance fields for virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3854–3864, 2022.

[131] Anton S Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. Deepfovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019.

[132] Taimoor Tariq, Cara Tursun, and Piotr Didyk. Noise-based enhancement for foveated rendering. *ACM Transactions on Graphics (TOG)*, 41(4):1–14, 2022.

[133] Lili Wang, Xuehuai Shi, and Yi Liu. Foveated rendering: A state-of-the-art survey. *Computational Visual Media*, 9(2):195–228, 2023.

[134] Hector Yee, Sumanita Pattanaik, and Donald P Greenberg. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Transactions on Graphics (TOG)*, 20(1):39–65, 2001.

[135] Behnam Bastani, Eric Turner, Carlin Vieri, Haomiao Jiang, Brian Funt, and Nikhil Balram. Foveated pipeline for ar/vr head-mounted displays. *Information Display*, 33(6):14–35, 2017.

[136] Matias Koskela, Atro Lotvonen, Markku Mäkitalo, Petrus Kivi, Timo Viitanen, and Pekka Jääskeläinen. Foveated real-time path tracing in visual-polar space. In *Proceedings of 30th Eurographics Symposium on Rendering*. The Eurographics Association, 2019.

[137] Martin Weier, Thorsten Roth, André Hinkenjann, and Philipp Slusallek. Foveated depth-of-field filtering in head-mounted displays. *ACM Transactions on Applied Perception (TAP)*, 15(4):1–14, 2018.

[138] Xin Zhang, Wei Chen, Zhonglei Yang, Chuan Zhu, and Qunsheng Peng. A new foveation ray casting approach for real-time rendering of 3d scenes. In *2011 12th international conference on computer-aided design and computer graphics*, pages 99–102. IEEE, 2011.

[139] Adam Siekawa, Michał Chwesiuk, Radosław Mantiuk, and Rafał Piórkowski. Foveated ray tracing for vr headsets. In *MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part I 25*, pages 106–117. Springer, 2019.

[140] Andrew T Duchowski, David Bate, Paris Stringfellow, Kaveri Thakur, Brian J Melloy, and Anand K Gramopadhye. On spatiochromatic visual sensitivity and peripheral color lod management. *ACM Transactions on Applied Perception (TAP)*, 6(2):1–18, 2009.

[141] Lei Xiao, Salah Nouri, Matt Chapman, Alexander Fix, Douglas Lanman, and Anton Kaplanyan. Neural supersampling for real-time rendering. *ACM Transactions on Graphics (TOG)*, 39(4):142–1, 2020.

[142] Xiaoxu Meng, Ruofei Du, and Amitabh Varshney. Eye-dominance-guided foveated rendering. *IEEE transactions on visualization and computer graphics*, 26(5):1972–1980, 2020.

[143] Chanhyung Yoo, Jianghao Xiong, Seokil Moon, Dongheon Yoo, Chang-Kun Lee, Shin-Tson Wu, and Byoungho Lee. Foveated display system based on a doublet geometric phase lens. *Optics Express*, 28(16):23690–23702, 2020.

[144] Seungjae Lee, Mengfei Wang, Gang Li, Lu Lu, Yusufu Sulai, Changwon Jang, and Barry Silverstein. Foveated near-eye display for mixed reality using liquid crystal photonics. *Scientific Reports*, 10(1):16127, 2020.

[145] Guanjun Tan, Yun-Han Lee, Tao Zhan, Jilin Yang, Sheng Liu, Dongfeng Zhao, and Shin-Tson Wu. Foveated imaging for near-eye displays. *Optics express*, 26(19):25076–25085, 2018.

[146] Akitoshi Yoshida, Jannick P Rolland, and John H Reif. Optical design and analysis of a head-mounted display with a high-resolution insert. In *Novel Optical Systems Design and Optimization*, volume 2537, pages 71–82. SPIE, 1995.

[147] Yu Guan, Chengyuan Zheng, Xinggong Zhang, Zongming Guo, and Junchen Jiang. Pano: Optimizing 360 video streaming with a better understanding of quality perception. In *Proceedings of the ACM Special Interest Group on Data Communication*, pages 394–407. 2019.

[148] Liyang Sun, Yixiang Mao, Tongyu Zong, Yong Liu, and Yao Wang. Flocking-based live streaming of 360-degree video. In *Proceedings of the 11th ACM Multimedia Systems Conference*, pages 26–37, 2020.

[149] Jiawen Chen, Miao Hu, Zhenxiao Luo, Zelong Wang, and Di Wu. Sr360: boosting 360-degree video streaming with super-resolution. In *Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, pages 1–6, 2020.

[150] Miguel Fabian Romero-Rondón, Lucile Sassatelli, Frédéric Precioso, and Ramon Aparicio-Pardo. Foveated streaming of virtual reality videos. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 494–497, 2018.

[151] Mattis Jeppsson, Håvard N Espeland, Tomas Kupka, Ragnar Langseth, Andreas Petlund, Qiaoqiao Peng, Chuansong Xue, Dag Johansen, Konstantin Pogorelov, Håkon Stensland, et al. Efficient live and on-demand tiled hevc 360 vr video streaming. *International Journal of Semantic Computing*, 13(03):367–391, 2019.

[152] Marc Lambooij, Wijnand IJsselsteijn, Marten Fortuin, Ingrid Heynderickx, et al. Visual discomfort and visual fatigue of stereoscopic displays: a review. *Journal of imaging science and technology*, 53(3):30201–1, 2009.

[153] Frank L Kooi and Alexander Toet. Visual comfort of binocular and 3d displays. *Displays*, 25(2-3):99–108, 2004.

[154] Takashi Shibata, Joohwan Kim, David M Hoffman, and Martin S Banks. The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of vision*, 11(8):11–11, 2011.

[155] Eli Peli, T Reed Hedges, Jinshan Tang, and Dan Landmann. 53.2: A binocular stereoscopic display system with coupled convergence and accommodation demands. In *SID Symposium Digest of Technical Papers*, volume 32, pages 1296–1299. Wiley Online Library, 2001.

[156] Philippe Hanhart and Touradj Ebrahimi. Subjective evaluation of two stereoscopic imaging systems exploiting visual attention to improve 3d quality of experience. In *Stereoscopic Displays and Applications XXV*, volume 9011, pages 93–103. SPIE, 2014.

[157] Petr Kellnhofer, Piotr Didyk, Karol Myszkowski, Mohamed M Hefeeda, Hans-Peter Seidel, and Wojciech Matusik. Gazestereo3d: Seamless disparity manipulations. *ACM Transactions on Graphics (TOG)*, 35(4):1–13, 2016.

[158] Michael Mauderer, Simone Conte, Miguel A Nacenta, and Dhanraj Vishwanath. Depth perception with gaze-contingent depth of field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 217–226, 2014.

[159] Andrew T Duchowski, Donald H House, Jordan Gestring, Rui I Wang, Krzysztof Krejtz, Izabela Krejtz, Radosław Mantiuk, and Bartosz Bazyluk. Reducing visual discomfort of 3d stereoscopic displays with gaze-contingent depth-of-field. In *Proceedings of the acm symposium on applied perception*, pages 39–46, 2014.

[160] Margarita Vinnikov and Robert S Allison. Gaze-contingent depth of field in realistic scenes: The user experience. In *Proceedings of the symposium on eye tracking research and applications*, pages 119–126, 2014.

[161] Robert Konrad, Emily A Cooper, and Gordon Wetzstein. Novel optical configurations for virtual reality: Evaluating user preference and performance with focus-tunable and monovision near-eye displays. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 1211–1220, 2016.

[162] Nitish Padmanaban, Robert Konrad, Tal Stramer, Emily A Cooper, and Gordon Wetzstein. Optimizing virtual reality for all users through gaze-contingent and adaptive focus displays. *Proceedings of the National Academy of Sciences*, 114(9):2183–2188, 2017.

[163] David Dunn, Cary Tippets, Kent Torell, Petr Kellnhofer, Kaan Akşit, Piotr Didyk, Karol Myszkowski, David Luebke, and Henry Fuchs. Wide field of view varifocal near-eye display using see-through deformable membrane mirrors. *IEEE transactions on visualization and computer graphics*, 23(4):1322–1331, 2017.

[164] Kaan Akşit, Ward Lopes, Jonghyun Kim, Peter Shirley, and David Luebke. Near-eye varifocal augmented reality display using see-through screens. *ACM Transactions on Graphics (TOG)*, 36(6):1–13, 2017.

[165] Douglas Lanman and David Luebke. Near-eye light field displays. *ACM Transactions on Graphics (TOG)*, 32(6):1–10, 2013.

[166] Hong Hua and Bahram Javidi. A 3d integral imaging optical see-through head-mounted display. *Optics express*, 22(11):13484–13491, 2014.

[167] Fu-Chung Huang, Kevin Chen, and Gordon Wetzstein. The light field stereoscope: Immersive computer graphics via factored near-eye light field display with focus cues. *ACM Trans. Graph. (SIGGRAPH)*, 34(4), 2015.

[168] Andrew Maimone, Andreas Georgiou, and Joel S Kollin. Holographic near-eye displays for virtual and augmented reality. *ACM Transactions on Graphics (Tog)*, 36(4):1–16, 2017.

[169] Liang Shi, Fu-Chung Huang, Ward Lopes, Wojciech Matusik, and David Luebke. Near-eye light field holographic rendering with spherical waves for wide field of view interactive 3d computer graphics. *ACM Transactions on Graphics (TOG)*, 36(6):1–17, 2017.

[170] Nitish Padmanaban, Yifan Peng, and Gordon Wetzstein. Holographic near-eye displays based on overlap-add stereograms. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019.

[171] Jannick P Rolland, Myron W Krueger, and Alexei Goon. Multifocal planes head-mounted displays. *Applied Optics*, 39(19):3209–3215, 2000.

[172] Kurt Akeley, Simon J Watt, Ahna Reza Girshick, and Martin S Banks. A stereo display prototype with multiple focal distances. *ACM transactions on graphics (TOG)*, 23(3):804–813, 2004.

[173] Jen-Hao Rick Chang, BVK Vijaya Kumar, and Aswin C Sankaranarayanan. Towards multifocal displays with dense focal stacks. *ACM Transactions on Graphics (TOG)*, 37(6):1–13, 2018.

[174] Changwon Jang, Kiseung Bang, Seokil Moon, Jonghyun Kim, Seungjae Lee, and Byoungho Lee. Retinal 3d: augmented reality near-eye display via pupil-tracked light field projection on retina. *ACM Transactions on Graphics (TOG)*, 36(6):1–13, 2017.

[175] Jae-Hyeung Park and Seong-Bok Kim. Optical see-through holographic near-eye-display with eyebox steering and depth of field control. *Optics Express*, 26(21):27076–27088, 2018.

[176] Changwon Jang, Kiseung Bang, Gang Li, and Byoungho Lee. Holographic near-eye display with expanded eye-box. *ACM Transactions on Graphics (TOG)*, 37(6):1–14, 2018.

[177] Ying Geng, Jacques Gollier, Brian Wheelwright, Fenglin Peng, Yusufu Sulai, Brant Lewis, Ning Chan, Wai Sze Tiffany Lam, Alexander Fix, Douglas Lanman, et al. Viewing optics for immersive near-eye displays: pupil swim/size and weight/stray light. In *Digital Optics for Immersive Displays*, volume 10676, pages 19–35. SPIE, 2018.

[178] Matthias B Hullin, Johannes Hanika, and Wolfgang Heidrich. Polynomial optics: A construction kit for efficient ray-tracing of lens systems. In *Computer Graphics Forum*, volume 31, pages 1375–1383. Wiley Online Library, 2012.

[179] Phillip Guan, Olivier Mercier, Michael Shvartsman, and Douglas Lanman. Perceptual requirements for eye-tracked distortion correction in vr. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–8, 2022.

[180] Jonathan W Peirce. Psychopy–psychophysics software in python. *Journal of neuroscience methods*, 162(1-2):8–13, 2007.

[181] Dale Allen and Robert F Hess. Is the visual field temporally homogeneous? *Vision research*, 32(6):1075–1084, 1992.

[182] Thomas P Weldon, William E Higgins, and Dennis F Dunn. Efficient gabor filter design for texture segmentation. *Pattern recognition*, 29(12):2005–2015, 1996.

[183] J-K Kamarainen, Ville Kyrki, and Heikki Kalviainen. Invariance properties of gabor filter-based features-overview and applications. *IEEE Transactions on image processing*, 15(5):1088–1099, 2006.

[184] John G Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *JOSA A*, 2(7):1160–1169, 1985.

[185] S Marĉelja. Mathematical description of the responses of simple cortical cells. *JOSA*, 70(11):1297–1300, 1980.

[186] Manuel Romero-Gómez, Juan Córdoba, Rodrigo Jover, Juan A Del Olmo, Marta Ramírez, Ramón Rey, Enrique De Madaria, Carmina Montoliu, David Nuñez, Montse Flavia, et al. Value of the critical flicker frequency in patients with minimal hepatic encephalopathy. *Hepatology*, 45(4):879–885, 2007.

[187] Andrew B Watson and Denis G Pelli. Quest: A bayesian adaptive psychometric method. *Perception & psychophysics*, 33(2):113–120, 1983.

[188] Jan J Koenderink, Maarten A Bouman, Albert E Bueno de Mesquita, and Sybe Slappendel. Perimetry of contrast detection thresholds of moving spatial sine wave patterns. ii. the far peripheral visual field (eccentricity 0°–50°). *JOSA*, 68(6):850–854, 1978.

[189] Mark F Bradshaw and Brian J Rogers. Sensitivity to horizontal and vertical corrugations defined by binocular disparity. *Vision research*, 39(18):3049–3056, 1999.

[190] Arian Mehrfard, Javad Fotouhi, Giacomo Taylor, Tess Forster, Nassir Navab, and Bernhard Fuerst. A comparative analysis of virtual reality head-mounted display systems. *arXiv preprint arXiv:1912.02913*, 2019.

[191] Philip A Stanley and A Kelvin Davies. The effect of field of view size on steady-state pupil diameter. *Ophthalmic and Physiological Optics*, 15(6):601–603, 1995.

[192] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE transactions on Image Processing*, 19(6):1427–1441, 2010.

[193] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 6(2), 2016.

[194] Auria Eisen-Enosh, Nairouz Farah, Zvia Burgansky-Eliash, Idit Maharshak, Uri Polat, and Yossi Mandel. A dichoptic presentation device and a method for measuring binocular temporal function in the visual system. *Experimental Eye Research*, 201:108290, 2020.

[195] MR Ali and T Amir. Critical flicker frequency under monocular and binocular conditions. *Perceptual and motor skills*, 72(2):383–386, 1991.

[196] Ruth Rosenholtz. Capabilities and limitations of peripheral vision. *Annual Review of Vision Science*, 2:437–457, 2016.

[197] Gyorgy Denes and Rafal Mantiuk. Predicting visible flicker in temporally changing images. 2020.

[198] Arien Mack. Inattentional blindness: Looking without seeing. *Current directions in psychological science*, 12(5):180–184, 2003.

[199] MJ Wright and A Johnston. Spatiotemporal contrast sensitivity and visual field locus. *Vision research*, 23(10):983–989, 1983.

[200] Moritz Kassner, William Patera, and Andreas Bulling. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*, pages 1151–1160, 2014.

[201] Jyrki Rovamo and Veijo Virsu. An estimation and application of the human cortical magnification factor. *Experimental brain research*, 37(3):495–510, 1979.

[202] Camilla Funch Staugaard, Anders Petersen, and Signe Vangkilde. Eccentricity effects in vision and attention. *Neuropsychologia*, 92:69–78, 2016.

[203] Bert Hoeks and Willem JM Levelt. Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research methods, instruments, & computers*, 25(1):16–26, 1993.

[204] Sebastiaan Mathôt, Lotje Van der Linden, Jonathan Grainger, and Françoise Vitu. The pupillary light response reveals the focus of covert visual attention. *PloS one*, 8(10):e78168, 2013.

[205] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000.

[206] Qi Zhang, Zhibang Huang, Liang Li, and Sheng Li. Visual search training benefits from the integrative effect of enhanced covert attention and optimized overt eye movements. *bioRxiv*, 2022.

[207] Ripan Kumar Kundu, Akhlaqur Rahman, and Shuva Paul. A study on sensor system latency in vr motion sickness. *Journal of Sensor and Actuator Networks*, 10(3):53, 2021.

[208] Timothy Terriberry. Derf's test media collection.

[209] Robin Thunström. Passive gaze-contingent techniques relation to system latency, 2014.

[210] Rachel Albert, Anjul Patney, David Luebke, and Joohwan Kim. Latency requirements for foveated rendering in virtual reality. *ACM Transactions on Applied Perception (TAP)*, 14(4):1–13, 2017.

[211] Lester C Loschky and Gary S Wolverton. How late can you update gaze-contingent multiresolutional displays without detection? *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 3(4):1–10, 2007.

[212] Oliver Wiedemann, Vlad Hosu, Hanhe Lin, and Dietmar Saupe. Foveated video coding for real-time streaming applications. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2020.

[213] Anastasios N. Angelopoulos, Julien N.P. Martel, Amit P. Kohli, Jörg Conradt, and Gordon Wetzstein. Event-based near-eye gaze tracking beyond 10,000 hz. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2577–2586, 2021.

[214] Jean-Baptiste Bernard, Scherlen Anne-Catherine, and Castet Eric. Page mode reading with simulated scotomas: A modest effect of interline spacing on reading speed. *Vision research*, 47(28):3447–3459, 2007.

[215] Christopher J Bockisch and Joel M Miller. Different motor systems use similar damped extraretinal eye position information. *Vision research*, 39(5):1025–1038, 1999.

[216] Eyal M Reingold. Eye tracking research and technology: Towards objective measurement of data quality. *Visual cognition*, 22(3-4):635–652, 2014.

[217] David Atchison and George Smith. *Optics of the human eye.* Butterworth Heinemann, 2000.

[218] David A. Atchison. Schematic eyes. In Pablo Artal, editor, *Handbook of Visual Optics, Volume I - Fundamentals and Eye Optics*, chapter 16. CRC Press, 2017.

[219] Geoffrey P Bingham. Optical flow from eye movement with head immobilized:"ocular occlusion" beyond the nose. *Vision Research*, 33(5-6):777–789, 1993.

[220] Heiko H Schütt, Stefan Harmeling, Jakob H Macke, and Felix A Wichmann. Painfree and accurate bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, 122:105–123, 2016.

[221] Božo Vojniković and Ettore Tamajo. Horopters–definition and construction. *Collegium antropologicum*, 37(1):9–12, 2013.

[222] Jacek Turski. On binocular vision: The geometric horopter and cyclopean eye. *Vision research*, 119:73–81, 2016.

[223] Piotr Didyk, Tobias Ritschel, Elmar Eisemann, Karol Myszkowski, and Hans-Peter Seidel. A perceptual model for disparity. *ACM Transactions on Graphics (TOG)*, 30(4):1–10, 2011.

[224] Neil A. Dodgson. Variation and extrema of human interpupillary distance. In Mark T. Bolas, Andrew J. Woods, John O. Merritt, and Stephen A. Benton, editors, *Stereoscopic Displays and Virtual Reality Systems XI*, volume 5291, pages 36 – 46. International Society for Optics and Photonics, SPIE, 2004.

[225] Felix A Wichmann and N Jeremy Hill. The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, 63(8):1293–1313, 2001.

[226] Andrew Glennerster, Brian J Rogers, and Mark F Bradshaw. Stereoscopic depth constancy depends on the subject's task. *Vision research*, 36(21):3441–3456, 1996.

[227] Whitman Richards and John F Miller. Convergence as a cue to depth. *Perception & Psychophysics*, 5(5):317–320, 1969.

[228] Jacob Leon Kröger, Otto Hans-Martin Lutz, and Florian Müller. What does your gaze reveal about you? on the privacy implications of eye tracking. *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers 14*, pages 226–241, 2020.

[229] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4):1633–1642, 2018.

[230] Robert F Dougherty, Volker M Koch, Alyssa A Brewer, Bernd Fischer, Jan Modersitzki, and Brian A Wandell. Visual field representations and locations of visual areas v1/2/3 in human visual cortex. *Journal of vision*, 3(10):1–1, 2003.

[231] Sathish Srinivasan. Ocular axes and angles: Time for better understanding. *Journal of Cataract & Refractive Surgery*, 42(3):351–352, 2016.

[232] Imene Salah Mabed, Alain Saad, Emmanuel Guilbert, and Damien Gatinel. Measurement of pupil center shift in refractive surgery candidates with caucasian eyes using infrared pupillometry. *Journal of refractive surgery*, 30(10):694–700, 2014.

[233] Brian A Wandell. *Foundations of vision.* Sinauer Associates, 1995.