PERCEPTION-DRIVEN RENDERING

Techniques for the Efficient Visualization of 3D Scenes including View- and Gaze-contingent Approaches

MARTIN WEIER

A dissertation submitted towards the degree
Doctor of Engineering (Dr.-Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland Universtiy

Saarbrücken, 2019

Dean of the Faculty

Prof. Dr. Sebastian Hack
Saarland University, Saarbrücken, Germany

Committee Chair

Prof. Dr.-Ing. Jan Reineke
Saarland University, Saarbrücken, Germany

Reviewers

Prof. Dr.-Ing. Philipp Slusallek
Saarland University, Intel VCI, and DFKI, Saarbrücken, Germany

Prof. Dr. Karol Myszkowski
MPI Informatik, Saarbrücken, Germany

Prof. Dr. André Hinkenjann
Hochschule Bonn-Rhein-Sieg, Institute of Visual Computing, Sankt Augustin, Germany

Academic Assistant

Dr.-Ing. Roland Leißa
Saarland University, Saarbrücken, Germany

Day of Colloquium

19. December 2019

Dedicated to my family and the loving memory of my brother Tobias Weier
⋆1990 – †2014

ABSTRACT

Computer graphics research strives to synthesize images of a high visual realism that are indistinguishable from real visual experiences. While modern image synthesis approaches enable to create digital images of astonishing complexity and beauty, processing resources remain a limiting factor. Here, rendering efficiency is a central challenge involving a trade-off between visual fidelity and interactivity. For that reason, there is still a fundamental difference between the perception of the physical world and computer-generated imagery.

At the same time, advances in display technologies drive the development of novel display devices. The dynamic range, the pixel densities, and refresh rates are constantly increasing. Display systems enable a larger visual field to be addressed by covering a wider field-of-view, due to either their size or in the form of head-mounted devices. Currently, research prototypes are ranging from stereo and multi-view systems, head-mounted devices with adaptable lenses, up to retinal projection, and lightfield/holographic displays. Computer graphics has to keep step with, as driving these devices presents us with immense challenges, most of which are currently unsolved. Fortunately, the human visual system has certain limitations, which means that providing the highest possible visual quality is not always necessary. Visual input passes through the eye's optics, is filtered, and is processed at higher level structures in the brain. Knowledge of these processes helps to design novel rendering approaches that allow the creation of images at a higher quality and within a reduced time-frame.

This thesis presents the state-of-the-art research and models that exploit the limitations of perception in order to increase visual quality but also to reduce workload alike - a concept we call perception-driven rendering. This research results in several practical rendering approaches that allow some of the fundamental challenges of computer graphics to be tackled. By using different tracking hardware, display systems, and head-mounted devices, we show the potential of each of the presented systems. The capturing of specific processes of the human visual system can be improved by combining multiple measurements using machine learning techniques. Different sampling, filtering, and reconstruction techniques aid the visual quality of the synthesized images. An in-depth evaluation of the presented systems including benchmarks, comparative examination with image metrics as well as user studies and experiments demonstrated that the methods introduced are visually superior or on the same qualitative level as ground truth, whilst having a significantly reduced computational complexity.

KURZFASSUNG

Ein wesentliches Ziel der Computergrafik ist es Bilder zu synthetisieren, die sich nicht von den realen visuellen Erfahrungen unterscheiden. Während moderne Ansätze der Bildsynthese die Erstellung digitaler Bilder von erstaunlicher Komplexität und Realismus ermöglichen, sind die Verarbeitungsressourcen ein limitierender Faktor. Dabei ist eine Effizienzsteigerung die zentrale Herausforderung. Das Erzeugen qualitiativ hochwertiger Bilder bei gleichzeitiger Interaktivität unterliegt fortwährender Komprommisse. Aus diesem Grund gibt es immer noch einen grundlegenden Unterschied zwischen der Wahrnehmung der physischen Welt und der von computergenerierten Bildern.

Gleichzeitig treiben die Fortschritte in den Display-Technologien die Entwicklung neuartiger Anzeigegeräte voran. Der Dynamikumfang, die Pixeldichten und die Bildwiederholraten nehmen ständig zu. Anzeigesysteme ermöglichen die Adressierung eines größeren Sichtfeldes, entweder aufgrund ihrer Größe oder in Form von Head-mounted Displays. Derzeit beobachten wir Forschungsprototypen von Stereo und Multiview-Systemen, Head-mounted Displays mit adaptierbaren Linsen bis hin zur Netzhautprojektion und Lightfield-/Holografie-Displays. Dabei stellt uns das Betreiben dieser Geräte von der Seite der Computergrafik vor große, bislang ungelöste Herausforderungen. Glücklicherweise hat das menschliche Sehsystem bestimmte Limitationen, so dass eine höchstmögliche visuelle Qualität nicht immer erforderlich ist. Visuelle Reize werden durch die Optik gefiltert, durch die Netzhaut erfasst und auf höheren Strukturen im Gehirn verarbeitet. Die Kenntnis dieser Prozesse hilft bei der Entwicklung neuartiger Rendering-Ansätze, die es ermöglichen, Bilder in höherer Qualität und in kürzerer Zeit zu synthetisieren.

Diese Arbeit diskutiert den neuesten Stand der Forschung und präsentiert die Modelle, die die Grenzen der Wahrnehmung ausnutzen, um eine verbesserte visuelle Qualität und eine ressourcenoptimiertere Synthese von Bildern zu ermöglichen – ein Bereich bekannt als Perception-driven Rendering. Aus dieser Forschung resultieren mehrere praktische Rendering-Ansätze, die es ermöglichen, sich einigen der grundlegenden Herausforderungen der Computergrafik zu stellen. Durch den Einsatz unterschiedlicher Tracking-Hardware, Anzeigesysteme und Head-Mounted Devices wird das Potenzial der vorgestellten Systeme aufgezeigt. Die Erfassung spezifischer Prozesse des menschlichen visuellen Systems kann durch die Kombination mehrerer Messungen mit maschinellen Lerntechniken verbessert werden. Verschiedene Abtast-, Filter- und Rekonstruktionsverfahren unterstützen die visuelle Qualität der synthetisierten Bilder. Eine eingehende Bewertung der vorgestellten Systeme einschließlich Benchmarks, vergleichender Untersuchungen mit Bildmetriken sowie Anwenderstudien und Experimenten zeigt, dass die vorgestellten Methoden bei deutlich reduziertem Rechenaufwand oftmals visuell überlegen oder auf Augenhöhe im Vergleich zur Referenz sind.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support, inspiration, and love of many people.

First of all, I would like to express my sincerest gratitude to my supervisors Prof. Dr. André Hinkenjann and Prof. Dr.-Ing. Philipp Slusallek for their continuous support, mentorship, and guidance. I am very grateful to have found two people with whom to share and discuss my ideas. They have put these ideas to the test and challenged me to strive for higher goals.

Exceptional thanks and my deepest gratitude are for to my main collaborator and friend Thorsten Roth. Some ideas would never develop into successful projects without someone to bounce them off and to encourage me to rethink them. To this end, I would also like to thank all the people at Saarland University that helped me to pursue my academic goals. Here, I especially want to mention the very fruitful collaboration with Arsène Pérard-Gayot. His deep understanding of performance-critical GPU programming helped in implementing some of the perception-driven rendering pipelines to be found in this thesis. Also, I am grateful to the Hochschule Bonn-Rhein-Sieg for their financial support, not only in terms of stipends but also for my employment as a research associate.

Nonetheless, there are numerous other people to be mentioned: Especially Jens Maiero and Ernst Kruijff with whom I not only shared an office but I count as friends. Furthermore, I would like to thank my other colleagues Katharina Stollenwerk, Oliver Jato, David Scherfgen, Timur Saitov, and all the others I had the pleasure to work with. I thank them for all their remarks, your proofreading skills, and all those little things that made my day.

I am also grateful to the anonymous reviewers for their thorough reviews on our papers. They definitely helped to improve our work and look at certain problems from other perspectives. I would like to thank all the other fantastic people in the computer graphics community that I had the pleasure to get to know and to work with. They made every conference an enjoyable experience.

A big thank you also goes to all my friends that managed to keep me in a good mood, made me laugh when I was feeling down and allowed to keep me focused on the important things in life.

Last but most importantly, this work would not have been possible without the continuous support and love of my family. Life has been challenging for us, to say the least, over the last several years I married, became a father and had to cope with the sudden death of my beloved brother. There is rarely a day where I do not think of him. My life included many ups and downs. I would like to thank my family for all their support and love throughout these times, always encouraging me to keep pursuing my dreams

– to my wife, my parents and in-laws, my brother, and
to my little son Noah and daughter Lea, who have brought so much joy to my life.

PUBLICATIONS

Some of the presented ideas and figures have appeared previously in other publications. Moreover, parts of this thesis are based on published works that I have co-authored. The passages of text from these works are explicitly marked and used with the explicit permission of the co-authors. I was the primary researcher and author of all the papers listed below.

- Chapter 2, Chapter 3 and Chapter 4 are a reproduction of our state-of-the-art report, and are significantly extended for this thesis to include physiological backgrounds, a more focused discussion of perceptual models, an approach to perception-driven rendering as regards efficient rendering techniques, as well as significant updates to capture the most current work relevant to this thesis.

  [Wei+17] Martin Weier et al. *"Perception-driven Accelerated Rendering."* In: *Computer Graphics Forum (Proceedings of Eurographics)* 36.2 (Apr. 2017).

- Chapter 5 is a composition of the material published in the following contributions:

  [WHS15] Martin Weier, André Hinkenjann, and Philipp Slusallek. *"A Unified Triangle/Voxel Structure for GPUs and its Applications."* In: *Journal of WSCG*. WSCG 24.No. 1-2 (2015), pp. 83–90. ISSN: 2464-4617.

  [Wei+14a] Martin Weier, Jens Maiero, Thorsten Roth, André Hinkenjann, and Philipp Slusallek. *"Enhancing Rendering Performance with View-Direction-Based Rendering Techniques for Large, High Resolution Multi-Display Systems."* In: *11. Workshop Virtuelle Realität und Augmented Reality der GI-Fachgruppe VR/AR*. Sept. 2014.

  [Wei+14b] Martin Weier, Jens Maiero, Thorsten Roth, André Hinkenjann, and Philipp Slusallek. *Lazy Details for Large High-Resolution Displays*. SIGGRAPH Asia. 2014. Poster.

  [Wei+13] Martin Weier, André Hinkenjann, Georg Demme, and Philipp Slusallek. *"Generating and Rendering Large Scale Tiled Plant Populations."* In: *JVRB - Journal of Virtual Reality and Broadcasting* 10.1 (2013).

- Chapter 6 is based on the work published in the following paper:

  [Wei+16] Martin Weier, Thorsten Roth, Ernst Kruijff, André Hinkenjann, Arsène Pérard-Gayot, Philipp Slusallek, and Yongmin Li. *"Foveated Real-Time Ray Tracing for Head-Mounted Displays."* In: *Computer Graphics Forum (Proceedings of Pacific Graphics '16)*. Oct. 2016.

- Chapter 7 is based on the work published in the following articles:

  [Wei+18a] Martin Weier, Thorsten Roth, André Hinkenjann, and Philipp Slusallek. *"Foveated Depth-of-Field Filtering in Head-Mounted Displays."* In: *ACM Transactions on Applied Perception (TAP)*. Vancouver, Canada, Aug. 2018. Best Paper Award, invited article.

[Wei+18b] Martin Weier, Thorsten Roth, André Hinkenjann, and Philipp Slusallek. *"Predicting the Gaze Depth in Head-mounted Displays using Multiple Feature Regression."* In: *Proceedings of the ACM Symposium on Eye Tracking Research and Applications (ETRA)*. Warsaw, Poland, June 2018.

Moreover, I have co-authored the following publications relevant to this thesis.

- Chapter 5 describes some ideas with the aim of abstracting the renderer to be used in a Docker environment, similar to our work in:

  [GWH17] Björn Ludolf Gerdau, Martin Weier, and André Hinkenjann. *"Containerized Distributed Rendering for Interactive Environments."* In: *Virtual Reality and Augmented Reality*. Ed. by Jernej Barbic, Mirabelle D'Cruz, Marc Erich Latoschik, Mel Slater, and Patrick Bourdot. Cham: Springer International Publishing, 2017, pp. 69–86. ISBN: 978-3-319-72323-5.

- Chapter 5 does present some ideas that have later been successfully applied to path tracing. This is discussed in:

  [Rot+15] Thorsten Roth, Martin Weier, Jens Maiero, André Hinkenjann, and Yongmin Li. *"Guided High-Quality Rendering."* In: *11th International Symposium on Visual Computing (ISVC)*. 2015.

- Chapter 6 and Chapter 7 make use of a ray tracing core that is likewise based on the work of [ALK12] and on the work of Arsène Pérard-Gayot and myself. Though substantially extended for this thesis, initially, the renderer was developed for comparison purposes to evaluate the work presented in:

  [Pér+17] Arsène Pérard-Gayot, Martin Weier, Richard Membarth, Philipp Slusallek, Roland Leißa, and Sebastian Hack. *"RaTrace: Simple and Efficient Abstractions for BVH Ray Traversal Algorithms."* In: *Proceedings of the 16th International Conference on Generative Programming: Concepts & Experiences (GPCE)*. ACM. Vancouver, BC, Canada, Oct. 2017, pp. 157–168.

- Chapter 6 deals with an in-depth evaluation of the eye tracking data, acquired in [Wei+16]. Parts of Section 6.3.2 are based on a data analysis presented in:

  [Rot+17] Thorsten Roth, Martin Weier, André Hinkenjann, Yongmin Li, and Philipp Slusallek. *"A Quality-Centered Analysis of Eye Tracking Data in Foveated Rendering."* In: *Journal of Eye Movement Research (JEMR)* 10.5 (2017).

CONTENTS

Part I

FIRST THINGS FIRST

Und jedem Anfang wohnt ein Zauber inne

In the core of every beginning there is some magic

Stufen
Hermann Hesse (⋆1877 - †1962)

1 INTRODUCTION

The rise of modern computing systems in the second half of the twentieth century and the ability to transform electrical signals into images was the beginning of computer graphics. From its early approaches using line drawings of 3D objects on random scan displays up to present day graphics on raster screens with ever-increasing resolutions, gamut and refresh rates, computer graphics has certainly come a long way. Nowadays, rendering methodologies allow images to be synthesized with astonishing visual realism and beauty. Still, achieving the goal of presenting a computer-generated scene in a convincing and compelling way that cannot be distinguished from the real world and potentially in real-time remains one of the most central challenges for computer graphics. While processing power is steadily increasing, researchers continue to push the boundaries by developing methods that can account for more and more phenomena of the real world. Examples include real-time global illumination, accurate depth-of-field, motion blur, or spectral effects. Likewise, as technology progresses, high-fidelity graphics demands for scenes with increasing complexity, and images need to be generated at higher refresh rates, lower latencies, and increasing resolutions. Researchers frequently discuss the efficiency of methods, either if they result in a higher number of Frames-per-Second (FPS) or if the visual quality is increased in comparison to the state-of-the-art in a shorter time frame. Increasing *rendering efficiency* is, therefore, an essential research goal. There still is a fundamental difference between the perception of the physical world and computer-generated imagery.

At the core of most graphics rendering systems lies the question of how to turn a description of an n-dimensional model into a representation that can be presented to an observer. Herbey, the target is the most essential perceptual channel, the human's sight. However, the visual system has some limitations and the highest possible visual quality is not always necessary. The knowledge of those limitations can be used to develop better and more efficient rendering systems, a field known as *perception-driven rendering*. Note that central to the field of perception-driven rendering is not the question of whether a method is targeting perception: Rendering systems commonly target one perceptual channel or the other, when for example rendering in color. Rather more, the central question is how to *exploit the limitations or use the potentials of perception to enhance the quality of a method whilst maintaining its performance and vice versa.* Perception-driven rendering is based on a close understanding of the Human Visual System (HVS) in order to improve the quality, the speed of generation, and comprehensibility of images.

In the last two decades, with the rise of flat-screen technology, display sizes and resolutions have continuously increased. Nowadays, 4k displays with up to 80 inches are able to increase the observer's Field of View (FoV) at typical viewing distances whilst still maintaining a high pixel density. Wall-size displays with 8k resolution and more are entering the markets. Large, high-resolution, projection-based displays and high-resolution tiled display walls have become well-established installations. At the same time, higher pixel densities are available on devices

with smaller form-factors, e.g. on tablets and smartphones. In the last few years, another resurgent trend, driven by the advances in displays is Virtual Reality (VR) and Augmented Reality (AR) technology. With the introduction of mass production, a range of high-quality Head-Mounted Displays (HMDs) with a wide FoV have become available on a commodity level. Novel technologies such as displays with a growing number of displayed views per pixel (ranging from stereo, multi-view, to holographic or lightfield displays) are advancing beyond the prototype stage. Multi-layered displays and adaptable lenses are integrated into HMDs to provide a better VR experience. AR headsets allow covering a wider part of the FoV. Prototypes for retinal displays and bionic contact lenses have begun to emerge. Additionally, the display's dynamic range and refresh rates are ever increasing and display latencies are continually reduced. All these advances in display technologies will increase the requirements on image synthesis techniques. Today, displaying seemingly photorealistic graphics at high refresh rates already is computationally demanding, especially for high pixel densities, a wide FoV, and when rendering in stereo. The achieved realism is greatly limited by hardware capabilities and many desirable but costly aspects of reality cannot be taken into consideration. The demand for high-fidelity graphics targeting ever-increasing display systems will cause significant issues, especially due to limited compute power and constrained bandwidth.

Although the HVS seems to allow high-quality images to be perceived which are not bound to a fixed frame rate, it does have several limitations. Visual input passes through optics, is filtered and (down-)sampled on the retina before it is transmitted over the optical nerves to enable high-level processing in the visual cortex. Here, processing can also rely on other perceptual channels and the brain's ability to access memory. Computer graphics greatly benefits from of a close understanding of the potentials and limitations of how images are processed and perceived in order to thereupon optimize rendering techniques.

Currently, several strong trends in the graphics community can be observed. Hardware supported ray tracing is becoming available. Yet resources must be spent wisely, this allows for more efficient and flexible sampling processes [Sti18; Bar19]. Besides this, techniques from the field of machine learning are entering the rendering pipelines and already allow for guiding computational resources and sampling processes more efficiently [Bem+19]. Convolutional neural networks make it possible to simulate the entire visual pipeline, e.g. to blindly estimate image quality without considering reference images [RW17] as well as to model attentional processes [Wol+19]. At the same time eye-tracking hardware to perform active measurement of the gaze is becoming available at a consumer level [Tra17; Cor19]. Important in this context are *gaze-contingent methods* that adapt their behavior, based on measurements of where a person is looking. Also, head-mounted devices and displays help to target the perceptual channels more directly. Knowledge about perceptual processes is gaining increasing importance.

The goal of the work presented in this thesis is to develop methods that exploit the limitations of the HVS to visualize complex models either in a time-constraint setting or potentially improving the quality of the rendering while still maintaining performance. These models are otherwise challenging due to either their geometric complexity or a high visual fidelity needs to be achieved - especially if low-latency real-time rendering is required. To this end, the central parts of the HVS that are involved when an image is turned into a percept are discussed. This understanding allows the limitations of the HVS to be defined and introduce the models used to describe them. A framework is developed from these theoretical considerations, showing various state-of-the-art techniques that can be used to increase rendering

efficiency. Often these approaches try to adapt the 3D scene (i.e., the sampled function) or the sampling process itself in order to provide more "visually-pleasing" results in a shorter timeframe. In addition, a great number of methods is dedicated to developing systems that attempt to reuse information across time and space, either by reusing samples temporally, or by means of post-processing techniques that filter otherwise disturbing visual artifacts. As the sampling of a higher dimensional signal with a limited sampling frequency is likely to cause spatial and temporal artifacts, a closer look is taken on the question why certain artifacts (such as noise or temporal instabilities) are perceptually distracting. This knowledge enables different sampling and rendering techniques to be discussed that are an integral part of modern computer graphic pipelines. Finally, this thesis showcases how these theoretical insights can be applied to real-world use cases. An evaluation of these methods is carried out either by performing and evaluating user studies or by using perception-driven image metrics demonstrating the validity of the method.

1.1 SUMMARY OF CONTRIBUTIONS

This thesis builds on the research carried out in a number of prior works. The major contributions and results are listed below:

- An overview of the building blocks, potentials, and limitations of the HVS (Chapter 2) with a discussion on their implications for image synthesis approaches (Chapter 3).
- An in-depth discussion of the state-of-the-art in perception-driven rendering covering gaze- and non-gaze-contingent methods to increase rendering efficiency (Chapter 4).
- A hybrid acceleration structure using voxel and polygonal information along with a perception-driven Level-of-Detail (LoD) selection scheme for view-directed rendering (Chapter 5).
- A gaze-contingent rendering framework exploiting the limitations of the retinal acuity of the HVS (Chapter 6).
- A machine learning approach to support gaze-depths measurements in HMDs using off-the-shelf eye tracking hardware (Chapter 7).
- A gaze-contingent rendering framework exploiting the limitations of the optical system using Depth-of-Field (DoF) to filter rendering artifacts (Chapter 7).
- User studies and experiments evaluating the quality of the approaches presented (Section 5.5, Section 6.3, Section 7.3, and Section 7.4).

1.2 OUTLINE OF THE THESIS

The thesis is organized into four parts and eight chapters. While the first part, including this introductory chapter, provides a brief introduction to the field of efficient and perception-driven rendering, the second part of the thesis describes the most relevant theoretical foundations to this work. Initially, an overview of the HVS is provided in Chapter 2. This chapter details the physiological parts that are involved in the vision process, giving the relevant per-

ceptual foundation and showing limitations that can be exploited in the context of perception-driven accelerated rendering methods. Chapter 3 introduces the (mathematical) models that are used to describe these limitations. Most important here are the discussions of different visual acuity and contrast sensitivity models as well as the presentation of eye models that make it possible to describe the HVS as an optical system. Following the models of the HVS, Chapter 4 looks at perception-driven rendering from a different perspective, considering sampling and image synthesis techniques. The chapter shows the fundamental problems of rendering and the necessity for more efficient methodologies. Finally, a general model of efficient rendering techniques is derived that allows efficient rendering methods to be structured. This model determines the systematic literature review of the state-of-the-art for efficient and perception-driven rendering in the corresponding sections.

The core contributions of this work are set out in part three. To begin with, Chapter 5 presents our LoD framework using hybrid polygonal and voxel data in a mixed scene representation. A gaze-contingent selection scheme is employed to adapt the LoD to the viewer's gaze. This enables to adapt the rendering quality based on the FoV and acuity limits of the HVS. Besides developing the rendering system, I performed benchmarks, a metric-driven evaluation and a user study to provide insights into this system. Later, Chapter 6 discusses our efforts to adapt the visual quality based on the retinal capabilities of the HVS using eye tracking. Besides developing the rendering and reprojection pipeline as well as benchmarking the system, I designed and executed the user study. Finally, our work presented in Chapter 7 exploits the optical limitations of the HVS by hiding artifacts using a DoF filter. Here, I developed both, the rendering and DoF filtering framework as well as the machine learning approach to derive more accurate gaze depths. I was also in charge for the execution, design, and evaluation of the user study.

The final part of the thesis in Chapter 8 comprises a discussion of the main contributions made by this research. Some interesting developments and trends, as well as possible avenues for future developments are contemplated here.

Part II

THEORETICAL FOUNDATION

This part constitutes the theoretical foundations for this thesis. To this end, Chapter 2 gives an overview of the physiological components of the human visual system. This is used to approach the processes of vision from a perspective that allows for focusing on the perceptual implications and findings. Continuing, Chapter 3 does present the models for these processes that are commonly used in the graphics community to aid rendering systems. We divide these in low-level models that describe individual properties and high-level models that try to derive representations of a more complete visual processing. Finally, Chapter 4 discusses related work, i.e., methods that try to enable more efficient rendering systems, especially those that explicitly exploit the limitations of vision and perception.

2 THE HUMAN VISUAL SYSTEM

Limitations and Potentials

Wär nicht das Auge sonnenhaft,
die Sonne könnt es nie erblicken.
Läg nicht in uns des Gottes eigne Kraft,
wie könnt uns Göttliches entzücken?

Were not our eye the sun's own kin
the sun behold our eye would never.
If not the Lord's own power dwelled within
could things divine delight us ever.

Zahme Xenien III
Johann Wolfgang von Goethe (⋆1749 - †1832)

In this chapter, the main parts and mechanisms of the Human Visual System (HVS) are described. First, the visual system is discussed from a physiological perspective. Here the components are described that are responsible for turning light into a percept. Following that, from the physiological view, a simplified model of the HVS is used to describe the process of vision from a perceptual perspective. In line with the main goals of this thesis, the limitations are highlighted that can be applied to optimize rendering techniques.

> *We argue that modulating ... an object based on its perceptual content first requires a principled perceptual model. The first step in developing such a model is to understand the fundamentals of the human visual system, including how it is designed and how it is believed to function.*
>
> *David Luebke et al. [Lue+03, p. 239]*

CONTRIBUTIONS BY THE AUTHOR This chapter is based on our state-of-the-art report:

> Martin Weier et al. "*Perception-driven Accelerated Rendering.*" In: *Computer Graphics Forum (Proceedings of Eurographics)* 36.2 (Apr. 2017).

In contrast to the previously published report, I complemented this thesis with a description of the physiological components and processes that constitute the HVS. This description is provided in Section 2.1. The high-level model of the basic components responsible for human perception, which is found in Section 2.2 (Figure 16), was developed by Michael Stengel, Thorsten Roth, Steve Grogorick, and myself. This model was also used to provide a structure for the state-of-the-art report. I made major contributions to all relevant sections following this abstraction in the report and I substantially extended and revised these when writing Section 2.2 for this thesis. In this context, I would particularly like to highlight my additional explanations on visual acuity and hyperacuity phenomena as well as on ocular motility.

2.1 PHYSIOLOGIC VIEW

The HVS involves numerous physiological components which acquire, transmit, and process visual information. For that reason, the structures, cells, and pathways of the visual system are briefly described here. A more comprehensive overview of the HVS can be found in the works by Adler et al. [Adl+11], Goldstein et al. [Gol01], and Snowden et al. [STT12], that built the foundation for the following overview.

2.1.1 *The Eye*

The eye transforms incoming light into electric signals that are transmitted to a higher level of neural processing. An overview of the eye from a physiological point-of-view is illustrated in Figure 1.

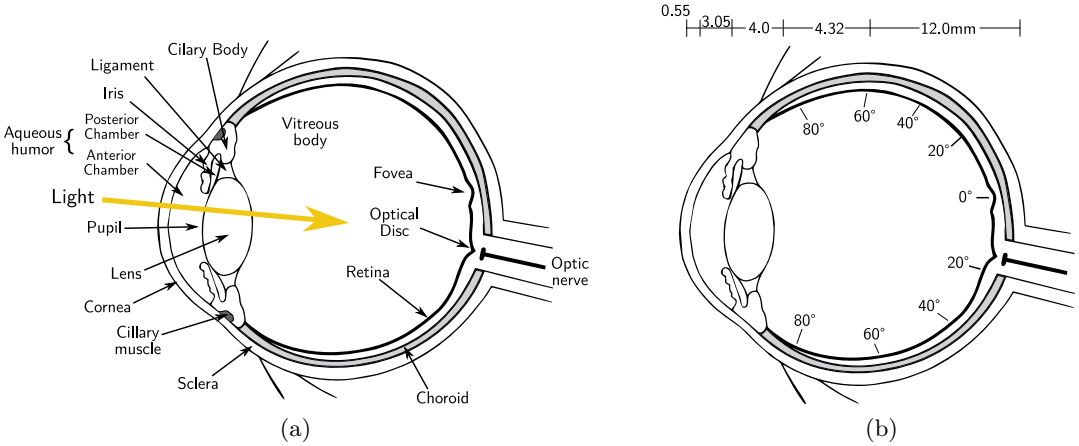


Figure 1: Schematics of the human eye. *Illustrations after Goldstein [Gol01, p. 54]*

The *cornea* and *sclera* constitute the outermost layers of the eye. Both tissues provide the structural integrity and protect the eyeball from physical injury. First, light entering the eye passes the transparent *cornea*. It covers 1/6th of the total surface's front of the globe. The whiter opaque sclera covers the remaining 5/6th. [Adl+11, p. 71] Interestingly, the cornea is the main lens of the eye. Three-quarters of the eye's focusing power come from the cornea (approx. 43 diopters), and only a quarter from the *lens* itself [STT12, p. 21]. The total optical power of the human eye is about 60 diopters [Gol01, p. 58]. Besides the refractive power, the transparency is a critical optical property of these structures. The cornea is the only transparent tissue in the human body. As such, corneal transparency has occupied scientists for over half a century. Nowadays, this feature is widely attributed to both the general transparency of the cornea's cells as well as the special lattice-like arrangement of collagen fibrils that support its structure [Adl+11, p. 79].

The spaces between the cornea, the iris, and the lens, namely the *anterior* (front) and *posterior* (rear) chambers are filled with an aqueous humor. This humor maintains the intraocular pressure, inflates the globe of the eye and supplies nutrients (and oxygen) to the

tissues that lack a direct blood supply [Sci10, pp. 39-40]. Moreover, its refractive power is another factor that contributes to the eye's focusing abilities.

In order to control the amount of light that can enter, the eye is equipped with an adaptable shutter, the *iris*. It consists of three main layers [Sci10, pp. 52-53]. The *anterior layer* of *endothelium* maintains a hydrated state of the tissues through the aqueous humor. The *two stroma layers* contain the blood vessels and sphincter and dilator muscles, that control the contraction and expansion of the iris. While the anterior stroma layer additionally contains the pigment cells that determine the color of the eye, the posterior layer is also heavily pigmented. Here, this pigmentation (by the pigment melanin) serves to prevent light from passing through the iris tissue in order to reduce the amount of light scattering in the eye [Sci10, p. 53].

The iris is controlled by a dilator muscle that is located circumferentially, in the mid-periphery of the iris and by a sphincter muscle around the opening of the pupillary border. The dilator is attached to the pupillary border (Figure 2) and made up of approximately 20 motor segments, connected together but innervated individually by branches of the ciliary nerve. Nonetheless, in a normal iris, these segments receive nerve excitation in a roughly simultaneous fashion to open and close the iris. [Adl+11, pp. 509-510] Besides, the different components of the posterior and anterior iris undergo structural alterations in order to accommodate changes in pupil diameter during contraction and dilation [Adl+11, p. 510]. The mechanical non-linearities are important because they impose limitations [Loe99]. The average pupil can be adjusted from a minimum diameter of 2 mm up to its maximum at 8 mm [Gol10].

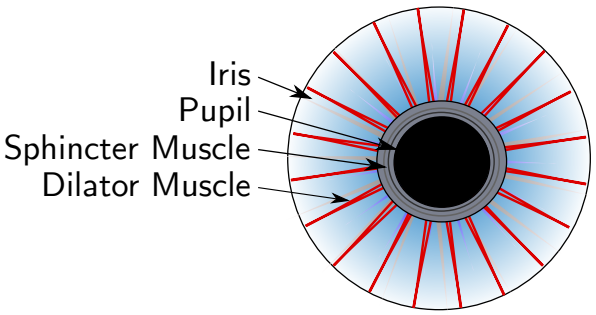


Figure 2: The iris and its muscular structures. The sphincter and dilator muscles control the amount of light that can fall into the eye.

The *lens* itself is a transparent body that is made out of fibrous cells, the lens fibers, that are enclosed in an elastic collagenous capsule. The lens's anterior surfaces is generally more flat than its posterior surface. It is held in place by a suspensory *ligament*; a series of fibers that connect the lens to the *ciliary body* of the eye also referred to as *zonules*. The *ciliary body* contains the *ciliary muscles* to control the shape of the lens. Besides the *ciliary body* produces the aqueous and vitreous humor.

This flexible suspension of the lens with the zonules and the ciliary muscles plays an important role when adapting the focus by changing the shape of the lens. This mechanical ability to compress and relax the lens is called *accommodation* [How12]. The accommodation may be as great as 12–16 diopters in a person under 20 but decreases with age. A person over the age of 55 is restricted to a range of less than one diopter, mainly due to a reduction in the elasticity of the lens and the capsule that holds the lens [Gol01, p. 59]. Although more recent investigations contributed to the understanding of the process of accommodation, its basics are still congruous with the original description by Helmholtz [Hel67][Adl+11, ch. 3]. This process is illustrated in Figure 3. When the eye is unaccommodated and focused for distance, the ciliary muscle is relaxed. Here, the zonular ligament fibers apply an outward-directed tension around the lens equator to hold the lens in a relatively flattened state.

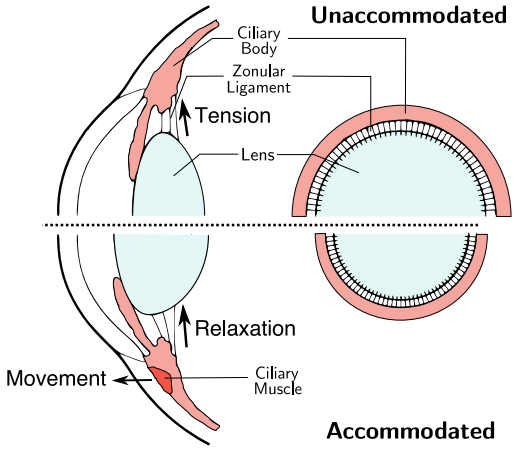


Figure 3: The process of accommodation. When the eye is focused at a distance, the zonular ligament applies a tension and flattens the lens. When the eye is accommodating, the forward force of the ciliary muscles relax the zonular ligament and the lens capsule becomes more round. *Figure adapted from Adler et al. [Adl+11, p. 49]*

If the eye needs to focus on near objects, the ciliary muscle contracts and the inner apex of the ciliary body moves *forward*. This movement of the apex stretches the posterior attachment of the ciliary muscles but *releases* the tension on the zonular fibers. This allows the lens to become more round through the force exerted by the lens capsule [Adl+11, p. 49]. This process shapes the lens anterior and posterior surface curvatures. This increases the lens's axial thickness and decreases the lens's equatorial diameter. Likewise, the anterior chamber depth and vitreous body depth decrease with accommodation. As a result of this process, the lens increases its optical power, thus refracting the incoming light rays stronger.

Behind the lens, the eye is filled with the *vitreous humor* and a semisolid structure forming the *vitreous body*. The solid structure is a collagen fiber scaffold. Embedded in the structure is the *vitreous humor*, a clear gel that provides nutrients to the structure and the retina. The vitreous body provides further structural support and serves to keep the underlying retina pressed against the choroid. The *retina* is the photosensitive nervous tissue that transmits the incoming light into chemical energy. The *choroid* is a vascular layer that supplies the retina with nutrients and oxygen. Essentially, the choroid is a layer of blood vessels and connective tissue sandwiched between the sclera and the retina. A dark layer of pigmented epithelium on the choroid helps to limit intraocular reflections that would otherwise disturb the perception [Sci10, p. 56].

Light rays eventually reach the *retina*. Here light passes through the different transparent layers to reach the light-sensitive outer segments of the photoreceptors. There are two types of photoreceptors, $6 \cdot 10^6$ *cones* and approximately 20 times as many *rods* [Gol13, p. 28]. Rods and cones contain large proteins called *opsin* and Vitamin A derivatives called *retinal*. These

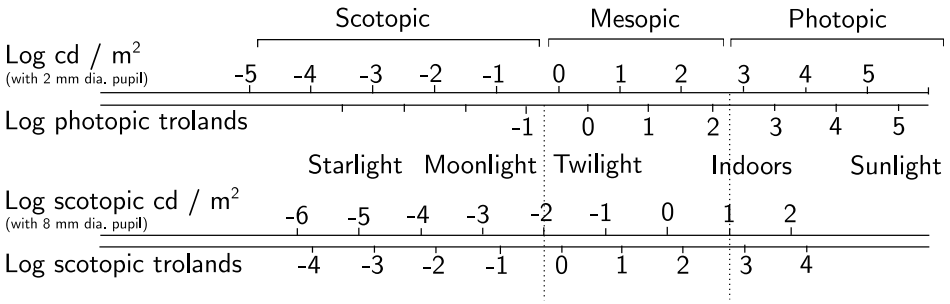


Figure 4: Scotopic, mesopic and photopic ranges for the macaque retina. *Illustration from Goldstein [Gol01, p. 64]*

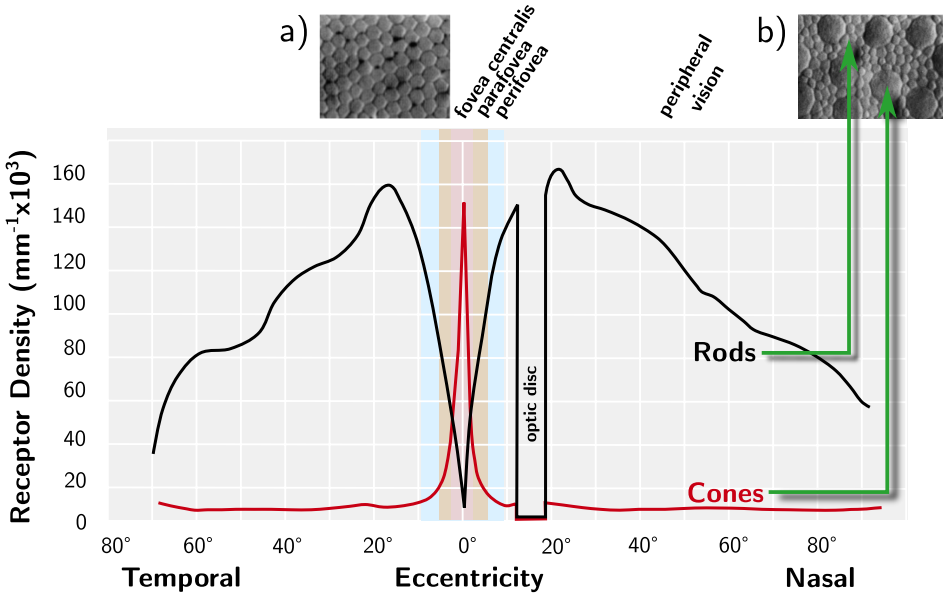


Figure 5: Retinal photoreceptor distribution. *Illustration adapted from Goldstein* [Gol13, p. 51] The fovea (a) contains only cones where the visual peripheral (b) contains rods and cones. These cell types are laid out in a Poisson-disc fashion. *Retinal microscopic recordings (a) and (b) from Curcio et al. [Cur+90]*

proteins form larger molecules, the photopigments. When light strikes such a photopigment, it initiates a reaction (<1 ms). This reaction results in the molecule splitting and subsequently generating an electric current.

Rods consist of only one type of photopigment, rhodopsin, and are responsible for the brightness sensation in lower-light conditions (scotopic vision) by providing monochromatic feedback. Cones are divided into three types for different wavelengths, namely L-cones (long wavelengths), M-cones (medium wavelengths) and S-cones (short wavelengths). Their responsiveness to different wavelengths is based on another form of light-sensitive pigments, photopsins, that differ in a few amino acids, depending on the sensitivity to the respective wavelengths [Gol01, p. 93ff]. As such, the cones are responsible for detailed color sensation (photopic vision). *"This duplex arrangement enables humans to see in a wide range of lighting conditions."* [Lue+03, p. 242] At scotopic levels, absolute sensitivity is high but since rods provide achromatic signals only, colors cannot be perceived [Fer+96]. In contrast, at photopic levels, sensitivity is dramatically lower but colors can be perceived due to the trichromatic nature of the cone cells. The region where both receptor types play a role is denoted as mesopic vision. The different levels and photopic ranges are illustrated in Figure 4.

The photoreceptors of different types follow the distribution pattern shown in Figure 5. The highest density of rods and cones is found in the *macula*. The central area of the macula is the *fovea* (approx. 5.2° around the central optical axis). It consists entirely of cones. The center of the *fovea*, the *foveola*, is the relative origin of our vision. It only consists of *M* and

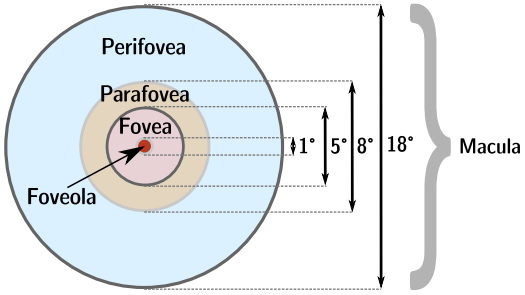


Figure 6: The regions of the macula around the foveola.

L cones and represents the central $1°20'$ of our visual field [Hen05]. Starting the *fovea*, the cone density drops significantly with increasing eccentricities (the angular distance to the optical axis) past the *parafovea* (approx. 5.2° to 9°) and *perifovea* (approx. 9° to 17°) [Cur+90]. A schematic overview is illustrated in Figure 6. These inner parts of the macula constitute *central vision*, while areas further away are referred to as *peripheral vision*.

Also, the density of the L-, M- and S-cone types differ. There are less S-cones than M and L cones. S-cones represent only about 7% of the cone population [Gol13, p. 69]. Thus, humans are much less sensitive to short 'blueish' wavelengths than to the 'reddish' and 'greenish': the medium and long wavelength of the spectrum. However, the S-cones are more spread out outside of the fovea and the ratio between the cone types varies greatly among different subjects [RW99]. The highest density of *rods* is approximately $15 - 20°$ around the *fovea*, this then decreases almost linearly. Just as the rods and cones have different densities across the retina, they have different spatial sampling distributions and follow a Poisson-Disc Distribution pattern [Wan95, ch. 3][Yel83].

Besides the rods and cones, the retina is a layered tissue composed of other cell types. A cross-section of the nervous structure of the retina is illustrated in Figure 7. Although there are very few dedicated pathways from the fovea for signals from individual foveal photoreceptors to the higher level neural structures, it is far more frequently the case that there are no true one-to-one connections for retinal photoreceptors. The rods and cones are connected to *bipolar cells* and *horizontal cells*. The bipolar cells are crucial retinal interneurons that transmit signals from the outer retina to the amacrine and ganglion cells (and less common interplexiform cells) [Gol01, p. 61]. There are various types of bipolar cells with distinctly different functions, they can for example:

- provide input to the high-resolution parvocellular stream that preserves specific information such as the type of photoreceptor input
- pool photoreceptor inputs for the lower-resolution, higher-gain magnocellular stream
- distinguish whether the light has increased or decreased [Gol01, p. 61]

The *horizontal* and *amacrine cells* participate in lateral interactions. They integrate potentials over (large) areas and provide feedback to the photoreceptors, adjusting the gain of the retinal circuits. *Ganglion* and *amacrine cells* form the ganglion cell layer. These cells further aggregate the output of the various classes of bipolar cells. Ultimately, they transmit parallel streams and amplify the local potentials from the bipolar cells to action potentials that can travel a longer distance through the visual pathways. [Gol01, p. 61]

There are also different types of ganglion cells [Gol01, p. 76]. The *parasol cells* (M ganglion cells) are mostly connected to rods and thus have little to no color information [HT00]. However, they have a fast response/refresh time. This way they contribute to the perception of movements, depth of objects, and small differences in brightness and contrasts.

The *midget cells* (P ganglion cells) are generally connected to the L- and M-cones and the *bistratified cell* (K ganglion cell) are connected mostly to the S-Cones. The P and K ganglion cells are much slower than the M ganglion cells. However, both are vital for the sensation of colors and structures. As presented in the next chapter the output of those ganglion cells directly map to higher level structures. In addition, there are less common ganglion types, for example to drive reflex-like movements of the eye (Section 2.1.2).

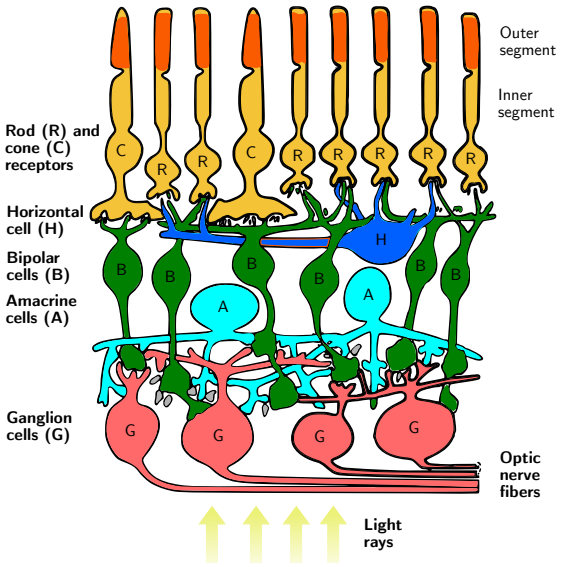


Figure 7: Cross section of the retina. Light passes through different layers of interconnecting neurons before it reaches the photosensitive rods and cones. *Illustration adapted from Adler et al. [Adl+11, p. 49]*

Another less frequent cell type is the *intrinsically photosensitive retinal ganglion cell*. As the name suggests, these cells are also responsive to photonic input. Thus they form a less-known third type of photoreceptor. However, their response to lighting changes is magnitudes slower compared to the other cell types. They adapt to the ambient lighting. Research has discovered that these cells have a vital role in controlling the circadian rhythm. Moreover, they take part in slow behavioral responses, also contributing to the regulation of the pupil size [WDB05; Eck+10]. Rods and cones are connected laterally by horizontal and amacrine cells and aggregated by ganglion cells. It is most unlikely that a given optic nerve fiber carries messages from only a single photoreceptor [Sci10, p. 92]. Hence, both visual acuity and contrast sensitivity cannot be described by the cone spacing alone but rather the density of the neural cells (Section 2.2.2). However, the ratio of ganglion cells to photoreceptors is highest in the fovea and decreases in a similar fashion to the decreasing density of rods and cones (Figure 8).

Ganglion cells have very distinct functions. This is one reason why their density, spacing, and overlap does not fully explain the visual acuity over the entire visual field [WR80]. One

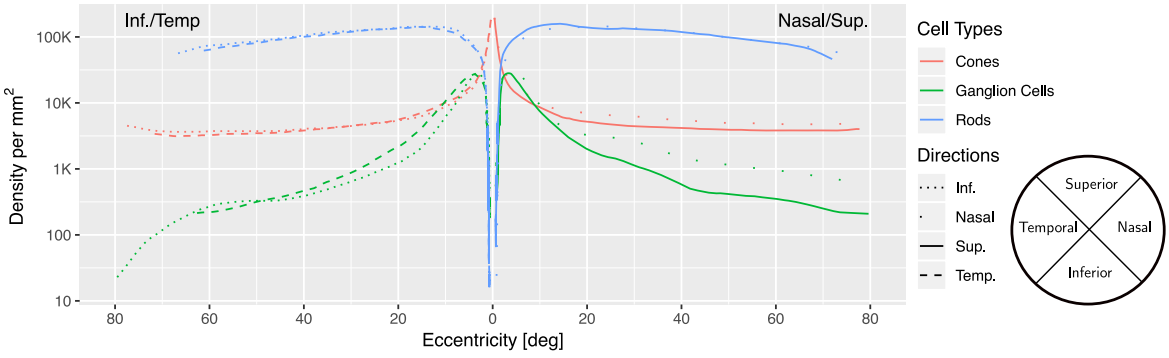


Figure 8: The retinal cell density of rods, cones and ganglion cells. Nasal, Temp., Sup. and Inf. indicate the directions moving away from the fovea. *Plotted data from Curcio et al. [Cur+90]*

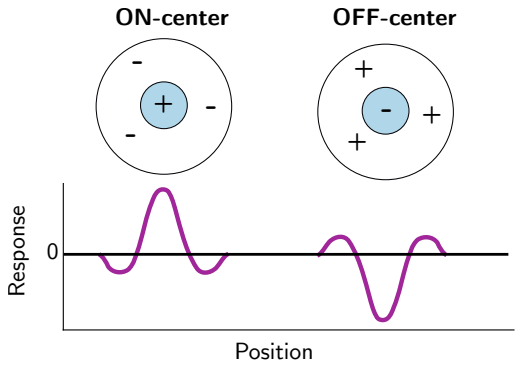


Figure 9: Receptive fields of ganglion cells and their inhibitory and excitatory behavior. *Illustration adapted from Snowden et al. [STT12, p. 50]*

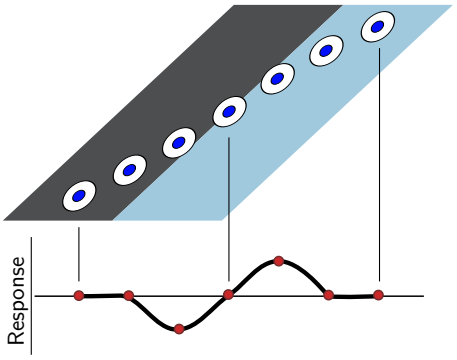


Figure 10: A series of ganglion cells and their inhibitory and excitatory behavior when sensing a brightness discontinuity. *Illustration from Snowden et al. [STT12, p. 54]*

of these functions of the ganglion cells becomes apparent when looking at the receptive fields. *"The receptive field of a [photosensitive] cell is the area on the retina over which the behavior of that cell can be directly influenced."* [STT12, p. 49] The receptive field of each neuron is roughly circular, with a small central and a larger surround region. There are two types of receptive fields that can be measured at ganglion-cell level: 'ON-center' and 'OFF-center' neurons. 'ON-center' neurons increase the firing rate when the light hits their center, and decrease when the light hits the surround. 'OFF-center' cells function in the opposite way. This is illustrated in Figure 9. Ganglion cells continually emit a background current. Consider now that both the excitatory and inhibitory part are equally stimulated: both stimuli would cancel each other out. Hence, even though the total intensity of light may greatly vary, these receptive fields only respond to relative ratios of intensity [STT12, p. 54]. *"So, the crucial message is that ganglion cells only signal the 'edges' or 'changes' in the pattern."* [STT12, p. 55] At the cellular level, this process generally happens in the horizontal cells that disable the spreading of potentials from excited neurons to neighboring neurons. This property is responsible for *lateral inhibition*. A series of receptive fields produces a high neural output for the edge as an excited neuron reduces the activity of its neighbors, as illustrated in Figure 10. In the last decade, researchers also discovered an additional active feedback mechanism between the horizontal cells and the cones that enable the HVS to actively boost contrasts along brightness discontinuities [Jac+11]. These mechanisms and *lateral inhibition* are one first form of processing in order to provide the sensation of a sharper image. In addition, these mechanisms are not only highly important for pattern recognition but also lead to our high responsiveness to image noise and jagged edges (Chapter 4).

Another factor that influences image quality on the retina is the *Stiles-Crawford effect*. As cones are less responsive to photons that touch them at an angle, light that is entering at the pupil's margin is perceived half as bright as the light entering the center of the pupil. As a result, light that passes through the edge of the pupil contributes less to image quality than light entering through the center [Adl+11, pp. 30-31].

The HVS can operate over an enormously wide range of intensities, from around $10^{-4}\,\mathrm{cd}\,m^{-2}$ under starlight conditions to around $10^{5}\,\mathrm{cd}\,m^{-2}$ under intense sunlight [Adl+11, ch. 20]. The adaptation to differences in brightness sensation mostly takes place on the retina. Only a very

minor part of about 1 log unit of this 9 log unit range is controlled by adapting the pupil's diameter [Adl+11, ch. 20]. The rest is achieved by adapting and switching between the rod-based and the cone-based pathways. This switching and the adaptation of the biochemical processes enable our photoreceptor systems to operate over a range of 5 log units (100,000-fold) or more. One aspect here is that photosensitive pigments need some time to recover and to adapt to different levels of illumination [Ade82; Bak49]. Likewise important in this process is that the electrical potentials generated by the photoreceptors can also be controlled by adapting the influx of calcium into the neural apparatus constituting the retina [NY88][Adl+11, p. 436ff]. This controls the action potentials of the neurons. Adaptation influences both visual acuity and color vision. A more in-depth discussion of the perceptual implications of adaptation take place in Section 2.2.2.

Finally, the already pre-processed electric signals are transmitted over the *optical nerve* to higher-level visual pathways. It carries the impulses from the ganglion cells of the retina to the visual centers in the brain. The nerve begins at the *optic disc* of the eye. As there are no photoreceptors in the optic disc, this region, which is approximately 1.5 mm across, is blind [Gol01, p. 55]. From here the optical nerve convergences the ganglion cells and then passes the signals out of the eye [Sci10, p. 71].

2.1.2 *The Visual Pathways and the Ocular Motility*

As the signal leaves the eye, the *optic nerve* enters the *cranium*, i.e. the skull. An overview of the visual pathways is presented in Figure 11. The neural fibers from the nasal of one eye and the temporal portion of the other eye, cross to the opposite side of the brain in the *optic chiasm*. However, the neural fibers from the other temporal or nasal portion of the optical fibers remain in the same hemisphere. The fibers forming the *optic tracts* now carry information, past the optical chiasm, into each brain hemisphere about the opposite hemifield of vision. When studying Figure 11, note, that a few fibers of the optical nerve project directly to the *suprachiasmatic nucleus*, which is located in the *hypothalamus*. The suprachiasmatic nucleus mainly acts as a master clock as it is highly involved in the circadian rhythm [Gol01, p. 56].

Around 10% of the optical fibers carry information to the *prectectum*, *superior colliculus*, and to the *pregeniculate* that are both closely located to the *Lateral Geniculate Nucleus (LGN)*. The remaining 90% of the axons in the optical tracts terminate in the LGN [Gol01, p. 56]. These are the critical ones for visual perception. However, before describing the details of the LGN, it is worthwhile looking at the other structures. The *prectum* is the structure that is responsible for the reflex-like controls of the lens and the pupil. The *superior colliculus* is responsible for the orienting movements of the head and the eye towards the Object of Interest (OoI). Signals from the superior colliculus also pass onwards to a structure called *pulvinar* towards the *posterior parietal cortex*. This structure is known to play a vital part in planned movements, spatial reasoning, and attention processes [STT12, p. 337-338]. The function of the *pregeniculate* that is directly located at the LGN remained unknown for a long time [Gol01, p. 56]. More recent research suggests that this structure plays a vital part in saccadic eye movements – a quick movement of the eye brought about by a brief but powerful activation of the eye muscles. The eye is subjected to a new orientation according to

the magnitude of this brief activation. Current research also suggests that the pregeniculate plays a role in the visual-ocular motor integration of the perceived scene [LF03].

The LGN receives the majority of the visual input. It is arranged in multiple layers that are segregated according to the origin of the retinal signal emerging from the retinal ganglion cells. These umbrella-like layers are illustrated in the inset in Figure 11. The *magnocellular cells* (M) located in the first and second layer are connected to the M ganglion cells, aggregating rods. Likewise, the parvocellular cells (P) are connected to the P ganglion cells, in turn, aggregating L- and M-cones, and the *koniocellular cell* (K) to K ganglion cells, aggregating S-cones (Section 2.1.1). Moreover, the different layers have different responsibilities based on the visual field and the eye from which the input emerges. *"Inputs from the nasal retina of the contralateral eye, which had crossed in the (optic) chiasm, synapse with cells in layers 1, 4, and 6, while inputs from the temporal retina of the ipsilateral eye contact cells in layers 2, 3, and 5. Each LGN layer contains an orderly, retinotopic map of the contralateral hemifield of vision, and the maps in the six layers are aligned."* [Gol01, p. 56] As such, the LGN mainly acts as a relay center that structures and distributes the visual inputs to a higher level of processing. From here, the split signals are distributed to the *optic radiations* and onward to the primary visual cortex. Before detailing the function and structure of the visual cortex in the next chapter, it is interesting to look at the ocular motility, i.e., the muscles and neural pathways which accommodate, adapt, and direct our vision.

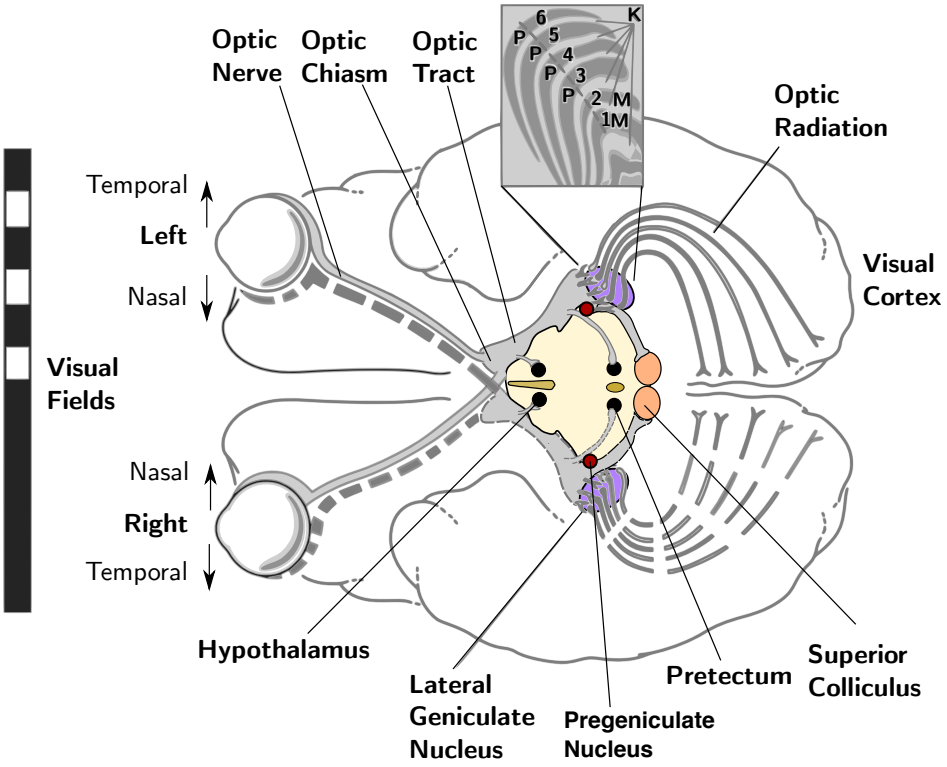
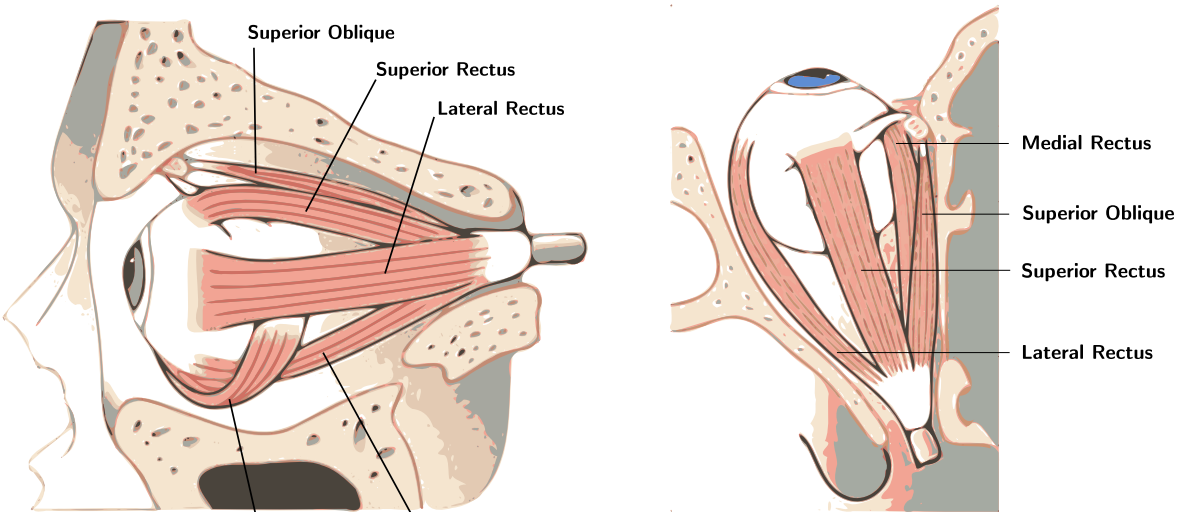


Figure 11: The visual pathways to the visual cortex. Signals travel through the optic nerve to the lateral geniculate nucleus (LGN) and from there to the optic radiations and on to the primary visual cortex. *Illustration adapted and extended from Goldstein [Gol01, p. 55]. Illustration of the LGN is adapted from Hendry and Calkins [HC98].*



(a) side (b) top

Figure 12: Schematics of the ocular muscles. *Illustrations after Rodieck [Rod98, p. 299]*

The accommodation is driven by the *accommodation reflex* to adapt the focus from a near object to a distant object and vice versa. While one of the main goals of accommodation is to focus the incoming light rays in order to maximize the retinal image contrast, current research suggests that focus changes also occur without such a monocular cue [Mar+17]. Influences of the optical vergence and measures based on the Stiles-Crawford effect may be other processes that help to sense the defocus. However, focus cues are generally interpreted behind the LGN in the visual cortex as only here the percept is interpreted as an image. While the signals that control the orientation of the eyeball also emerge from the hypothalamus (superior colliculus, prectum), the new accommodative state is generally driven by those higher level structures. Only reflex-like responses are triggered by the *prectum*.

The *pupillary light reflex* adjusts the pupil. This reflex is controlled by the retinal illumination that is signaled by the ganglion cells [Gol01, p. 59]. The *prectum* in the midbrain processes this information and sends signals to the ciliary nerve in order to contract or relax the iris' sphincter and dilator muscles. As the pupil size also affects image quality, the accommodation and pupillary light reflex are tightly entangled. An example is a small pupil which improves the image's focus on the retina when the retinal illumination is sufficiently high.

The eyeball itself is rotated by six external muscles (extraocular muscles). Each muscle is attached to the eyeball and the skull structure forming the eye's orbit. The anatomical structure of the muscular system is illustrated in Figure 12. The *lateral* muscles connect straight to the eye, while the *oblique* muscles are looped around and run obliquely [Rod98, p. 299]. The *medial and lateral* as well as the *superior and inferior* muscles are complementary pairs of flexors and extensor muscles. The *lateral* and *medial rectus* rotate the eyeball horizontally, while *superior* and *inferior rectus* rotate the eyeball vertically. The *superior oblique* and *inferior oblique* form a third complementary pair of muscles, the purpose of which is to rotate the eye around its direction of gaze. However, these torsional movements are very limited. Their main purpose is to maintain the visual horizon on the retinas based on the orientation of the head and the direction of gaze – an important process to locate objects in space and stabilize the eye movements [Rod98, p. 302]. Besides the state of the muscles of the eye,

image stabilization and locating objects also requires knowledge about the head rotation and the vestibular apparatus. An introduction to these processes and movements can be found in Rodieck's work [Rod98, p. 303pp]. However, in order to direct the gaze towards an OoI, at least the eyeball must be moved. If the eye is resting in a relaxed state, the complementary muscles are equally activated at a constant rest level. In case of neural activation of a complementary pair of muscles, it is always reciprocal. When the eyeball is moved, the reciprocity increases the rated of activation in one muscle, whilst decreasing the rate of activation of the complementary one. Besides voluntary movements, the eye also performs somewhat involuntary movements, such as saccades. A more detailed introduction to such movements is given in Section 2.2.3. Having directed the gaze to the OoI and the action potentials from the retina have passed the LGN via the *optic radiations*, the information can be interpreted using higher-level processing in the visual cortex.

2.1.3 *The Visual Cortex*

Finally, a visual signal passes from the optic radiations onward to the primary visual cortex, an area in the brain often referred to as V1 or *striate cortex* due to its stripy look. As illustrated in Figure 13 this area is directly located in the occipital lobe on the back of the head. Similar to the LGN the V1 area contains a retina-optic map of the contralateral hemifield. This means that things close together on the retina will trigger neighboring bits of the visual cortex. Moreover, the left V1 maps the right visual field and the right V1 maps the

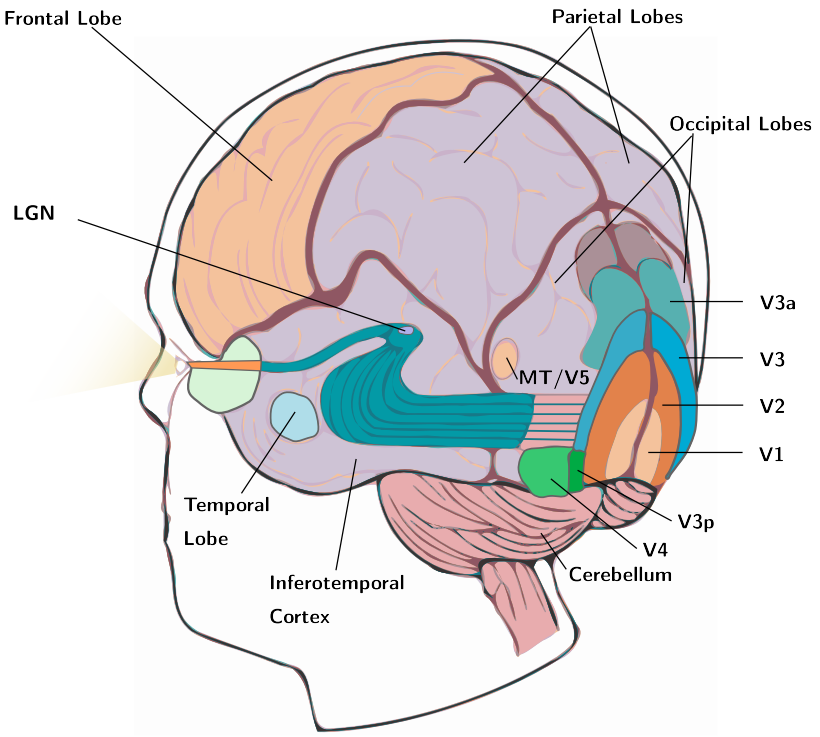


Figure 13: Approximate positions of different areas of the brain that are responsible for vision. Signals from the LGN reach V1 (striate cortex) and are in turn processed by the remaining extrastriate visual areas in the human brain. *Image redrawn from Snowden et al. [STT12, p. 89]*

left visual field with minimal overlap. However, within the map, the central area of the visual field is represented by a greater amount of neural cells so that it receives a disproportionately large representation [Gol01, p. 56]. This property is often modeled as magnification factor (Section 3.1.2) and directly relates to the higher density of ganglion axons on the retina. Unlike the retinal cells and the cells in the LGN, the neurons in visual cortex are selective on the orientation and the direction of a moving stimulus [BC69]. Hubel and Wiesel [Hub88] significantly improved our understanding of these cells and their functions. They distinct three basic types:

1. **Simple Cells** are tuned to inputs from different orientations. According to Hubel and Wiesel these cells can be thought of as bar or edge detectors. They aggregate the center-surround cells from the LGN to be responsive to a pattern in a distinct direction. These aggregates then have discrete ON and OFF regions.
2. **Complex Cells**, such as simple cells, are also most responsive to bars and edges. Complex cells mostly are aggregates of simple cells. This way, the receptive fields of a complex cell is larger and the cells do no longer have discrete ON and OFF regions. This makes their receptive fields phase invariant, i.e. it gives more robust responses to moving stimuli. Complex cells respond regardless of the exact location in the receptive field. Also, several complex cells respond optimally only to movement in a certain direction.
3. **Hypercomplex Cells**, or end-stop cells, in turn, aggregate complex cells. The hypercomplex cells are also selective for the specific orientation, motion, and direction of stimuli. However, they also decrease in firing strength when the length of a stimulus, such as a colored bar or line, does change.

While simple and complex cells respond to patterns and structure in still and moving images, hypercomplex cells enable to better perceive corners and curves in the environment by identifying the ends of a given stimulus [HW04]. All these responses to discontinuities in the visual input already explain a great portion of the eyes' responsiveness to (temporal) noise. Thus, methods that reduce noise and jagged edges are highly important in computer graphics (Chapter 4). The responses to different directions also build the basis for processes such as the evaluation of the contrast sensitivity and the *cortex transform* used in image metrics (Section 3.2.1). However, all this does likely not give the complete picture of the processes in V1 and this area is only the first of more than 30 cortical areas in the brain that process visual information [Gol01, p. 56][CHF92]. Moreover, the percept is influenced by other sensory channels such as sound, taste, and smell. Not least, memory plays a critical role in pattern and object recognition. Every area that is directly involved in vision following V1, is part of the *extrastriate cortex*.

An overview of some components of the process of vision is illustrated in Figure 14. This structure has tempted researchers into the seductive notion that each area is specialized for a particular aspect of vision. Commonly, models assume that V1 and V2 might be involved in processing basic visual features, V3 and MT/V5 are involved in motion detection, spatial localization, and hand and eye movements, V3/VP is involved in shape perception, V4 is responsible for color vision and V9 for eye movements etc. [STT12, pp. 90-91, p. 86]. Although there might be some truth to these ideas, generally, vision processes are more involved and include the interplay between various areas in the brain. A precise description of their structure and responsibilities is not only beyond the scope of this thesis but generally highly complex

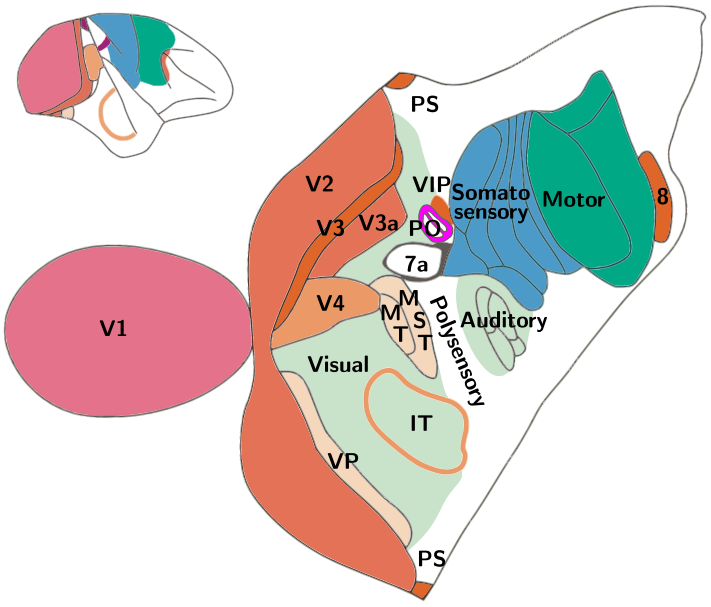


Figure 14: A map of the visual areas in the primate brain. The brain has been flattened so that both the areas on the surface of the brain (sulci) and those hidden in the folds (fissures) are visible. *Illustration based on Van Essen [CHF92] redrawn from Snowden et al. [STT12, p. 90]*

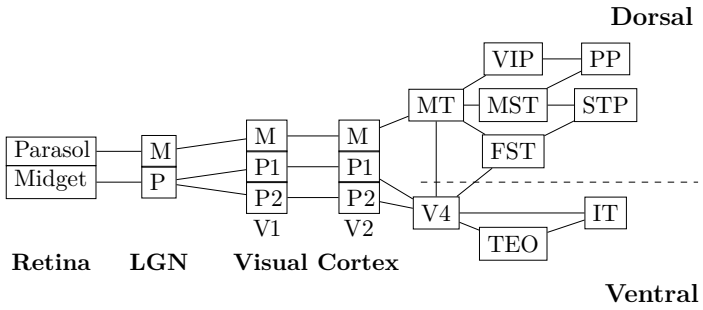


Figure 15: Schematics of the dorsal and ventral pathways in the visual cortex. *Image redrawn from Goldstein [Gol01, p. 57]*

from a physiological view and to a great extent not understood in all its details. A commonly used generalization on how streams of visual input are processed has been introduced by Van Essen et al. [CHF92], Merigan and Maunsell [MM93], and Mishkin et al. [MUM83]. A version of this model is illustrated in Figure 15. Here, visual information is processed in two streams: a *dorsal* and a *ventral* stream [Gol01, pp. 60ff]. The dorsal streams get the input mainly from the magnocellular layer of the LGN, and projects from V1 to V2 over to V3 and onward to MT and MST, as well as directly from V1 to MT. The ventral stream, receiving inputs mainly from the parvocellular but also magnocellular layers of the LGN, projects from V1 to V2, onward to V3, and V4 to reach IT as well as TEO [Gol01, p. 58]. Investigations and lesions of macaque brains led to the assumptions that the dorsal stream is concerned with the location in space and motion. Therefore, it has been described as the "where" stream. In contrast, the ventral stream was found to be concerned with object identification, form, and color, and has been called the "what" stream. [Gol01, p. 58] However, at this point the image information is a percept that is reconsidered from a perceptual view in the next section.

2.2 PERCEPTUAL VIEW

Based on the physiological structure of the HVS a simplified model of its basic functions is illustrated in Figure 16. The model was developed in the previously published state-of-the-art report [Wei+17] in order to provide an overview of the visual perception system. The model conglomerates different physiological components of the HVS and is used to highlight the stages a visual stimulus passes through before it is turned into a percept. Following the model, this chapter discusses the findings related to the perceptual implications of the HVS. As humans are equipped with two eyes, light constitutes two data streams that enable stereo vision. The optical system projects the stimuli onto the *retina* (the "sensor"). Human ocular motility allows our focus of attention to be on an OoI. The sensor turns light into electric potentials that are transmitted along the visual pathways from the eye to the LGN and the visual cortex. Here, different parts of the brain are involved in processing and interpreting the signals until a final mental representation, the percept of the environment is produced. These processes have access to our memory. Likewise, attentional mechanisms may (re-)direct our movement or impair our visual sense. In the following, the components of Figure 16 are discussed in greater detail.

2.2.1 *Optics*

The HVS is characterized by several unique *optical* qualities that are a result of both the position and shape of the eyes. With binocular vision and both eyes looking straight ahead, humans have a horizontal Field of View (FoV) of almost 190°. The FoV of a single human eye is approximately 95° away from the nose and 60° towards the nose horizontally. Vertically it is about 60° upward and 75° downward. If eyeball rotation is included, the horizontal FoV extends to 290° [HR95, p. 32]. While the human eye will receive visual stimuli over the full extent of the FoV, the way stimuli are processed in different parts of the visual field is highly affected by the spatially varying properties of the retina. However, a first influence

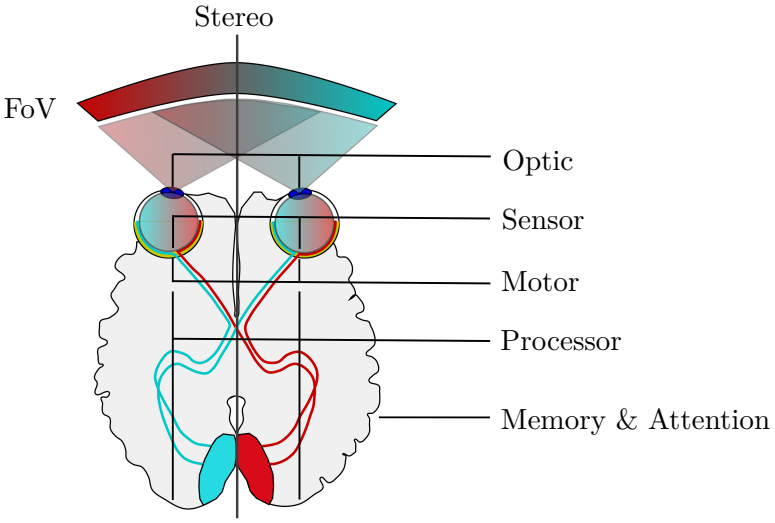


Figure 16: A high-level model of the basic components responsible for human perception. *Adapted from Weier et al. [Wei+17]*

that is affecting the spatial acuity of the HVS are the eye's optics. As discussed in Chapter 4 *aliasing* occurs if a signal contains frequencies higher than the observer's Nyquist rate [Sha49]. In human vision, undersampling effects occurs for spatial frequencies higher than approx. 60 Cycle per Degree (cpd) [Wan95, p. 24]. A cpd is a unit to describe spatial frequency. It is defined as one period in the alternating pattern of black and white spaces (sinusoidal grating pattern) at the projected size of 1°. However, the eye's optics in the cornea and lens act as a low pass filter with a cutoff frequency between 60 cpd and 80 cpd. At this spatial frequency the optics do not transmit the sinusoidal variations in the luminance of the object [Adl+11, p. 632]. This way, the signal that cannot be properly sampled and reconstructed is effectively removed through optical prefiltering, which is one efficient way to combat aliasing. The pupil is an additional important factor and serves as an aperture. This adjustment mostly affects the sharpness of the image, as the pupil can control only about one magnitudes of light intensity difference. This adjustment is largely triggered by a pupillary light reflex [Adl+11, ch. 25]. However, as discussed in Section 2.1.1, eye's adaptation to differences in brightness sensation (dark and light adaptation) mostly takes place on the retina. In addition, the size of the pupil does affect image quality. As all optical systems, the image quality is limited by the Rayleigh criterion [Adl+11, pp. 630-632]. When all factors of the optical and retinal capabilities are considered, a pupil with a diameter of 2–3 mm provides the best image quality [Gol01, p. 59]. However, in order to increase the amount of incoming light, the pupil can be as large as 8 mm, leading to a reduced focal range and non-optimal optical distortion.

Moreover, a healthy human being has two eyes. The distance between the eyes, the Interocular Distance (IOD), results in two streams of visual stimuli from slightly different viewpoints, which are combined in the brain by a process called stereopsis and enable perception of depth also referred to as *stereo vision* [Pal99, Chapter 5.3]. Usually, a gender-dependent mean IOD ranges between 62 mm to 65 mm [Dod04]. Depth perception is additionally enabled by visual cues such as parallax, occlusion, color saturation, and object size [CV95; Hel+10].

2.2.2 *Sensor*

The light is projected onto the retina, the photosensitive layer of the eye. As presented in the previous Section 2.1.1, the photosensitive and interconnection cells on the retina are not evenly distributed. Their density decreases at increasing eccentricities and is directly related to *visual acuity*, the "keenness of sight". As the density of the photoreceptors, visual acuity of the eye decreases significantly outside the small foveal region, where humans are able to generate a sharp image (acuity is already reduced by 75% at an eccentricity of 6°).

Visual acuity can be expressed as Minimum Angle of Resolution (MAR). Normal vision corresponds to 1 MAR, i.e. the eye is considered to have a minimum angular resolution of 1 minute of arc ($\approx 0.017°$). This minimal feature size relates to a spatial frequency of a sinusoidal grating pattern of alternating black and white spaces at 60 cpd, roughly the same as the eye's optical cut-off frequency (Section 2.2.1). Also, these upper limits of visual acuity can be calculated from the foveal cone spacing [Gol01, p. 69]. The entire relationship is illustrated in Figure 17. Under ideal conditions with a sharp bright white line in front of a uniform dark background, a feature can be detected with a limit of about 0.5 arc minutes [Adl+11, p. 627]. However, there is more to visual acuity than the visibility of a feature. According to Adler et al. [Adl+11, ch. 33] it can be distilled down to four widely accepted criteria, presented in Table 1.

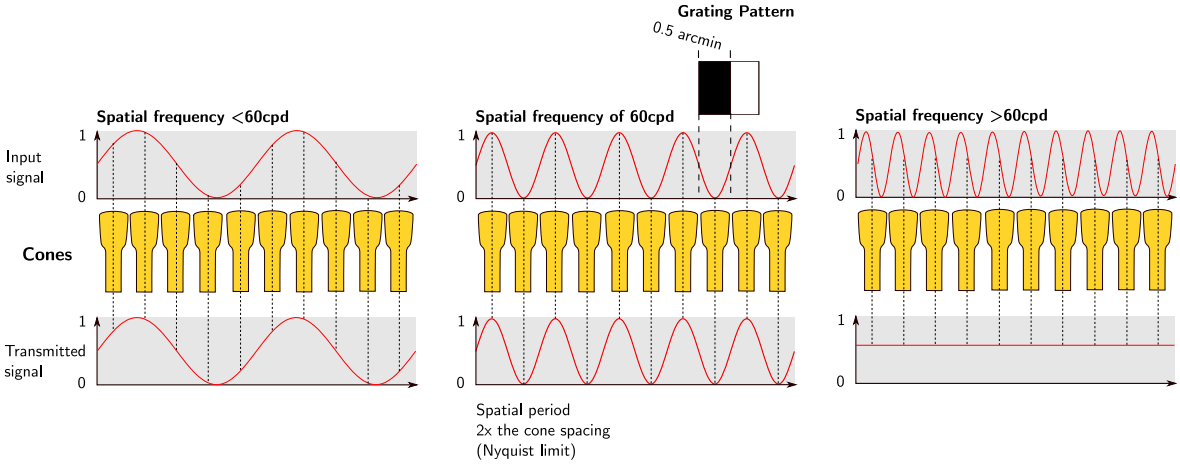


Figure 17: Spatial sampling of the cones with a frequency lower 60 cpd (left) at the Nyquist limit of 60 cpd and exceeding 60 cpd (right). For the latter case, the sampling is limited by the cone spacing for the grating patterns as the input signal cannot be reconstructed properly. *Image adapted from Adler et al. [Adl+11, p. 628]*

The acuity limits are usually measured using high contrast grating patterns, images, or letters under photopic luminance conditions, which corresponds to typical daylight or display use cases. Models to describe the spatial acuity of the eye can be found in Section 3.1.2.

Interestingly, besides the measured and modeled spatial acuity, phenomena such as *hyperacuity* transcend the intuitively given limits of the visual acuity. Judging relative positions of objects can be performed with a precision that is finer than what can be explained by the size and spacing of the cones alone. This becomes apparent when looking at the *minimum discriminable acuity* in Table 1. The misalignment of two lines can be detected with remarkably high precision (Vernier acuity), magnitudes higher than what can be explained by visual acuity. First experiments using two lines in order to improve the precision of scales date back to Pierre Vernier (⋆1580 - †1637), whose scale was used to aid ship's navigators in determining lengths and angles [Adl+11, p. 629]. The Vernier acuity plays a major role in the visibility of aliasing artifacts in digital images [Lue+03, p. 258]. Nowadays, pixel densities are available on smartphones that can easily exceed the visual acuity of the eye in their everyday use. However, humans usually prefer displays with a high number of pixels-per-inch (PPI). Images appear "crisper" and the text readability is increased. The reason for this is the eye's performance with hyperacuity. Although a model developed by Geisler [Gei84] showed that such and even higher acuities are theoretically possible by considering photon effects and the cone spacing alone, in general hyperacuity is only explainable when the pattern of the photoreceptors, and as a result, a pattern of the absorbed photons, is known to the HVS. Hence, the hyperacuity phenomena are most probably only happening due to

Type of acuity	**Measured**	**Acuity (deg.)**	
Minimum visible	Detection of a feature	0.00014°	
Minimum resolvable	Resolution of two features	0.017°	
Minimum recognizable	Identification of a feature	0.017°	C
Minimum discriminable	Discrimination of a change in a feature	0.00024°	

Table 1: Different types of visual acuity and their limits. *Table after Adler et al. [Adl+11, p. 628]*

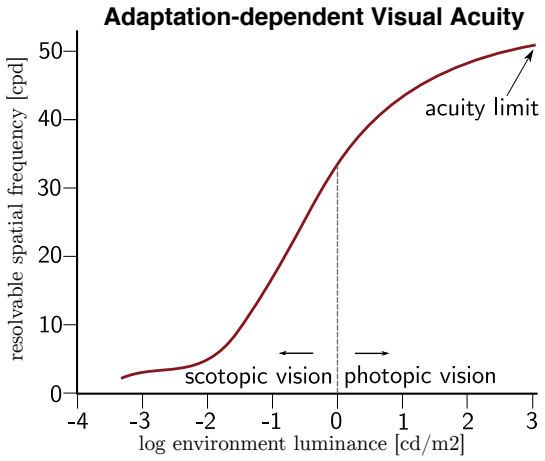


Figure 18: Adaptation-dependent acuity. Spatial acuity increases non-linearly from scotopic to photopic vision. *Image adapted from Ferwerda et al. [Fer+96]*

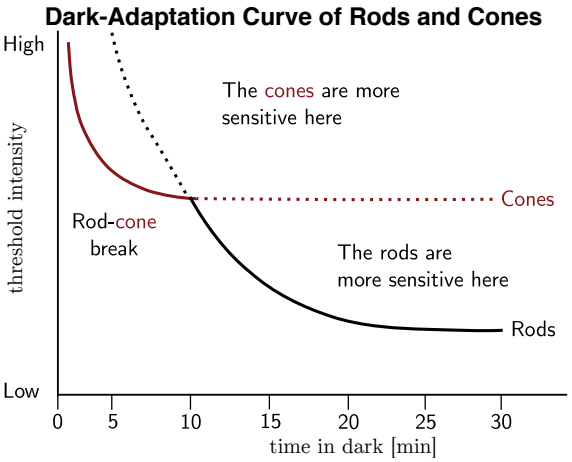


Figure 19: Dark adaptation curve. Cones recover rapidly if they are bleached, gaining their maximum sensitivity after about 10 min. in the dark. Rods recover at a much slower rate, only gaining full sensitivity after 30–40 min. *Plot from Snowden et al. [STT12, p. 33]*

sophisticated information processing in the visual cortex and possible because the cortex has an idea of the positional information of the photoreceptors [Adl+11, p. 630]. Note, that as the visual acuity, the hyperacuity performance degrades rapidly with eccentricity [LKA85; SB85]. Also, there is a fundamental difference between resolution and localization. The hyperacuity localization of individual peaks of intensities or sharp borders is accomplished with a "sub-pixel precision" by an operation utilizing output differences – not between individual contiguous photoreceptor elements within the distribution of cells activated by a single target feature, but from parameters derived from all the photosensitive elements of the activated distribution [Wes12]. When a feature is spread across several photosensitive cells, each with a gradient in the response but outputting only a single spatial value, the position of the image center can be located more exactly than what is given by the spacing of the photoreceptors alone. Likewise, a temporal change in intensities does further support localization.

Visual acuity is not only determined by the density of photoreceptors, but also by the presence of bipolar and retinal ganglion cells in sufficient numbers. Hence, further factors such as the overall lighting and contrast of the stimuli are greatly influencing acuity [BSA91]. The reduction of perceivable spatial detail under various lighting conditions is visualized in Figure 18. The highest perceivable spatial frequency of a sinusoidal grating pattern reduces from approx. 60 cpd at photopic levels down to 2 cpd for scotopic vision, illustrated in Figure 18. The eye's sensitivity to contrast can be described by a Contrast Sensitivity Function (CSF) for the spatial and temporal domain [Wan95, p. 33]. One common variant to measure perceivable contrast are sine wave patterns of changing black and white stripes whereby spatial frequency increases from left to right and contrast increases from top to bottom. The CSF is defined as the reciprocal of the smallest visible contrast expressed as a function of temporal and spatial frequencies (Figure 20). The region under the curve is commonly called *the window of visibility* [Adl+11, pp. 613–621]. The resolvable acuity limit of 60 cpd corre-

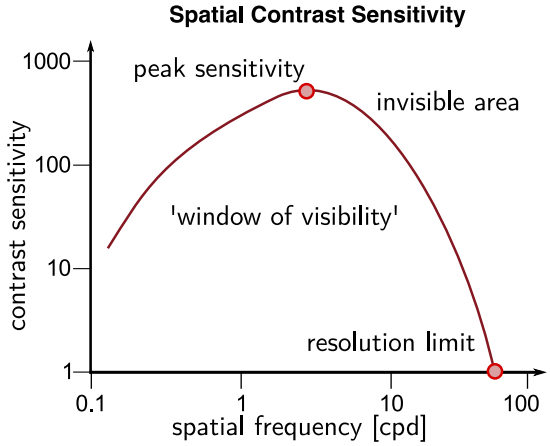


Figure 20: Spatial contrast sensitivity function (CSF). The CSF denotes the threshold contrast required for a given spatial frequency of sinusoidal pattern to be perceivable (visually detectable). All sinusoidal patterns with a contrast higher than the threshold are placed in the *window of visibility* under the CSF curve. *Image adapted from Snowden et al. [STT12, p. 115]*

sponds to the lowest contrast sensitivity value. Very high (>60 cpd) and very low frequencies (<0.1 cpd) cannot be perceived at all. As illustrated in Figure 17, cone spacing and optical filtering can explain the upper limit. However, the lower limit cannot be directly derived from the eye's physiology [Adl+11, pp. 613–621]. Contrast sensitivity depends on the number of neural cells responding to the respective grating pattern [RVN78]. From the fovea to the periphery, sensitivity decreases significantly at all frequencies. The decrease is fastest for high frequencies [RVN78]. Commonly used models for the CSF are presented in Section 3.1.3.

A property commonly exploited in graphics to increase the perceived quality of images is based on the spatial arrangement of the photosensitive cells and the resulting spectral properties. These properties suggest turning regular patterns into less perceivable high-frequency noise [Yel83; PC85; WC83]. The varying distributions of rods and cones also affect the *sensitivity to colors* in different parts of the visual field [Noo+83]. While the fovea is tuned to chromatic red/green stimuli, those stimuli are significantly less salient in the periphery. Here, the S-cones sensitive to 'blueish' wavelength dominate (Section 2.1.1). Hence, contrast sensitivity also depends on the chromaticity of the stimulus. Blue-yellow and achromatic stimuli result in a less-pronounced decrease in terms of contrast threshold [Mul85]. The sensitivity to red-green color variations decreases more steeply toward the periphery than the sensitivity to luminance or blue-yellow colors. Besides the different densities of the cones, neural processes are also of importance in this context [HPG09]. Information on perceptually-driven color models is given in Section 3.1.4.

Retinal photoreceptors can adapt to stark changes in light intensity. It enables humans to perceive visual information robustly over seven orders of magnitude of brightness intensities. *Dark adaptation* describes the change of vision from brightness to darkness. Being exposed to a very bright light, the eye can see gradually dimmer objects over time. This process is illustrated in the dark adaptation curve plotted in Figure 19. The brightness required to trigger a neural response falls for a few minutes but appears to stay constant, before it rapidly falls more. Whiles cones can respond to lighting changes rather quickly, rods take a much longer time to adapt. Highest responsiveness is reached after 30 minutes. After our eye has

been exposed to very intense illumination, the visual threshold is greatly elevated and may take tens of minutes to recover fully. Quite noticeable are afterimages that appear to be "imprinted" on the retina fainting over time. Conversely to dark adaptation, *light adaptation* describes the process of reducing the sensitivity of the HVS as light intensity increases. *Light adaptation* occurs when moving from a dark into a bright environment. At first, bright light dazzles us because the rods are set to be highly sensitive to dim lighting conditions. Now, as the eyes are stimulated by intense lighting, the rods and cones are stimulated. The currently prevailing high levels of the photopigment (rhodopsin and photopsin) are now broken down, leading to a saturation of the neural signals. This results in the glare and dazzle. Adaptation influences the performance of the HVS, such as color perception, spatio-temporal contrast sensitivity and the amount of perceivable detail [LSC04].

As the sensitivity of the retina decreases, the retinal neurons undergo rapid adaptation, inhibiting rod function in favor of the cone system. Within one minute the cones are sufficiently excited by the bright light to take over fully. Both visual accuracy and color vision continue to improve over a range of five to ten minutes. Adaptation comes at the expense of a reduced acuity at lower light levels. During daytime, contrast sensitivity is lower but visual acuity and color vision excels. However, besides counteracting the stimuli saturation, light adaptation must also be considered as a way to provide the best possible performance at a particular level of illumination [Adl+11, p. 429]. Commonly used models to describe adaptation and Tone Mapping Operators (TMOs) that implement these for High Dynamic Range (HDR) imaging, are presented in Section 3.1.5.

Similar to the drop in acuity with the eccentricity that can be observed in stereopsis, depth perception is significantly reduced in the periphery [PR98]. Lastly, acuity is greatly influenced by the motion of a stimulus. Objects appear to blur if they move quickly along the visual field. While this loss in acuity can to some degree be attributed to precision issues caused by our limited ocular motility precision [Mur78], studies by Tyler and Nakayama [Tyl85; K90] give strong evidence that the photoreceptors themselves limit our sensitivity to temporal details.

2.2.3 *Motor*

In order to explore and scan the environment by shifting attention from one OoI to another, our eyes are constantly moving. Likewise, accommodation does adapt the eye's lens in order to adjust and to set the OoI into focus. The primary goal of moving the eyes is to stabilize and move the projection of the OoI onto the macula so that the focused object is perceived with high detail. The most important types of motion are the *vestibular-ocular reflex*, *saccades*, *Smooth Pursuit Eye Motion (SPEM)*, and coupled *vergence-accommodation* motion. A survey on the properties and implications of human eye motion is provided by Kowler [Kow11]. An overview of the angular velocities and durations of the various types of movements is presented in Table 2.

The *vestibular-ocular reflex* uses acceleration information from the vestibular system, the orientation of the head and the retinal velocity (optic flow) in order to keep the orientation of the eyes aligned with the OoI. This process is fairly robust even for fast head movements and is quick with a latency of 7-15 ms [Adl+11, p. 222]. *Saccades* denote the motion when rapidly jumping from one OoI to another. A saccade can reach peak angular velocities of up to $900°/s$ and can last for several milliseconds. However, even at lower angular velocities

there is a dramatic decline in visual acuity during the movement. This is known as *saccadic suppression* [Vol+78; WDW99; Ros+01]. Generally, increasing the retinal velocity does limit the visual acuity (Section 3.1.2). When viewing a typical natural scene, the HVS triggers around 2 to 3 saccades per second. The spacing between fixations is, on average, 7° viewing angle. Fixating objects at larger eccentricities (> 30°) is highly uncomfortable and humans will likely start turning the head to doing so. While freely moving the head, saccades and the vestibular-ocular reflex are reported to commonly result in motions less than 15° around the normal line of sight [BAS75; Bar79]. Hence, a Comfortable Viewing Angle (CVA) is considered to be identical (approx. 15°) [Def99, p. 17].

In contrast to saccades, *fixation* describes the process in which visual information is perceived while our gaze is mostly at rest, fixated, and focused on an OoI. Fixation durations typically vary between 100 ms and 200 ms (Table 2) but are seldom reported to be as long as 1.5 seconds [WDW99, p. 72]. The duration is assumed to correspond directly to the complexity of the visual input. Also, during fixations, the eyes are not completely motionless. They perform tiny but important movements known as *tremor motion*. If tremor motion is inhibited, the perceived image fades away [Adl+11]. Hence, this unconscious motion is highly important to refresh the retinal image.

Type	**Duration (ms)**	**Amplitude** ($1° = 60'$)	**Velocity**
Fixation	200-300	-	-
Saccade	30-80	$4 - 20°$	$30 - 500°/s$ (Peak $900°/s$)
Smooth Pursuit	variable	variable	$10 - 30°/s$ (Peak $100°/s$)
Vergence	300-600	-	-
Microsaccade	10-30	$10 - 40'$	$15 - 50°/s$
Tremor	10	$< 1'$	$20'/s$
Drift	200-1000	$10 - 60'$	$6 - 25'/s$

Table 2: Velocities and durations of typical eye motions.

SPEM is the unconsciously triggered tracking reflex when a moving object attracts our attention. This motion enables humans to track slow-moving targets in order to fixate the object onto the macula. Interestingly, small eye movements up to $2.5°/s$ have hardly any effect on visual acuity [Adl+11, p. 9]. Researchers have found that the peak velocity for SPEM is $100°/s$ [WDW99, p. 148]. However, the success rate depends on the speed of the target and decreases significantly for angular velocities $> 30°/s$. The increased visual acuity when objects are moving does provide an explanation for the results of the user experiments presented in Chapter 6.

Creating a clear and focused image on the retina is driven by two main processes: Firstly, the *vergence* movement, which is the rotation of the eyeballs in opposite directions in order to fuse a focused object into a single percept and secondly, the *accommodation*, adjusting the power of the eye's lens to create a sharp retinal image of the fixated object. This way, accommodation describes the natural counterpart of adjusting a camera lens so that an object in the scene is set into focus. Importantly this process happens unconsciously and without any effort in less than a second at photopic illumination levels [Gol10, p. 289]. The maximally allowed change a user can still keep focus on an object is reported to be approx. 0.7 m/sec and the range of accommodation is considered to be approx. 0.2 m - 6 m for a healthy adult [TM89]. However, the speed and range of accommodation are dependent on a variety of influences including age, visual acuity, and a wide range of physiological factors [TM89; GPB80; LS10]. Results and measurements widely vary. This constitutes a challenge

that needs to be addressed when computing Depth-of-Field (DoF) in order to filter images as presented in Chapter 7.

Vergence and *accommodation* are highly entangled in the process of stereopsis. Typically, stereoscopic displays affect vergence by providing binocular disparity cues as a separate image for each eye. As the images are shown on the screen, the eyes focus on the screen's distance. This results in a conflict, known as the *vergence-accommodation conflict* [Gol10, p. 1040]. Accommodation and vergence motions are coupled with the fixation process for binocular vision so that both eyes' gaze aims at the same point in the distance. However, due to their IOD, both eyes perceive an OoI from slightly different viewpoints. The difference of the per-eye gaze directions can be quite large when looking at an object close-by. *Vergence* moves the point of intersection of both gaze lines to the point of focus and enables humans to optimize the FoV overlap for a wide range of distances. As presented in Chapter 7, measuring vergence movements can be used to compute gaze points in the 3D space.

2.2.4 *Processor*

Retinal stimuli processing is followed by neural information processing in the *visual cortex* of the brain. Analogous to the decrease in the density of rods and cones, over 30% of the primary visual cortex is responsible for the central 5° of the visual field, while the periphery is under-represented [HH91]. Perception research has targeted cognitive processing of images and perceptual differences between central and peripheral vision. A common approach to neural processing that corresponds with the neural design of the visual cortex is the multi-channel model. It was inspired by work from Entroth-Cugell as well as Campbell and Robson [ER66; CR68]. When walking through a forest, for example, the silhouettes of the trees provide coarse information. From here, we can focus on individual trees or single leaves. The multichannel model suggests that the visual system extracts all of these different scales of information from a scene simultaneously, but analyses each stream independently and in parallel. Later these streams are combined using the higher vision processes in order to assemble the final percept for the particular scene [Lue+03, p. 247-248]. However, experts disagree on the exact number of channels, as can be seen in work by Caelli and Moraglia [CM85] and Heeley [Hee91], and visual acuity is eccentricity-dependent. Hence, other authors have pointed out the importance of features in the peripheral vision for perception and scene understanding.

Thorpe et al. [Tho+01] have shown that peripheral vision provides a rich source of information, crucial to the perception and recognition of features, objects, and animals. Gilchrist et al. [To+11] point out that the influence of color changes in the periphery is greater than that of orientation changes. Furthermore, the HVS makes extensive use of contextual information from peripheral vision, facilitating object search in natural scenes [Kis+14]. During this process, preprocessing of visual stimuli probably occurs. There is evidence that basic visual features, such as object size, color, and orientation, are pre-processed before actual attention is placed on the object by moving it into central vision [WB97]. Hence, humans may be aware of certain aspects of the scene content (shapeless bundles of basic features) in the periphery but have to pay attention to the shape in order to recognize its form and all its features.

Besides the process of stereopsis, the ability to interpret depth cues in the visual input to improve stereo vision and the sense of spatial localization is deeply rooted in the visual cortex. Depth cues can be static (e.g., occlusion, perspective foreshortening, texture and shading

gradients, shadows, and aerial perspective) or dynamic (e.g. motion parallax). Cues can also be obtained from memory, such as for relative sizes of familiar objects [Gol13, ch. 10][Pal99, ch. 5.5] (Section 2.2.5). Moreover, depth cues are dependent on the object's distance to the eye and dominant cues may prevail or compromise 3D scene interpretation [Did+11]. Some considerations on the question of if DoF does influence depth perception is presented in Chapter 7.

A phenomenon that can only be observed with peripheral vision is known as *crowding*. Objects become more problematic to recognize (rather than to detect) when distracting stimuli surround them. Crowding is studied by using well-defined stimuli such as letters or sine wave patterns [Bou70; To+11]. The effect of crowding can also be observed for more complex content such as faces [MMP05] and complex stimuli in natural images [RLN07; PT08; BNR09].

Finally, vision is affected by cross-modal effects. In particular, Virtual Reality (VR) systems often provide non-visual cues such as audio, vibration, or even smell. These effects have been studied in psychological experiments on various interplays between cues [SS01; Pai05; SS03; WP04]. When sensory channels are substituted or combined, several implications occur: These channels are no longer seen as separate but may affect each other through integration of sensory signals inside multimodal association areas in the brain [Sut02, p. 36–64][Pai05][LN07]. As yet, cross-modal effects are not fully understood. The research on multisensory factors still needs to be continued in order to fully understand its importance, but various cross-modal effects can already be identified. Vision plays an important role for the bias of stimuli, since it predominantly alters other modalities [LTJ86; Nar+10]. Sound, on the other hand, alters the temporal, but also other aspects of vision, like those that affect disambiguation [SSL97]. Finally, tactility may alter vision, but may also be influenced itself by audio [BSJ04; Bre+05]. Hence, theoretically, other modalities could be used to further alter perception and as a consequence optimize visual representation. [Wei+17]

2.2.5 *Memory and Attention*

The processing of visual information is highly dependent on knowledge and patterns, stored in the memory. How this knowledge is stored, is still being discussed. According to Smith and Kosslyn [SK13b], a representation is a physical state that stands for an object, event or concept, and must be constructed intentionally to carry information. Representations may encode information in different forms, including those similar to images or feature records, but also amodal symbols, and statistical patterns in neural networks. These representations are connected: Different formats here work together to represent and simulate objects [SK13b]. Moreover, the brain preserves certain features of the retinal image over time, despite their motion and potentially varying occlusions [Yan95; LE13]. As previously noted, there is also evidence that basic visual features such as color, size, and orientation are parsed and preprocessed before the central gaze is directed in that direction.

While attention is still not fully understood, research indicates that it has three components: orienting to sensory events, detecting signals for focused processing as well as maintaining a vigilant or alert state [PB71]. Attention is essential for processing visual stimuli and search behavior [TG80]. It can occur in information-processing tasks in various ways [WC97]: *selective attention* is the choice of which events or stimuli to process; *focused attention* is the

effort to maintain processing of these elements whilst avoiding distraction from other events or stimuli; *divided attention* is the ability to process more than one event or stimulus at a given point in time.

The focus of attention also affects perception on a cognitive level. A critical perceptual effect for certain tasks is the effect of *cognitive tunneling* (or *visual tunneling*) and *inattentional blindness* [Miu86]. Observers tend to focus attention on information from specific areas or objects. However, a strong cognitive focus on specific objects leads to an exclusion/loss of information for areas in the periphery of highly-attended regions. Several studies conducted by Thomas et al. [TW01] are concerned with the detection of perceptual differences and show the effects of visual tunneling during terrain rendering. Further studies show the same effects as a dramatic decrease in the size of the visual field and the loss of information in the user's peripheral vision [TW06; WA09; LMS10]. A user study for this thesis, presented in Chapter 6, provides evidence of the presence of visual tunneling under certain experimental conditions.

MODELS FOR VISION AND VISUAL PERCEPTION

A Review for Computer Graphics

3

> *. . . all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind. . .*
>
> George E.P. Box and Norman R. Draper [BD07, p. 414]

Perceptual models are commonly used in computer graphics to approximate functions and properties of the Human Visual System (HVS) using mathematical descriptions. These models steer perceptual rendering algorithms and are able to judge the perceptual quality of images. In this chapter, a selection of models for the HVS and perceptual processes are presented that are relevant for computer-graphics applications. These models can be either low-level: only describing certain aspects of the HVS including early processes, or high-level: describing an entire system. For the latter, a focus is placed on perception-driven image metrics that model the entire visual pipeline to judge image and video quality in order to detect visual artifacts. Lastly, attentional processes have been modeled to explain which parts of the scene trigger our attention and how this is carried out.

CONTRIBUTIONS BY THE AUTHOR This chapter is a reproduction of our state-of-the-art report

> Martin Weier et al. "*Perception-driven Accelerated Rendering.*" In: *Computer Graphics Forum (Proceedings of Eurographics)* 36.2 (Apr. 2017).

and follows this reports general structure. As in the state-of-the-art report, I made major contributions in Section 3.1 and Section 3.2. However, especially the latter section was significantly revised by various co-authors of the report. In this thesis, I updated Section 3.2 again in order to include the most relevant related work and broaden the basis of the discussion by providing more insights into the most important approaches. Section 3.1 has been extensively updated. The subsection on "Optical Properties and Aperture" is new. The subsections "Spatial Acuity" and "Brightness and Contrast Sensitivity" have been largely extended and rewritten. The explanations in Section 3.3 must be acknowledged to Michael Stengel, a co-author of the state-of-the-art report. Hence, only an abstract of the section in the state-of-the-art report is found in this thesis. Also, I gave credit to Michael in this abstract.

3.1 LOW-LEVEL MODELS

Low-level models target particular aspects of the HVS. The following sections present the most commonly used low-level models to describe the optical apparatus, the spatial acuity and temporal resolution along with models for brightness, contrast, color sensitivities, adaptation as well as visual masking.

3.1.1 *Optical Properties and Aperture*

Various researchers have modeled the eye as an optical system. Although, concrete numbers differ, it can be assumed that an average healthy young adult has a focal range from roughly +56 D to +70 D with a near point of 0.2 m [MC04] [LS10] [Ed17, ch. 16]. Usually the distance from the lens to the retina is assumed to be between 170 mm and 230 mm. The first know sources that discussed the optical properties of the eyes date back to the Greek physicist Galen (⋆130 - †210 a.D.) [KK95]. Much later, Johannes Kepler (⋆1571 - †1630) was one of the first to detail the role of the crystalline lens and cornea. Points in space were imaged onto the retina to form an inverted version. Two centuries later Listing (⋆1808 - †1882) developed a first schematic eye model. This model was later refined by Gullstrand (⋆1862 - †1930) based on the known works on the eye and optical instruments by Helmholtz (⋆1821 - †1894) [Hel24]. This model became known as *Gullstrand's schematic eye* and is still widely used for education and in academia. A description of this eye model and its parameters can be found in the book by Katz and Kruger [KK95]. In the subsequent years, numerous more modern versions have been developed in order to better account for various conditions such as spectacle wear, contact lenses and refractive surgical procedures [LB97]. Unfortunately, these eye models are rather complex. Gullstrand and the other modern schematic eye models often contain six or more refracting surfaces. However, calculations can be greatly simplified by treating the eye as a black box. Donders (⋆1818 - †1889) made an early attempt on such a simple model [Khu08, p. 34]. His *reduced eye* model replaces the several refracting bodies of the eye with an ideal single lens. However, Donders took great libertine in rounding numbers. A bit more advanced version of his model for simple calculations, e.g. to determine the object-image relationships, can be derived by adapting the cardinal points [KK95]. This reduced eye has an optical power of 58.2 D with a nodal point of 17.2 mm (Figure 21). Using this model, the retinal image size i can then be computed by $i = 17.2\,\text{mm} \cdot \phi$, with ϕ being the height of the object in radians of the visual angle.

Unfortunately, all, even the complex eye models, mostly fail to describe the processes that are involved when accommodating the lens. Often, these models assume the eye to be relaxed. A highly complex model that can be used to describe accommodation processes is the Arizona eye model [Sch04]. Other popular modern models are the Navarro and the Liou-Brennan model. A comparison of these can be found in the work by Zoulinakis et al. [Zou+17] and the book by Artal [Ed17, ch. 16]. However, their amount of different surfaces, media with different refracting indices and complex lens shapes are too computationally complex to be evaluated inside performance critical graphics pipelines. They are still successfully used to describe and develop optical systems using ray or wavefront-based optical evaluation. More on schematic eye models can be found in the work by Katz and Kruger [KK95] and the book by Artal [Ed17, ch. 16]

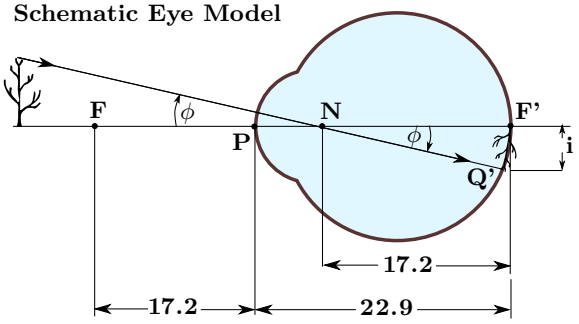


Figure 21: A reduced eye model with a single nodal point N. The model allows for computing the retinal image size. *Illustration redrawn from Katz and Kruger [KK95]*

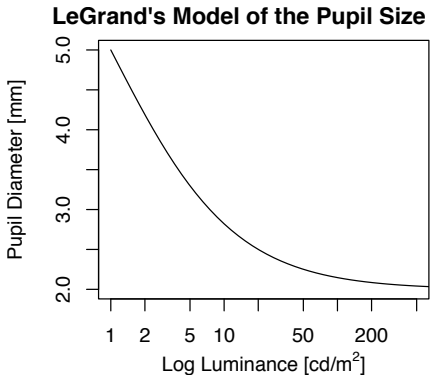


Figure 22: Plot of Le Grand's model (Equation (1)) to estimate the pupil size based on the incident luminance.

Another approach to model the optical properties of the eye is by determining an optical Modulation Transfer Function (MTF). These functions are usually developed by using psychophysical experiments. Here, linear systems theory provides analytical tools for assessing transformations that systems make between inputs and outputs [Gol01, p. 59]. Often the experiments make use of studies showing varying sinusoidal grating patterns to human subjects. These patterns can be used as fundamental frequencies to derive higher level harmonics by applying Fourier theory. *"Thus, for lenses, the MTF can be determined by measuring the transfer of contrast at each spatial frequency."* [Gol01, p. 59] While these models commonly describe the entire HVS and not only the pre-retinal optics, they can be derived from here. Early models of the pre-retinal optics were introduced by Campbell and Green [CG65]. They measured the MTF of the entire HVS with gratings patterns on a CRT. A pre-retinal MTF was then derived by comparing results with measurements when the optics were bypassed by using interference fringes [Gol01, p. 60]. Other models, such as the one by Williams et al. [Wil+94], implement various improvements and use two-pass measurements in order to reduce errors. Barten [Bar99, pp. 27-29] presents an MTF using a simple Gaussian description. This model is used to represent the eye's sensitivity to contrasts, introduced in Section 3.1.3. Although denoted as optical MTF, Barten states, that it does not only include the optical influences but also takes the effects of the stray light in the ocular media, the diffusion in the retina, and the discrete structure of the photoreceptors on the retina into consideration. More on MTF models can be found in the book by Artal [Ed17, pp. 182, 319].

Another part that can be modeled is the aperture, i.e. the pupil. A commonly used model to determine the pupil size was introduced by Le Grand [Gra69, p. 99]. Given the average luminance of the image in cd/m^2, the pupil diameter d in mm can be estimated with

$$d = 5 - 3\tanh(0.4 \log L) \tag{1}$$

A plot of this function is shown in Figure 22. However, if eye trackers are available, the pupil size can often be measured directly.

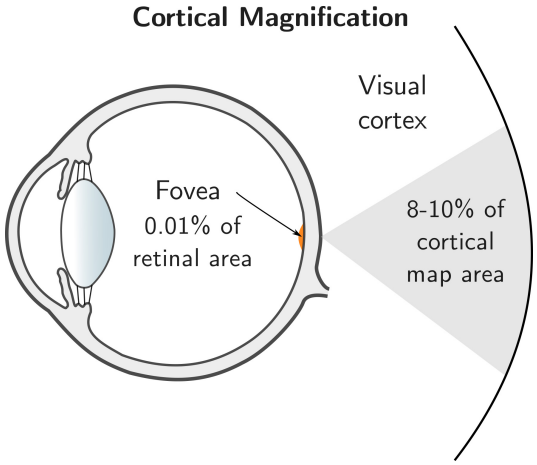


Figure 23: Cortical magnification. The cortical magnification maps the small area of the fovea to a much larger area on the visual cortex. *Image adapted from Goldstein [Gol10, p. 82]*

3.1.2 *Spatial Acuity*

One important aspect of vision is the acuity of the eye. Acuity models are often used to adapt image resolution, sampling patterns and rendering quality based on the user's gaze. Strasburger et al. [SRJ11] provide a historical summary and survey describing how the visual acuity drops for peripheral vision. Green [Gre70] shows that acuity differs from what may be expected from physical cone spacing at eccentricities $> 2°$. Weymouth et al. [Wey58] determine that visual acuity decreases roughly linearly with eccentricity for the first $20° - 30°$. Visual performance decreases more rapidly for higher eccentricities [FWK63]. According to Weymouth [Wey58] and Strasburger et al. [SRJ11, ch. 3], when measured in terms of a Minimum Angle of Resolution (MAR) rather than acuity (its reciprocal), a linear model matches both anatomical data (such as receptor density) and performance results on many low-level vision tasks. Nonetheless, the slope of the MAR function is user-dependent and cannot be precisely predicted [SRJ11, p. 19]. Acuity is also affected by eye adaptation in very bright and dark areas, by eye motion as well as cognitive factors [Gol10, p. 60]. Thus, any linear model remains an approximation.

The model by Guenter et al. [Gue+12] describes the MAR ω in degrees per cycle as

$$\omega(e) = me + \omega_0$$

The factor m is the slope of the MAR, e the eccentricity, and ω_0 the smallest resolvable angle in the fovea. Given 60 Cycle per Degree (cpd) as the upper limit for resolvable details, ω_0 can be computed as $\omega_0 = 1/60$. Finally, the visual acuity in cpd can be expressed by $M(e) = \omega(e)^{-1}$. A plot of the function is illustrated in Figure 24a. Generally, the foveal visual acuity in healthy, non-elderly adults (with corrected-to-normal vision) substantially exceeds 20/20 on a Snellen chart, which equals 30 cpd.

The Snellen chart, developed by Hermann Snellen (⋆1834 - †1908), is used to determine the visual acuity using letters on a grid. Visual acuity was quantified by Snellen comparing the vision of a patient with that of his assistant, who had good vision [Adl+11, p. 9]. Patients

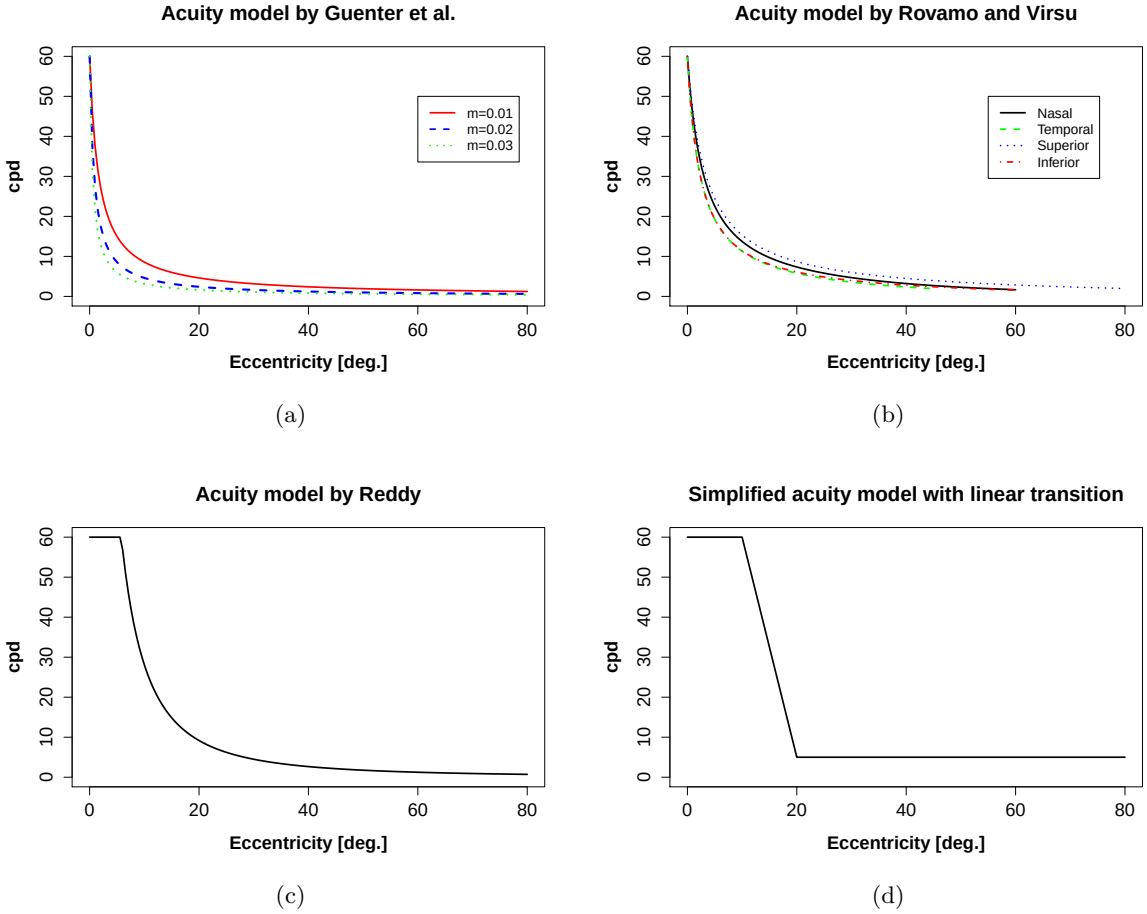


Figure 24: Different models of the retinal visual acuity. The model described by Guenter et al. [Gue+12], (a) is a hyperbolic function of the visual acuity. When considering its reciprocal, the Minimum Angle of Resolution (MAR), it becomes a linear function. The model by Rovamo and Virsu [RV79] (b) uses additional quadratic terms to express the MAR. Hence, the hyperbolic visual acuity is a bit flatter/less bulbous. Reddy (c) discusses a simplified model in the context of computer graphics. We often rely on an even simpler but better parameterizable linear approximation (d). All of these plots assume a maximal baseline acuity of 60 cpd.

must identify the differently scaled letters on the chart. This makes it possible to determine the Snellen fraction S [Sch04, p. 19]:

$$S = \frac{\text{Greatest distance subject can just read a given line on the chart}}{\text{Greatest distance a "normal" observer can just read the same line}}$$

"... 20/200 (6/60) vision meant that the patient could see at 20 feet (6 m) what Snellen's assistant could see at 200 feet (60 m). ... The essence of correct identification of the letters on the Snellen chart is to see the clear spaces between the black elements of the letter. The spacing between the bars of the letter "E" should be 1 minute for the 20/20 (6/6) letter. [Adl+11, p. 9]" The other letters for the different Snellen fraction can be computed using simple trigonometric functions. However, the size change of the letters between two rows for different Snellen fractions might not be constant. Also, the original chart has a different number of letters on each line for each Snellen score. In order to overcome these drawbacks, more elaborate eye charts such as the Bailey-Lovie [BL13] and ETDRS chart [Kai09] are used

in practice today. Both have a logarithmic reduction in letter size per row and a constant number of letters on each line [Sch04, p. 19]. Still, all are used to determine a Snellen fraction.

Generally, acuity does not exceed a Snellen score of 20/10 [Gue+12; Col01]. Guenter et al. assume an average acuity for ω_0 in between the 20/20 and 20/10 foveal acuities. They found 48 cpd, i.e. a 20/12.5 on the Snellen chart, to be a good estimate of the foveal acuity for adults below 50. Hence, for practical uses they assume $\omega_0 = 1/48$. According to Guenter et al. [Gue+12] such a simple linear model works well for "central" vision (angular radius $< 8°$). For peripheral vision, the MAR rises more steeply. Hence, models with additional quadratic terms have been used. Rovamo and Virsu [RV79] model the fall-off as

$$\begin{aligned}
\text{Nasal: } M_N(e) &= (1 + 0.33e + 0.00007e^3)^{-1} && (0 \leq e \leq 60°)\\
\text{Superior: } M_S(e) &= (1 + 0.42e + 0.00012e^3)^{-1} && (0 \leq e \leq 45°)\\
\text{Temporal: } M_T(e) &= (1 + 0.29e + 0.000012e^3)^{-1} && (0 \leq e \leq 80°)\\
\text{Inferior: } M_I(e) &= (1 + 0.42e + 0.000055e^3)^{-1} && (0 \leq e \leq 60°)
\end{aligned}$$

This model is plotted in Figure 24b. For the plot these functions were scaled using a base acuity of 60 cpd. The model has later been adapted by Reddy [Red01] to better suit rendering needs.

$$M(e) = \begin{cases} 1 & e \leq 5.79° \\ 7.49/(0.3e + 1)^2 & e > 5.79° \end{cases}$$

The model is illustrated in Figure 24c. It has likewise been scaled using 60 cpd as maximal acuity. This model assumes a symmetrical radial fall-off, ignoring differences between principal half meridians in the visual field (Nasal, Superior, Temporal, and Inferior). Moreover, it explicitly creates a region where visual acuity is assumed to be maximal. Most recently, this model has been transformed into a Probability Density Function (PDF) to guide sample generation in stochastic rendering processes [Kos+17] (Section 4.2.3). However, all of the aforementioned models do not consider higher-level influences to acuity such as the retinal velocities. A simple method to incorporate velocities has been introduced by Reddy [Red97; Red01]. The velocity-dependent maximum acuity is given by:

$$G(v) = \begin{cases} 60.0 & \text{where } v \leq 0.825°/s \\ 57.69 - 27.78 \log_{10}(v) & \text{where } 118.3 \geq v > 0.825°/s \\ 0.1 & \text{where } v > 118.3°/s \end{cases}$$

Finally, visual acuity at a velocity v and eccentricity e can be computed by $H(v, e) = G(v)M(e)$. A plot of Reddy's model at various retinal velocities is presented in Figure 25.

In this thesis often an even more straightforward linear model is assumed for peripheral vision, entirely without a hyperbolic fall-off, for two main reasons: First, reconstruction and the reduction of temporal artifacts in the periphery is critical for perception. Improving reconstruction and counteracting artifacts works better at slightly increased sampling rates. Secondly, a linear model can be parameterized more efficiently, and sampling rates and quality can be controlled more intuitively. Figure 24d shows an illustration of this model as it is used for the systems presented in Chapter 6 and Chapter 7.

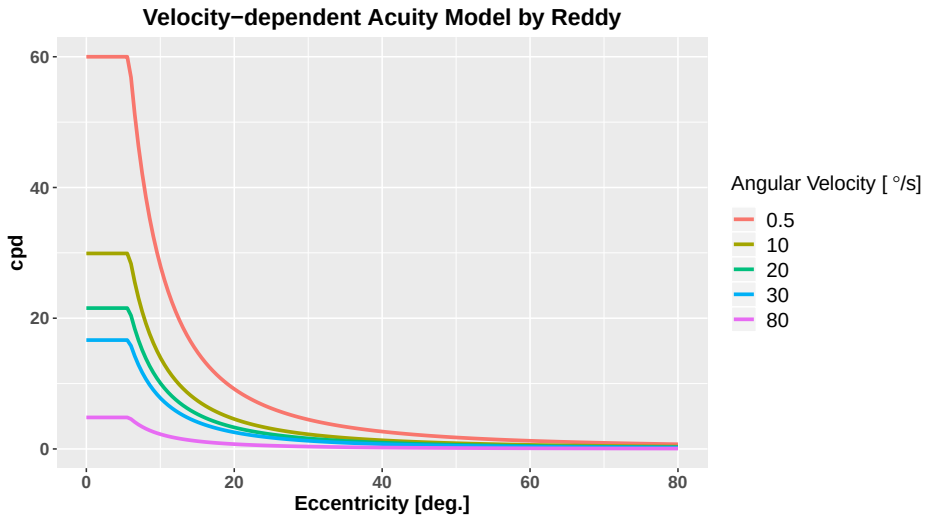


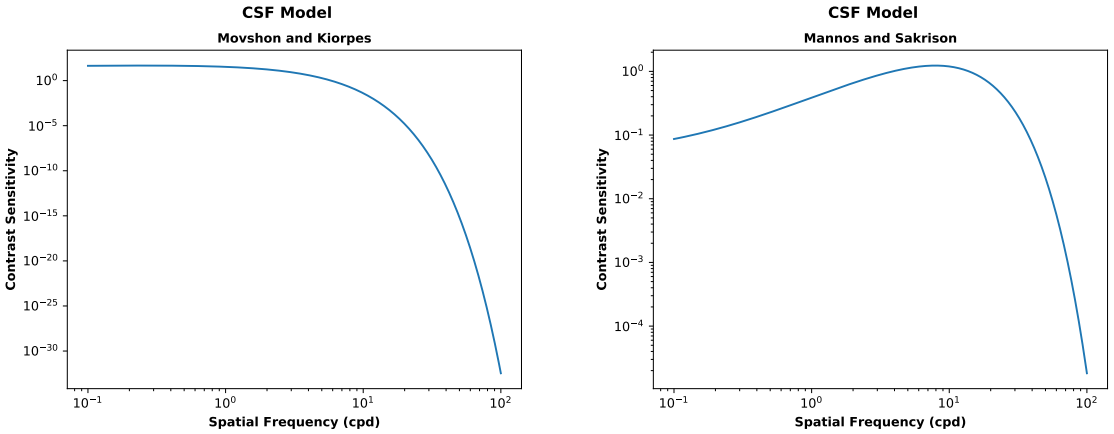
Figure 25: Reddy's acuity model extended with a model for retinal velocity.

A motivation for the often-used linear model has been provided through the concept of *cortical magnification* by Whitteridge and Daniel [DW61], and Cowey and Rolls [CR74]. According to them, a magnification factor M can represent a mapping from the visual angle to a cortical diameter in millimeters (Figure 23). The Cortical Magnification Factor (CMF) M is largest for those areas corresponding to the fovea and decreases with eccentricity for peripheral areas.

Resulting in the linear CMF, the M-scaling hypothesis claims that performance degradation with eccentricity can be canceled out by applying spatial scaling to stimuli. For example, in order to compensate for the loss in acuity when attempting to read letters in the periphery, these letters have to be enlarged by scaling them with the linear CMF to be equally readable again. This method has been successfully demonstrated by Cowey and Rolls [CR74] and motivated researchers to unify fovea and periphery [SRJ11, ch. 3.1]. Moreover, prior studies indicate a strong relationship between the CMF and the eye's contrast sensitivity, as discussed in the next section. Horton et al. [HH91] calculate the CMF factor for any given eccentricity:

$$M(e) = \frac{A}{e + e_2} \tag{2}$$

A is the cortical scaling factor, and e_2 is the eccentricity at which a stimulus subtends half the cortical distance that it subtends in the fovea [Swa+16]. Horton et al. [HH91] found the values $A = 17.3\,\text{mm}$, $e_2 = 0.75°$ to be a good fit. Dougherty et al. [Dou+03] found $A = 29.2\,\text{mm}$ and $e_2 = 3.67°$. Such strong supporters of the M-scaling concept claim that *"...a picture can be made equally visible at any eccentricity by scaling its size by the magnification factor."* [RVN78, p. 56] However, other researchers have pointed out difficulties of the M-scaling concept [WM78; BKM05]. First, the linear CMF model only approximates the complexity of the HVS, as peripheral vision is not a scaled-down version of foveal vision [BKM05]. Second, several studies exist in which the CMF concept is less convincing or fails, such as stereo acuity, two-point separation in the far periphery or contrast sensitivity for scotopic vision [SRJ11, ch. 3.4]. Besides, due to variations in the measurements for different visual tasks and to inter-individual differences, it is still an open question whether M-scaling is also applicable to near-foveal regions [SRJ11, p. 10].



(a) (b)

Figure 26: Simple CSF models by (a) Movshon and Kiorpes [MK88] (b) and Mannos and Sakrison [MS74].

3.1.3 *Brightness and Contrast Sensitivity*

Visual acuity usually is assessed under optimal lighting conditions. It is a measure of size and does not consider the contrast of a target. Sensitivity to the spatial contrast of the HVS is expressed by a Contrast Sensitivity Function (CSF) [Sch56; AL73; AET96]. An early approach to model the CSF was presented by Movshon and Kiorpes [MK88]. They design an empirical function that can be fitted to different measured CSF models.

$$csf_{movshon}(f) = a \cdot f^b \cdot e^{-c \cdot f}$$

Here f is the spatial frequency of the visual stimulus in cpd. A plot of this function employing widely used empirically determined values for $a = 75$, $b = 0.2$, and $c = 0.8$ is presented in Figure 26a. Mannos and Sakrison [MS74] propose a practical model that works well to describe achromatic and chromatic contrast sensitivity.

$$csf_{mannos}(f) = 2.6 \cdot (0.0192 + 0.144f)\, e^{-(0.114f)^{1.1}}$$

Again f is the spatial frequency of the visual stimulus in cpd. A plot of the function is presented in Figure 26b. Although the above approaches have been successfully used in various fields, they often over-estimate the actual contrast sensitivity. A complete model of the CSF also depends on influences such as eccentricity, temporal effects, and retinal velocity, making it a function of a higher order.

Hence, various CSF models with various degrees of freedom exist in the literature. One approach to develop a model is to fit functions to measurements. Gervais et al. [GHR84], for example fit splines to psychophysical data. Daly et al. [Dal98] use precomputed CSF data for a specific illumination level and support spatial frequency and retinal velocities. However, closed form solutions such as mathematical models can be evaluated more efficiently and require less data at runtime.

A renowned elaborate mathematical model was developed by Barten [Bar99]. This model allows accounting for various influence such as eccentricity and retinal illumination. Therefore,

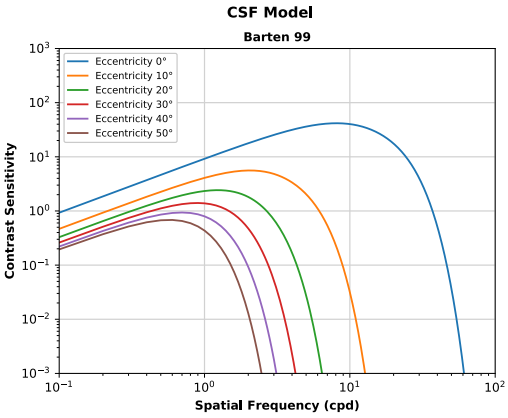


Figure 27: A model of Barten's CSF [Bar99] at a luminance level of 150 cd/m^2 at various eccentricities.

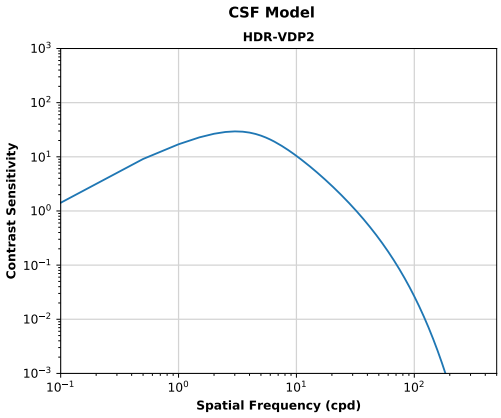


Figure 28: A plot of the CSF introduced by Mantiuk et al.[Man+11] at a luminance level of 150 cd/m^2 at various eccentricities.

it uses various other mathematical models, such as for the pupil's diameter (Section 3.1.1), an optical MTF (Section 3.1.1) and model for photon and neural noise as well as lateral inhibition. Barten also extends the model to the temporal domain. Although a complete description of the model is beyond the scope of this thesis, a plot of this CSF for different eccentricities is shown in Figure 27.

Barten's model has shown its potential to match the results of several measurements of the CSF from literature. Unfortunately, the model does not appear to work well for graphics applications – at least for inspecting static images [Man+11]. Mantiuk et al. [Man+11] tested Daly's [Dal98] and Barten's [Bar99] model in their visual difference metric HDR-VDP2 (Section 3.2.1). Unfortunately, they could not fit these models against their experimental data. Daly and Barten are not likely less accurate – both have shown their validity matching many CSF measurements from experiments – but rather that their functions may capture conditions that are different from visual inspection of static images [Man+11]. Similarly, an attempt to fit Barten's model to the model of Mantiuk was made in this thesis in order to get an idea of how the CSF would behave when adapting to eccentricity. Initially, the idea was to get a parameter set for Barten, based on the model by Mantiuk et al., that can than in turn be used to vary the eccentricity parameter in Barten's model. However, even with optimizing parameters not respecting their physical and physiological constraints, the fit remains poor. Mantiuk's model is given as:

$$csf_{mantiuk}(f) = p_4 S_a(l) \frac{\mathbf{MTF}(f)}{\sqrt{(1 + (p_1 f)^{p_2}) \cdot (1 - e^{-(f/7)^2})^{p_3}}}$$

Here, $MTF(f)$ is a fit to an experimentally determined MTF of the eye, S_a is the joint luminance-sensitivity curve for cone and rod photo-receptors and p_1 - p_4 are experimentally determined parameters. These parameters were fitted separately for each adaptation luminance L level. Values in between luminance levels are linearly interpolated. A plot of the model by Mantiuk et al. at a luminance level of 150 cd/m^2 is shown in Figure 28. Please note the difference to Barten's model as illustrated in Figure 27 at an approximately equal luminance level. It becomes apparent that more contrast for high spatial frequencies is detectable according to the model by Mantiuk et al.

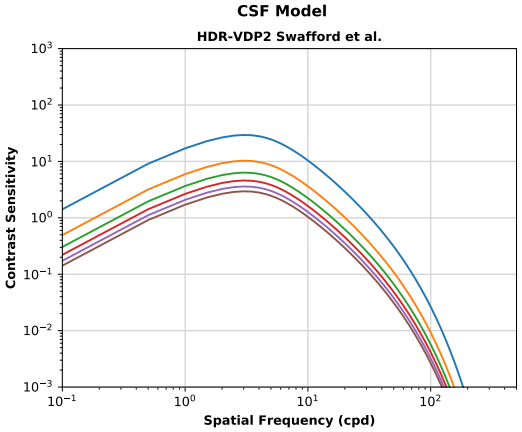


Figure 29: A plot of the CSF introduced by Swafford et al. [Swa+16] at a luminance level of $150\,cd/m^2$ with $\alpha = 0.5$ at various eccentricities.

It can be argued, that an explanation for the behavior of the model by Mantiuk et al. is deeply rooted in perception. Contrast sensitivity is commonly measured using sine-wave patterns. While those contain several complete periodic cycles of contrast, computer graphics is generating images with aperiodic regions of detail. However, the aperiodicity and the shown number of cycles does substantially influence contrast sensitivity. For single cycle-patterns and low spatial frequencies, this can yield a reduction of sensitivity by up to 60% [Lue+03, p. 254].

Besides its original use in the image metric HDR-VDP2 (Section 3.2.1), Mantiuk's model is broadly used for other image quality assessment tasks [Val+14; ASG15] and Tone Mapping Operators (TMOs) [Yan+12; Her+12; PM13; EMU15]. Unfortunately, Mantiuk's model cannot predict the CSF at different eccentricities. Therefore, a simple adaptation of the model that can be parameterized with eccentricity was developed by Swafford et al. [Swa+16]. As there is a strong relationship between the M-scaling concept and the degradation of contrast sensitivity, the idea is to scale Mantiuk's model by the CMF. The model can be described as follows:

$$csf^{e}_{swafford}(f) = csf_{mantiuk}(f) - csf_{mantiuk}(f) \cdot \left(1 - \frac{M(e)}{M(0)}\right)^{1+\alpha}$$

The contrast at an eccentricity e is given by down-scaling the CSF by Mantiuk et al. This is expressed using the CMF at the center $M(0)$ and at the eccentricity in question $M(e)$ (Equation (2), previous section). The parameter α is a tunable parameter to attenuate the influence of the eccentricity. Looking at the plot in Figure 29 makes apparent, that, though contrast sensitivity is decreased overall, it probably does not capture the loss of high-level contrasts at high eccentricities. While Swafford et al. report that this model is also adapted to support HDR-VDP2 multi-scale decomposition, Swafford et al. did not perform a user study, as such its validity is hardly determinable. Hence, it is probably worthwhile to develop other models for image assessment tasks.

The eye's sensitivity of contrast is tightly linked to visual acuity and both are important for many graphical applications. The HVS is highly sensitive to regular structures and patterns [Wan95, ch. 7], and contrast builds the basis for perceptual pattern recognition tasks in images. A general discussion on the pattern sensitivity of the HVS can be found in

works by Wandell [Wan95, ch. 7] and Shapley et al. [Sha+90]. However, evaluating elaborate CSF-models in performance critical rendering pipelines is computationally demanding as this requires orientation-tuned channels (cortex transform), Fourier or wavelet transforms of the rendered images. Nonetheless, CSF models build the basis for numerous high-level vision models in image and video metrics (Section 3.2). More details on CSFs are presented in the work by Johnson and Fairchild [JF02] and by Lukac [Luk12, p. 17].

3.1.4 *Color Sensitivity*

The different sensitivities, distribution, and densities of cone types highly affect human color perception – the sensation of visible light with wavelengths λ ranging from 390 to 750 nanometers. A visible stimulus S depends on wavelength-dependent illumination $I(\lambda)$ and object reflectance $R(\lambda)$ [Luk12]. When a stimulus is observed, the cones respond by integrating energy over all wavelengths:

$$(L, M, S) = \int_{390}^{750} l(\lambda), m(\lambda), s(\lambda))I(\lambda)R(\lambda)d\lambda,$$

where $l(\lambda)$,$m(\lambda)$, and $s(\lambda)$ describe the spectral sensitivities of the L-, M-, and S-cones.

The most commonly used color model in computer graphics is the RGB model. This model *additively* combines the three primary colors red, green and blue (r,g,b). However, this model is not considering perceptual implications. The color space depends on the underlying device and the primaries for r, g, and b. A later standardization by the Commission Internationale de l'Eclairage (CIE) has targeted this matter. CIE-RGB [Bod+07, p. 29-30] uses special reference colors for r, g, and b. Assigning these references became possible by determining CIE-XYZ [Bod+07, chp. 3], a device independent and non-negative color space [Win12, p. 24-26]. This was the first attempt to encompass the retinal response of the HVS with the goal to cover all perceivable colors with positive coordinates. CIE-XYZ is often used in practice, but the color space is perceptually irregular, as is RGB, HSV or many of the standard color spaces, and does not consider the different sensitivities of the L, M and S cones. In these spaces, e.g., the distance for perceptually equally-different green colors are smaller than for red or blue. Perceptual color spaces, such as CIE-LAB [Bod+07, pp. 61ff], CIE-LUV [Bod+07, pp. 64ff], CIEDE2000 [Bod+07, pp. 91ff], are almost linear and can be converted from CIE-XYZ values [Win12, p. 28-29]. Within these color spaces, perceptual differences between any two colors are directly related to the Euclidean distance. As the response of the cones cannot be readily measured, the LMS color space [Bod+07, p. 233] was designed to relate to the spectral responses of the LMS-cone types directly.

Several linear transformations from CIE-XYZ to LMS space have been proposed based on empirical measurements. A transformation is performed by multiplying the XYZ values with an empirically derived matrix to account for the different spectral responses of the cones. Common models for the LMS space are the Hunt model [Hun91; Hun94], LLAB [LLK96], CIECAM97 [Bod+07, pp. 269-270], and CIECAM02 [Bod+07, pp. 270-271]. A detailed presentation of perceptual color models is given by Fairchild [Fai05], Gonzales et al. [GW07], and Lukac et al. [Luk12]. Hering [Her20] developed the idea of color opponency in 1892. He found that certain colors cannot be mixed, e.g., there is no reddish green. This was later empirically validated [HJ57] and proved beneficial in several image processing tasks [BBS09]. Interestingly, both CIE-LAB and CIE-LUV provide color opponency in their color channels.

Colors, in addition, highly affects the ability of the HVS to perceive contrast. An early approach to account for the eye's sensitivity to different contrasts per wavelength and color used in computer graphics was introduced by Mitchell [Mit87]. Contrast is detected using separate, perceptually inspired thresholds for the RGB colors. Other approaches in computer graphics such as the one presented by Bolin et al. [BM95; BM98] use a conversion to transform the CIE-XYZ color space to an LMS space. The LMS values are used to detect the regions that have strong perceivable contrasts. More information on these approaches can be found in Section 4.2.1.

3.1.5 *Adaptation Models*

Adaptation allows for perceiving the environment over a High Dynamic Range (HDR) exceeding 24 *f-stops* [MMS15], where the illuminance reaching the sensor (retina) is doubled between two f-stops. This adaptation means that the HVS can perceive visual stimuli with an illuminance of more than the 2^{24}-fold of the minimum perceivable illuminance. The actual perceivable dynamic range depends, however, on the peak brightness of a scene, which is very limited on common Low Dynamic Range (LDR) displays. A TMO provides models to approximate the appearance of High Dynamic Range (HDR) images on low-dynamic-range display devices or prints. Detailed information can be found in the survey papers by Reinhard et al. [Rei+10], Eilertsen et al. [Eil+13], Fairchild [Fai15], and Mantiuk et al. [MMS15]. An evaluation on TMOs from a perceptual perspective has been carried out by Michael Stengel in our previously published state-of-the-art report [Wei+17].

In addition to those general considerations of adaptation, different areas on the retina need varying times to adapt to new lighting situations due to previous visual stimuli (simultaneous and successive contrast) [ARH78]. When viewed simultaneously or in quick succession, different objects having the same color appear to have different colors when viewed, for example in front of a different colored background. Retinal photoreceptors need time to refresh, which leads to bleaching processes [Gut+05]. Hence, the image stays locally "imprinted" in the visual system for some time, resulting in perceivable *afterimages*. Ritschel and Eisemann provide a model for this process for real-time applications [RE12]. It also has been refined to model color transitions when the afterimage disappears [Mik+13]. After-image-like effects can also be used to increase the perceived brightness of a light [ZC01]. Similarly, perceived brightness, as well as perceived color can be altered by flickering; an effect called *apparent brightness*. This has been used to improve perceived color saturation of images beyond the display capabilities [MFN16].

3.1.6 *Visual Masking*

Another phenomenon affecting sensitivity is visual masking, which happens when the perception of one stimulus, the *target*, is affected by the presentation of another stimulus, called a *mask*. A profound overview and survey on physiological findings for visual masking can be found in the work by Legge and Foley [LF80], Breitmeyer and Ogmen [BO00] as well as Enns and Lollo [ED00]. The effects of visual masking occur spatially and temporally (backward masking). Several methods attempt to model visual masking. Spatial visual masking

is typically considered by determining a background over which potential target patterns are superimposed. A survey on image processing with information on visual masking can be found in the work of Beghdadi et al. [Beg+13]. Often, a transducer function captures the non-linearity of visual masking as a function of the contrast level. Effectively this function models a hypothetical response of the HVS to the input contrast, which is scaled in perceptually uniform Just Noticeable Difference (JND) units. By computing a difference between the original and distorted signals expressed by means of the transducer in the JND units, the visibility of distortion (the difference over 1 JND) as well as its magnitude can immediately be derived. A more straightforward approach is to directly scale the input contrast by the corresponding threshold value that can be derived from the CSF that in turn is elevated proportionally to the masking's signal contrast.

Visual masking is widely used for image [Wat93; HK97; ZDL02] and video compression [AKF13]. In addition, visual masking is commonly modeled in image and video quality metrics. For example, the Daly's Visible Differences Predictor (VDP) [Dal93] employs the simpler threshold elevation approach, while the Sarnoff Visual Discrimination Metric (VDM) [Lub95] is based on a transducer. Both approaches are discussed in more detail in Section 3.2.1 and are used from methods in Section 4.1.2 and Section 4.2.2. Employing transducers is also common in computer graphics applications in the image contrast [Fer+97] as well as stereo disparity [Did+11] domains.

3.1.7 *Temporal Resolution*

In addition to spatial contrast sensitivity models, temporal changes may have a strong effect on the visibility of a pattern. "The *critical flicker frequency* (CFF, also flicker fusion frequency) describes the fastest rate that a stimulus can flicker and just be perceived as a flickering rather than stable" [Adl+11, p. 700]. Figure 30 shows the estimated adaptation-dependent temporal sensitivity for different retinal illuminance values at photopic levels with an achromatic flickering stimulus. The retinal adaptation levels are measured in *Troland* $\mathrm{T} = L\frac{cd}{\mathrm{m}^2} \cdot p_a \mathrm{mm}^2$, taking the size of the pupil area p_a and the luminance L into account. In Figure 30, temporal frequency along the x-axis is plotted against the modulation ratio of the flickering stimulus. The modulation ratio represents the extent that the sinusoidally modulated light deviates from its average direct current component [Adl+11, p. 705]. It can be seen that at low frequencies the modulation sensitivity is approximately equal for all adaptation levels. For higher flicker frequencies, modulation sensitivity strongly depends on the retinal illuminance values. However, the temporal CSF plotted in Figure 30 does not show the complete picture. Many other properties result in deviation from the presented CSF behavior.

Kelly [Kel61] compared chromatic flickering with achromatic stimuli and explored spatial contrast sensitivity in combination with temporal contrast sensitivity resulting in a spatio-temporal CSF [Kel79]. The mathematical model can be used to describe how a feature moving across the visual field also affects the perception of detail, which leads to motion blur. As the visual acuity, the CSF is also a function of the retinal velocity. The surface produced by evaluating the CSF over a range of velocities and cpds is called the *spatiotemporal threshold surface*. Kelly's [Kel79] measurements showed that the CSF remains essentially constant for

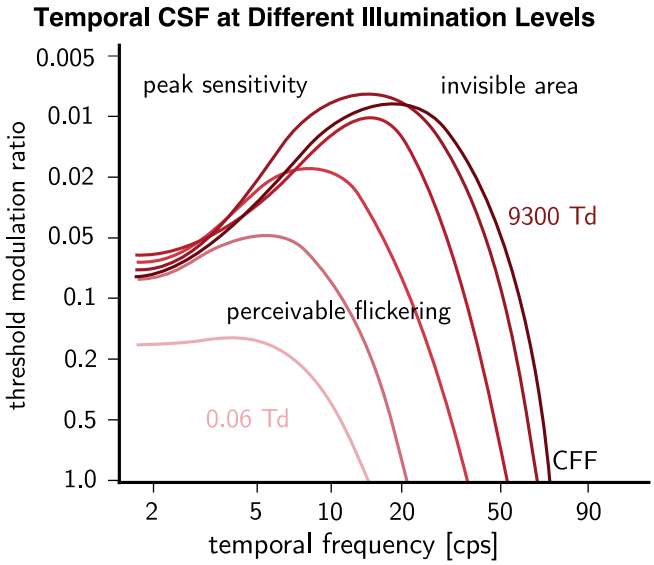


Figure 30: Temporal contrast sensitivity function (CSF) for different retinal adaptation levels. Each curve represents the threshold modulation ratio (percentage deviation of the average value) of a just-detectable flicker stimulus for a given adaptation level (in Trolands) plotted against flicker frequency (in cycles per second, cps). Low levels of retinal illuminance result in a low-pass CSF, whereas higher levels reshape the CSF into a more band-pass curve. *Image adapted from Adler et al. [Adl+11, p. 705]*

the first $0.1°/s$ before it is the target to fall-off. These properties later led to the development of the retinal-dependent visual acuity model by Reddy (Section 3.1.3).

If a light flickers faster than the speed the HVS can resolve, the flashing light is perceived as stable rather than seeing a sequence of flashes. The CFF is dependent on different features. For photopic lighting conditions, the CFF increases linearly with the logarithm of luminance of the flickering light over a dark background. This is known as the Ferry-Porter law [Por02]. The Granit-Harper law states that the CFF increases linearly with the size of the stimulus area [GA30]. Rovamo and Raninen [RR88] have shown that for constant stimulus size and luminance, the CFF increases with eccentricities up to 55°. Towards the far periphery, the CFF decreases again. Hence, the mid-peripheral vision has better temporal resolution than foveal vision and far-peripheral vision. If the CFF is plotted against the number of stimulated retinal ganglion cells, the resulting function is linear across all eccentricities [RR88]. This directly relates to the number of Frames-per-Second (FPS) needed to be rendered in order to perceive an animation rather than a sequence of individual images. A high number of frames combats the flickering and decreases the motion-induced blur. Therefore, temporal upsampling is often used to increase the frame rate artificially [Did+10b] (Section 4.3).

One can ask if there is a dependency between the HVS' ability to detect motion and certain eccentricities. McKee [MN84] conducted several experiments showing that the peripheral visual field has no special ability to detect motion; the threshold to detect motion and accelerations are not better in the periphery than in the fovea. However, they are not worse either. Interestingly the threshold to detect motion is much smaller than the MAR. This means the peripheral visual field is as good as the fovea when it comes to motion detection. A fact that must be regarded when designing gaze-contingent rendering systems, where temporal artifacts must be avoided in the peripheral visual field.

3.2 HIGH-LEVEL MODELS

High-level models integrate low-level components and methodologies to find a perceptual measure for image and video quality and to detect perceptual differences. These perceptual measures and models can be rather generic and have the potential to be directly embedded into rendering software. Perceptual image metrics can be categorized into *full-reference metrics* and *No-reference quality metrics*.

Full-reference metrics require ground truth, the reference. These metrics usually results in a single value describing the overall perceptual difference of two images or frames. This way, they are sufficient for comparing different rendering algorithms. Often these metrics also enable to compute a map that provides localized information about the strength of the perceptual differences. However, it is challenging to use such a map to guide rendering algorithms, for example to put more rendering effort in certain image regions. The reason for this is that no reference is available. One option here is to compare frames from the subsequent rendering stages to gain insight and adapt rendering convergence. However, this is a computationally demanding process.

No-reference quality metrics can directly judge the quality of single images or videos without any reference. While usually less reliable, they are often better suited for rendering applications. Being able to "blindly" estimate image quality allows for guiding image synthesize approaches in a more flexible (progressive) manner.

In the following sections, representative full- and no-reference quality metrics are described in greater detail.

3.2.1 *Full-Reference Metrics*

Image metrics generally attempt to derive an abstract measure for judging image quality. Prominent examples include computing a Mean Squared Error (MSE) of pixel differences and the associated Peak Signal-to-Noise Ratio (PSNR). Because MSE and PSNR do not take the limitations of HVS into account [HG08], they perform poorly when estimating the image quality as perceived by a human observer. Therefore, more elaborated versions have been developed. For example, simple contrast sensitivity can be evaluated using *single-channel measures* [Lue+03, p. 287-288]. To this end, an image is transformed by computing a local contrast image that is followed by a transformation of this contrast image to the frequency domain, and an application of the respective CSF model. However, *multi-channel measures* are required if factors such as eccentricity, retinal velocity, or visual masking are to be included. A multi-channel processing splits image input into numerous different (potentially orientation-tuned) channels [Lue+03, p. 288-289]. While this increases computation and storage requirements immensely, it is a common approach of elaborate perceptual metrics. Also, based on these methodologies more sophisticated PSNR methods are in existence that employ contrast sensitivity models, such as PSNR-HVS [Egi+06], and account for visual masking, for example PSNR-HVS-M [Pon+11].

In contrast to the aforementioned approaches, the Structural Similarity (SSIM) index [Wan+04] has become one of the most popular and influential quality metrics in recent years. It emphasizes less on the precise perceptual scaling but is still sensitive to the differ-

ences in the image brightness, contrast, and structure. In particular, the structure modeling component plays an important role in achieving a high accuracy [Čad+12]. SSIM is based on a top-down assumption that the HVS is highly adapted for extracting structural information from the scene. Therefore, it can be expected that a measure of structural similarity is a good approximation of perceived image quality [WSB03]. Approaches such as the Multi-Scale SSIM (MSSSIM) [WSB03] compute these structural similarities of the image at various scales. This multi-level processing mimics vision processing on the retina, the Lateral Geniculate Nucleus (LGN) and the early stages in the visual cortex (Section 2.1). Other modern video quality metrics, such as the Visual Information Fidelity (VIF) index [SB05], rely on natural-scene statistics and employ an information-theoretic approach in order to measure the amount of information that is shared between pairs of frames. A survey on video quality metrics can be found in work by Wang [Wan06].

Currently, meta-models such as the Video Multi-Method Assessment Fusion (VMAF) by Netflix [VMA17] are gaining increasing popularity. Here, some of the aforementioned metrics are fused using machine learning in order to compute a single abstract value which is representing the video quality with respect to ground truth. The model was trained to provide quality scores for 1080p images at natural viewing conditions. VMAF also includes the computation of a Mean Co-Located Pixel Difference (MCPD) that enables to capture motion and image inconsistencies more accurately. However, all of the above metrics do not model the entire visual pipeline but rather make assumptions about the most influential features, or only use simple models of the HVS such as contrast sensitivity. They often fail in capturing just visible (near-threshold) differences. Likewise measuring the magnitude of supra-threshold differences as well as scaling them in more meaningful JND units is highly challenging.

A more complete model of the HVS is implemented in Daly's Visible Differences Predictor (VDP) [Dal93]. VDP allows to compare two input images and derive a *difference map*. To this end, each input image undergoes identical processing. First, the retinal response and luminance adaptation are simulated. Then, the images are converted into the frequency domain, where CSF (Section 3.1.6) filtering is performed. This step scales pixel values into perceptually meaningful detection threshold units. Such perceptually-scaled input images are decomposed into spatial and orientation channels (cortex transform) to account for per-channel visual masking (Section 3.1.6). VDP is a prime example for a multi-channel measure. Its basic processing steps mimic the processes in the LGN and in visual cortex region V1. As discussed in Section 2.1.3 neighboring parts of the retina trigger neighboring parts in the LGN and in V1. Here, the output of the photosensitive cells is aggregated. As the different cell types discussed in Section 2.1.3 are tuned to subdivide the input in streams that are orientation sensitive to input, the *cortex transform* used in VDP results in numerous different channels representing different spatial frequencies and in channels that represent the input captured from different orientations. Finally, per-channel differences of the two image in questions are transformed into the probability of perceiving the differences by means of a psychometric function and then accumulated in an aggregated difference map. VDP is particularly sensitive in detecting image differences near the visibility thresholds (Section 4.2.2). Since rendering artifacts, such as Monte-Carlo noise, typically cannot be tolerated, VDP is a useful tool for guiding such artifact suppression below the visibility level. If the goal is to measure the magnitude of clearly visible (supra-threshold) artifacts, the precision of VDP is limited.

Various researchers have extended Daly's original VDP. Jin et al. [JFN98] and Tolhurst et al. [Tol+05] extend the model to consider the eye's color sensitivities (Section 3.1.4). To this

end, they follow the color-opponent theory by Hering [HJ57] and use chromatic CSFs [Mul85; MS99] in order to account for color information in separate channels. In practice, input images are transformed into luminance, red-green, and blue-yellow channels, and then VDP processes those separately using a corresponding CSF. Eventually, the results from all channels are combined to compute the final difference.

Mantiuk et al. [Man+05] improve the prediction of perceivable differences in HDR images (HDR-VDP). They integrate several aspects of high contrast vision such as light scattering by the eye optics, the nonlinear light response for full-range luminance, and local adaptation. In particular, light scattering is important for HDR signals as strong light sources or highlights can lead to significant glare, even for remote image regions. In a follow-up article, Mantiuk et al. [Man+11] introduce HDR-VDP2. It improves on the original metric, among others, by using another visual model for varying luminance conditions [Bar99]. Moreover, they have calibrated and validated the model using several image quality and contrast databases. Swaffort et al. further improve HDR-VDP2 by providing an image metric to assess gaze-contingent rendering quality [SCM15; Swa+16]. It adds measurements for the peripheral vision degradation at increasing eccentricities by a CSF. Both CSF models have been discussed in Section 3.1.3. Narwaria et al. [Nar+14] improve the accuracy of HDR-VDP2 prediction by employing a comprehensive database of HDR images along with their corresponding subjective ratings. The same group provides a quality measure for HDR videos [NDL15]. This measure is based on a spatio-temporal analysis focused on fixation behavior when viewing videos. The performance of the method is verified using an HDR video database and their subjective ratings.

Despite VDP being widely used for various rendering systems (Section 4.2.2), it is computationally expensive due to the individual processing of each channel and band within the frequency domain [LMK01]. The Sarnoff Visual Discrimination Metric (VDM) [Lub95] is simpler, requires less computational effort and Graphics Processing Unit (GPU)-based implementations are available [WM04]. In contrast to VDP, where image filtering is performed in the frequency domain, VDM uses convolutions and down-sampling in the spatial domain only [Čad04, p. 18]. A transducer function (Section 3.1.6) is applied to account for visual masking [BM98]. VDM derives two measures from the input images. The first is a single value representing the strength of the perceptual difference. The other is a map containing the locations of regions with a high predicted visual difference. In contrast to VDP, VDM is an example of a metric specifically designed to account for the magnitude of supra-threshold image differences. However, this comes at the expense of precision loss, when near-threshold differences need to be judged.

A metric specifically designed for realistic image synthesis and inspired by VDP and VDM was introduced by Ramasubramanian et al. [RPG99]. Their metric attempts to predict thresholds for detecting artifacts in order to spend most computational effort in regions with the highest visibility of artifacts. The metric models the adaptation processes, contrast sensitivity and visual masking. The key idea is to precompute the most expensive metric components for direct lighting as a per-pixel contrast threshold elevation map. Such a map is directly used to guide the costly computation of indirect lighting. Similarly, Walter et al. [WPG02] analyze texture information to find a tolerance measure for visual error. The tolerance measurements can be stored as a standard mip-map, along with the texture, and efficiently used as a lookup table during rendering.

Ramanarayanan et al. [Ram+07] pointed out that, even though VDP can predict various visible image differences, they usually do not matter to human observers. The authors attempt to focus on visual equivalence and determine whether two images convey the same impression regarding scene appearance. A couple of psychophysical experiments along with a validating study led to the Visual Equivalence Predictor (VEP) metric. Later, Křivánek et al. [Kři+10] investigated visual equivalence for instant radiosity (virtual point light) algorithms and proposed many useful rendering heuristics, which were otherwise challenging to formalize into a ready-to-use computational metric. Vangorp et al. [Van+11] propose a perceptual metric for measuring the perceptual impact of image artifacts generated by approximative image-based rendering methods. They consider artifacts such as blurring, ghosting, parallax distortions, and popping. For the evaluation, the authors generated viewpoint-interpolated image datasets containing different levels of distortions and respective artifact combinations. All of the aforementioned metrics assume that both reference and test images are perfectly aligned. However, human perception compensates for geometric transformations. For example, a human can easily tell that an image is identical to its rotated copy. Kellnhofer et al. [Kel+15] present a metric that quantifies the effect of transformations not only on the perception of image differences but also on saliency and motion parallax.

Cadik et al. [Čad+12] compare a large number of state-of-the-art image quality metrics and evaluate their suitability for detecting rendering artifacts. This investigation includes SSIM, MSSSIM, and HDR-VDP2. The authors conducted user experiments that show that the most problematic features for existing metrics are an excessive sensitivity to brightness and contrast changes, calibration for near-visibility threshold distortions, lack of discrimination between plausible/implausible illumination and a poor spatial localization of distortions for multi-scale metrics. Based on these observations, the authors have developed a test dataset in order to support the development of future metrics. The current trend is to employ machine learning methods to derive full reference metrics [ZK15; ZWF16; APY16]. So far, the existing metrics are generally successful in predicting the Mean Opinion Score (MOS) value, i.e., a scalar that characterizes the overall image quality, without producing detailed error maps. While regarding the computation performance such metrics can be a viable option for rendering applications.

3.2.2 *No-Reference Metrics*

All of the methods in the previous section are comparative approaches that assume the reference image is given as an input. However, in the vast majority of computer graphics applications, the goal is to synthesize a new image. In such a situation, the reference image is missing. Here, it is desirable to have a method that can blindly estimate the quality of an image or a video (Figure 31).

Chandler and Hemami [CH07] quantify the visual fidelity of natural images based on near-threshold and supra-threshold properties of human vision. Their Visual Signal-to-Noise Ratio (VNSR) uses contrast thresholds to detect image distortions. This detection stage is coupled to a wavelet-based analysis of visual masking in order to determine whether the distortions are visible. If the distortion is above the threshold, a second stage uses low-level vision models and accommodates different viewing conditions and contrasts to compute the final VNSR value.

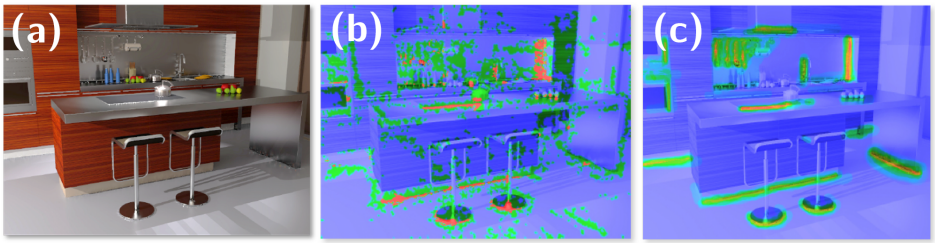


Figure 31: Example of a no-reference metric. No-reference metrics derive a measure (b) of perceived image quality based on a single image (a). Results can be close to the ground truth (c) often determined in psychophysical experiments. *Image from Herzog et al.* [Her+12]

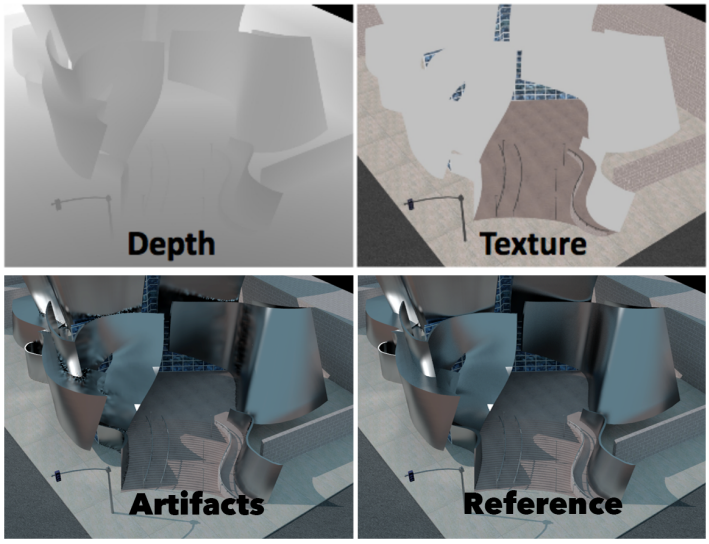


Figure 32: NoRM – training [Her+12]. An example scene from the data set used to train a Support Vector Machine (SVM) to derive the no-reference metric NoRM. Each case consists of a reference color image and a test image with different rendering artifacts. Moreover, specific 3D scene information that is readily available in rendering such as the depth and diffuse material/texture buffers are employed. *Image from Herzog et al.*[Her+12]

Stokes et al. [Sto+04] attempt to predict the perceptual importance of the indirect illumination components with respect to image quality by conducting a series of psychophysical experiments. Their idea is based on the observation that the different direct and indirect illumination components are probably not equally important with respect to their contributions to the visual quality. Their metric is solely based on simple measures of scene reflectances that are gathered during the computation of the direct illumination component. Hence, a lightweight progressive update during the integration of the indirect illumination component is possible. However, this metric cannot detect local artifacts which would sometimes be desirable for local image enhancement.

Such local error maps, as also shown in Figure 31 (b), are generated by the no-reference metric NoRM as proposed by Herzog et al. [Her+12]. They use a machine learning system, trained with various types of rendering artifacts that are locally marked by the subjects in a perceptual experiment. At both, the training and error prediction stage, they actively use feature descriptors based on 3D scene information (Figure 32, top row) in order to compensate

for the lack of a reference image. They also employ low-level models of the HVS in order to predict the perceived strength of rendering artifacts in the error map (Figure 31 (b)).

Since it is difficult to build a general-purpose no-reference quality metric, many attempts have been made to focus on specific artifacts such as ghosting [Ber+10] or camera-shake blur [Liu+13] that cause specific and relatively easy to isolate changes in the image signal. A closed source toolkit by the MSU Graphics & Media Lab [Vat+11] enables the estimation of artifacts such as temporal noise, blocking artifacts, and overall brightness flickering. A more detailed introduction to the noise estimation metric can be found in Appendix A.1.

Support Vector Machines (SVMs) and other classic machine learning techniques have been employed in order to derive a number of novel no-reference metrics, similar to NoRM using neural networks. These metrics typically rely on natural image statistics and are focused on predicting various incarnations of noise and compression artifacts such as ringing, blur, or blocking [MB10]. Here, the work by Liu et al. [Liu+13] provides a more complete survey. Most recently, Wolski et al. [Wol+18] present another large dataset of images and marked visible image distortions. The authors also propose a statistical model for a meaningful interpretation of such data and to be used as input to a neural network in order to derive an image metric.

Similar to the full-reference metrics (Section 3.2.1), deep machine learning may provide a viable tool for robust, no-reference artifact detection in the following years. Recent examples in the field of blind image quality assessment by training deep neural networks include works by Kang et al. [Kan+14], Bianco et al. [Bia+16], Bosse et al. [Bos+16; Bos+18], and Bemana et al. [Bem+19].

3.3 ATTENTIONAL MODELS

Attentional models are another way to describe higher-level processes of vision. Often, these models are used to detect and quantify those components of the scene that catch our attention. Components are for example points and features that are likely to be fixated by our gaze or are in the scan-path that provides order to those features as our gaze wanders through the environment. All this is centered around a *stimulus's saliency*. The saliency refers to the visual "attractiveness" or importance of components and features in the environment. In this section, an overview of the most common and successful models to describe attention are presented. It is based on the work by Michael Stengel presented in the previously published state-of-the-art report [Wei+17]. An even more comprehensive review of the state-of-the-art in modeling visual attention can be found in work by Scholl [Sch01], and Borji and Itti [BI13].

Saliency models can be categorized in *bottom-up* and *top-down* models. *Bottom-up* models are driven by low-level features and stimuli such as parts that show high discontinuities, e.g. in contrasts. *Top-down* models are driven by high-level attention, for example cognitive processes such as subjects solving a task or the intention of the subject understanding the scene. Historically these two-fields appear to have largely ignored each other [Lue+03, p. 285]. However, neuroscience has confirmed that there are "feed-forward" links in the brain from the centers of higher cognition to the centers of lower-level visual processing [Lue+03, p. 286]. Also higher-level cognitive processes, such as *visual tunneling*, can largely influence perceptibility (Section 2.2.5)

Attention can be guided by keeping track of additional information in so-called pre-attentive object files – usually, those are stored per shape and object and are created before actual attention is placed upon them. This way they may direct scan movements of the eyes in the conscious, attentive processing of information [PRC84; PRH90] and have often be used to predict fixation locations [JDT12; VDC14; Byl+15]. One information that can be stored is, if and when an object was visible before as "new" objects may likely attract attention [WB97].

Saliency is generally recorded in saliency maps, often visualized as a greyscale image or heat map. Such a map describes with which probability a particular image region is paid attention to by the HVS. The saliency map can be thought of as a summary of the conspicuities of all visual stimuli. Several approaches have been developed to compute such maps. One approach is to look at common gaze patterns usually followed by healthy adult humans. However, research has found that there are differences between cultural environments [CBN05] and gender [VCD09; SI10]. Another promising approach is to "learn" visual saliency from large amounts of eye tracking data [ZK13] or employ object or scene knowledge. One example for the latter is that humans are similarly attracted by faces and objects that are located in the line of sight of such faces [Gol10, p. 823]. *"A strong interrelationship exists between pre-attentive object files and saliency: pre-attentive segmentation (the process of creating "figural units") is based either on perceptual grouping (object shapes are integrated with surface details) or saliency"* [Edw09, pp. 57–59].

3.3.1 *Bottom-up Saliency Models*

Early processes in vision greatly affect our percept. Hence, it can be assumed that the salience of a stimulus is likewise affected by low-level features such as color, orientation, brightness and contrast of the stimulus.

The development of bottom-up saliency models was motivated by feature integration theory [Tre88] which states that these individual stimulus features can be added linearly resulting in a normalized saliency map (Figure 33 (b)). A renowned biologically inspired bottom-up model was introduced by Itti and Koch [IKN98]. It rebuilds some of basic stages of a human's visual processing, for example by measuring local contrast on different scales, simulating the receptive fields of ganglion cells in the retina and neurons in the visual cortex (Section 2.1). Based on this model, various approaches have been developed that target different aspects of the HVS. These aspects include: depth, motion, proximity and habituation components [LDC06], high edge density regions [MRW96], binocular disparity [JOK09], local contrast, orientation and eccentricity [Opr+09], to name a few. A more detailed list can be found in our previously published state-of-the-art report [Wei+17].

According to the bottom-up theory, the detection of objects across the visual field is assumed to be subconscious and does not depend on attention [WDW99]. However, while low-level features provide strong hints which parts are salient, they do not give information on the sequence, order and duration of fixations, i.e. the *scan path*. However, in order to determine the scan path, models do not only rely on low-level features but attentional mechanisms [Oli+03; Hen+07] to be found in top-down models presented in the following section.

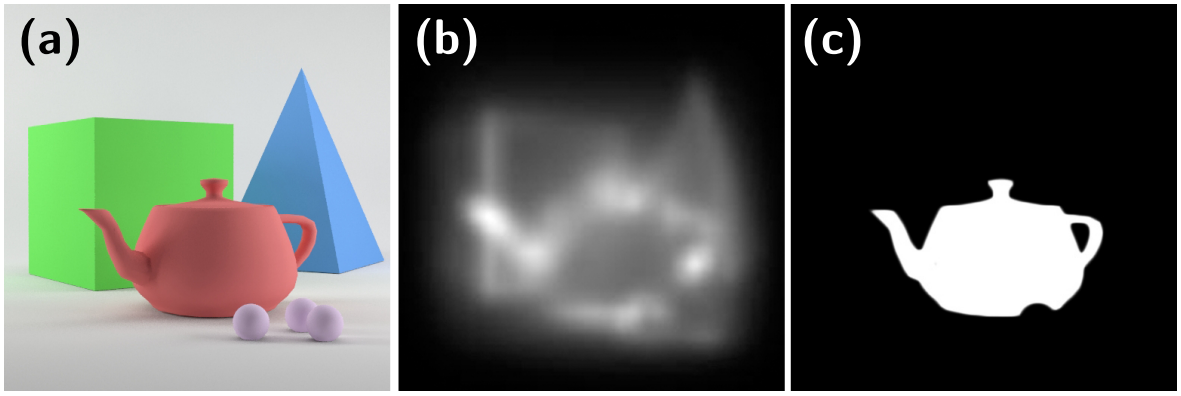


Figure 33: Bottom-up and top-down saliency. Given the input image (a) the method by Harel et al. [HKP07] based on bottom-up saliency can predict fixations in a free-viewing task (b). Top-down prediction in a visual search task for the teapot can result in (c). *Image from Weier et al. [Wei+17]*

3.3.2 *Top-down Saliency Models*

Top-down models make use of attentional as well as cognitive models and observations to derive saliency. However, these approaches are usually combined with bottom-up approaches in order to increase prediction accuracy. In this manner various models have been developed that include: task-related feature values [IK01], color opponent images and task information [GVC03], an importance map for task-relevant objects combined with a bottom-up saliency computation step [Sun+05], gaze behavior for natural scenes including face detection [Cer+08], and many more [Wei+17]. To this end, besides using pre-attentive object files, methods commonly derive scene knowledge from figure-background segmentation [FWM15], face and person detection [VJ04; FMR08], object detection [Cer+08] or manually defined task-specific location bias [CCW03] (Figure 33 (c)).

In addition, deep convolutional networks trained on large image data sets have shown great improvements in fixation prediction [VDC14; KTB14; KAB15]. Also, trained networks can directly incorporate models that account for the influence of high-level (faces, text) and abstract features. In this field, Kümmerer et al. [KTB14] reuse neural networks to decrease the computational effort in creating a network for saliency prediction. Kruthiventi et al. [KAB15] developed a location-biased convolutional filter. This enables learning location-dependent patterns of fixations.

Apart from these learning-based approaches, findings in cognitive science remain important to improve modern high-level saliency predictors. Koulieris et al. [Kou+14a] make use of the *scene schema* [HH99] and the *singleton hypothesis* [TG02] in order to improve saliency prediction. The scene schema hypothesis states that salient objects are those that are not expected to occur in that scene, e.g., a lawn mower in the kitchen is considered to have a high salience. The singleton hypothesis is based on the observation that the HVS is more sensitive to features that are singular across the visual field while suppressing prevalent features [Wol94]. Hence the singleton hypothesis states that the viewer's attention is drawn by stimuli that are *locally unique* and *globally rare*.

3.3.3 *Attention Model Quality*

Free-viewing tasks commonly evaluate the prediction accuracy for bottom-up saliency. Participants look at photographs and watch videos (ideally for the first time) [JDT12], while their gaze is recorded. This way, the model predictions can be compared to real-world measurements. However, there is controversy about the role of bottom-up versus top-down mechanisms in the context of gaze prediction [JDT12; VDC14; Byl+15]. Free-viewing experiments assume controlled conditions in order to be comparable. These conditions are difficult to achieve since participants may be biased by the cognitive load when performing the tests. It is clear, that attention models for passive gaze prediction do not provide exact solutions and are by far less exact than gaze measurements by an eye tracker. Nonetheless, knowing the approximate gaze location may be sufficient for some applications. For free-viewing tasks, saliency prediction based on convolutional networks learned from gaze-labeled natural images often outperforms traditional "hand-crafted" saliency predictors [VDC14; KWB14; Byl+15]. Learning the ground truth gaze data from only two observers already gives more accurate results than the best tested bottom-up gaze predictors.

Saliency prediction rarely results in a single, distinct salient region. Likewise, scene-viewing models have been primarily designed to predict *potential* fixation locations rather than the sequence of fixations. However, estimating the order and location, i.e. the *scan-path* of an observer is even more challenging. Hence, saliency researchers ignored it for a long time [Nut+10]. In fact, the experiments by Henderson et al. [Hen+07] confirm that scan paths generated by bottom-up saliency maps do not correlate well with ground truth. However, physiological knowledge can help to increase quality. Le Meur et al. [LC16; MPE16] exploit this bias of saccade motion in combination with bottom-up feature detection. When performing a saccade, humans are biased towards making either horizontal or vertical saccades. Using this knowledge results in a saccadic model for free-viewing scenarios that allows for predicting spatial fixation and scan-paths. The saccadic model by Trukenbrod and Engbert predicts fixation durations by varying the estimated fixation time with respect to the estimated foveal processing effort of the salient region [TE14]. Other research present approaches for extracting the scan path from a given video using machine learning on gaze data in combination with a perceptually inspired color space [Boe+06; Dor+10; DVB12].

To conclude, successful saliency models balance the complex interaction of low-level and high-level processes in visual perception. To this end, deep learning has shown great potential when obtaining high-quality saliency estimates and scan-paths. Also hybrid approaches that improve eye tracking quality might be a field for future research. In order to reduce the effects of tracking latency when using eye tracking for gaze-contingent rendering, Arabadzhiyska et al. [Ara+17] present a system-theoretic approach in order to predict saccade landing positions to slightly shift the measured gaze point ahead in time. Such approaches enable to meet the performance requirements in order to capture the critical eye movements more accurate (Appendix A.2). Likewise, robust, accurate, fast and temporally stable scan-path prediction remains a topic of ongoing research [Vig+12; VDC14; Hen+13; NE15].

3.4 CONCLUSION

This chapter presented a general overview of the HVS and its limitations as well as the associated models to describe key visual processes or mechanisms. After two centuries of intense research, the capabilities of human vision have been precisely measured. In particular, low-level knowledge on the eyes optical abilities (Section 2.2.1) and knowledge such as sensitivity, distribution, and interconnection of retinal photoreceptors are well understood (Section 2.2.2). Also, models for low-level vision features (Section 3.1) such as spatial and temporal contrast sensitivity and adaptation are fairly well studied. Examples that are convincingly adapting images to model perceptual properties can be found in adaptive local tone-mapping and brightness adaptations. These are widely used in games and movies (Section 3.1.5). However, integrating all existing models into high-level models (Section 3.2) and suitable for (interactive) rendering is more involved. Often, these models describe high dimensional functions involving parameters such as environment lighting and display properties. As such, these models have been mainly developed and evaluated for synthetic lab setups. The author believes that more research must place focus on specific measurements using real-world or synthetic images to assure the validity of each model. One example that supports this claim can be found in the work by Mantiuk et al. [Man+11] on the CSF function for HDR-VDP2. The widely accepted CSF model by Barten [Bar99] does appear to perform sub-optimally when applied to natural images (Section 3.1.3). Moreover, often the process of calibrating spatio-temporal models is cumbersome and error-prone. The entire process is even more challenging when no reference images or information of the output devices are available (Section 3.2.2). As it can be seen in the following chapters, methods that exploit limitations of the HVS may either rely completely on a model or take active measurements such as eye tracking into account in order to increase the success of the technique. Using eye tracking, the Point-of-Regard (PoR) can be derived from the location of the pupil's center mapped into screen space. Unfortunately, though steadily improving, eye tracking does have its pitfalls and accuracy is never high and latency low enough to respond to eye movements sufficiently well. This is especially apparent if 3D PoRs need to be measured (Chapter 7). Another approach in order to perform selective rendering is using saliency and predicting fixations as well as scan paths. However, higher-level perception such as attention (Section 2.2.5) is more difficult to measure and still not well-understood due to the complexity of the involved parts of the brain (Section 2.2.4). In addition, individual differences between subjects may vary widely. Most models derived on those measurements neither provide temporal stability nor are they able to provide a distinct gaze direction. This is a problem, as models for attention strongly depend on the gaze direction due to differences in foveal and peripheral vision. Nonetheless, knowledge of human perception can greatly improve the performance and quality of image synthesis techniques. In the next chapter, an extensive overview of such methods is presented.

PERCEPTION-DRIVEN EFFICIENT RENDERING

A Matter of Sampling

4

Nobody will ever solve the anti-aliasing problem.

Jim Blinn [Bli02, p. 166]

The most fundamental goals in synthesizing realistic images are *efficiency* and *realism*. Often these two goals cannot be considered separately. If the efficiency of an algorithm is increased, a greater proportion of computational resources can be spend on creating more realistic imagery. However, the field of efficient rendering techniques is broad. Methods that accelerate rendering can, for example, improve performance by an optimized utilization of the hardware. Noteworthy are also the tremendous efforts invested into the area of compilers, that allow high-level language constructs to be translated to highly-optimized hardware specific machine code. There is still at the very core of each rendering system the question of how to turn an n-dimensional representation into an image that can be displayed on the screen. Frequently this representation is defined over a continuous domain. When rendering a realistic scene at a high-quality, the continuous function is the amount of incident radiance on the image plane emitted from the 3D scene towards the observer. Unfortunately, it is almost always only possible to achieve computationally efficient sampling from this higher dimensional function in a point sampling manner. Hence, a great body of work focuses on how to make more efficient and less error-prone sampling possible. However, eventually images are displayed on physical devices such as computer screens, and these screens are formed by pixels, i.e., a small discrete number of objects emitting light a single wavelength at a time. Here, the extent of a pixel potentially covers a certain area of the view plane and thus covers a certain part of the (potentially adapted) function. Ideally, to drive each pixel correctly multiple point samples from the extent of the pixel need to be computed; these are then combined in a *reconstruction* process in order to compute a final pixel color.

Imagine a triangle in front of a white background that is moving across the scene. If it is touching a pixel's cell but not covering its center, the white color is emitted although the pixel intensity should be varied smoothly as it moves over the extent of the pixel. However, a naive rendering system treats a pixel as either covered or not covered. This leads to the triangle's outline appearing to be jagged. If the extent of a triangle covers multiple pixels, each with a different color, then only a single color from the triangle at the pixel's center is used to color that pixel. The matter is illustrated in Figure 34. Now, as the 3D scene or the virtual camera is moving, a pixel can suddenly change its color as the triangle at the pixel's center is changing, leading to temporal instabilities in the form of a highly-disturbing flickering. The entire process of computing and presenting digital images is likely to cause such artifacts due to the under-sampling of the higher dimensional signal, that is *aliasing*. If this under-sampling occurs spatially, it is referred to as *spatial aliasing*. Spatial aliasing artifacts manifest themselves in various forms, showing jagged edges or by temporal flickering [PH04, p. 280].

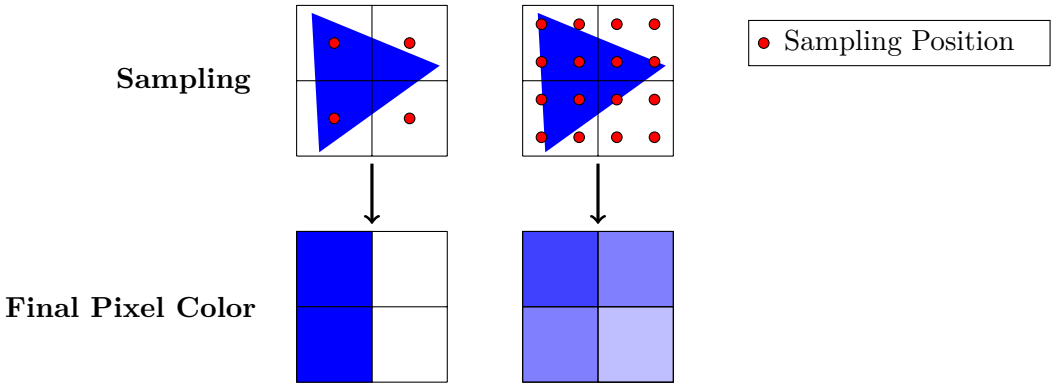


Figure 34: The sampling process of a triangle at two different sampling rates. Note that the higher number of samples per pixel yields a more accurate representation of the triangle when outputted to screen.

Both these artifacts are two different aspects of aliasing. We see that an approach to mitigate one issue might be far less efficient in the case of the other. One simple way to counteract both artifacts is by increasing the sampling frequency, by taking more samples over the extent of the pixel. However, this increases the computational costs and results in less efficient rendering. The question that remains is what sampling rate, that is how many samples are needed to enable a full reconstruction of the function. In order to answer this question, rather than looking at radiance in a 3D scene, we first look at a simple one-dimensional function $f(x)$ as the theoretical considerations are identical for higher dimensional cases. Now, let us assume that the function $f(x)$ is band-limited by $\mathbf{B}$, i.e. the function does not contain any frequencies greater than $\mathbf{B}$ such that $\mathcal{F}\{f(x)\}(\lambda) = 0, \forall \|\lambda\| > \mathbf{B}$. We refer to $\mathbf{B}$ also as (baseband) bandwidth of the frequency in Hertz with $\mathcal{F}$ being the Fourier transform. Using these prerequisites a minimum sampling rate for a perfect reconstruction can be determined using the renowned Nyquist theorem by Shannon, Nyquist, and Whittaker [Luk99].

Nyquist Theorem. A band-limited signal $f(x)$ with a (baseband) bandwidth of $\mathbf{B}$ can be equally expressed by a sum

$$f(x) = \sum_{n=-\infty}^{\infty} f(\frac{n}{2\mathbf{B}}) sinc(2\mathbf{B}x - n)$$

The sampling rate of $\lambda = 2\mathbf{B}$ is referred to as the Nyquist rate. For a perfect reconstruction of a bandlimited signal, the sampling rate λ must be *at least twice* the signal's bandwidth $\mathbf{B}$. In the case of $\lambda = 2\mathbf{B}$, a perfect reconstruction is only possible if the sampling is not happening exactly at the zero crossings from $f(x)$. Formal proof of the Nyquist theorem can be found in Marks's work [II91, pp. 33ff].

However, although often used to reinforce arguments in the context of computer graphics and well-suited to provide initial insight, the Nyquist theorem is hardly usable for image synthesize processes in general. Nyquist's sampling theorem states that the sampling frequency needs to be more than twice the maximum frequency of the sampled function. In order to establish this maximum frequency, the function must be bandlimited, which however, rarely is the case with a 3D scene. Discontinuities in depth along edges or shadows produce discontinuities in the signal. Essentially, the spatial frequency here is infinite. Therefore, there is

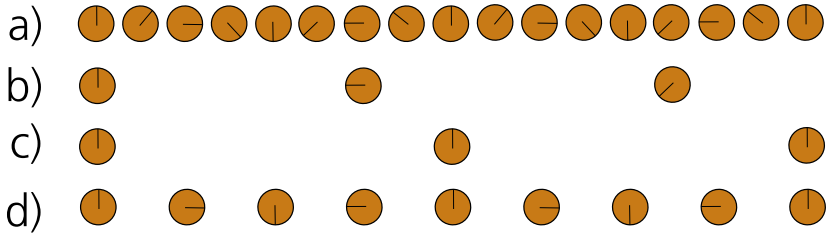


Figure 35: Temporal aliasing known as the wagon wheel effect. A (a) wheel turning at a certain speed that is captured at a low frame rate appears to be (b) moving backwards or (c) standing still. In order to (d) correctly capture the forward movement of the wheel, sampling frequency must be sufficiently high. *Image after Akenine-Möller et al. [AHH08, p. 118]*

always a position to be selected in between two other samples. Admittedly when representing and rendering the scene we are in reality limited by the maximum accuracy of the datatypes used. Even more important however is the insight that, after a certain number of samples has been drawn, the Human Visual System (HVS) is not likely to notice any change in the image if more samples are drawn. Likewise, if the display resolution is high, pixels might not be computed as these cannot be resolved by the HVS due to limited visual acuity at a specific eccentricity. All these constitute first cases where perception-driven approaches have been successfully used both to mitigate some of the problems as well as to increase rendering efficiency. Hence, *"the key insight about aliasing in rendered images is that we can never remove all of its sources. So we must develop techniques to mitigate its impact on the quality of the final image"* [PH04, p. 294] – ideally to a point where potential artifacts are not visible to a human observer.

Another form of aliasing apart from spatial aliasing is temporal. A sequence of single images displayed with a certain frame rate conveys a smooth motion. *Temporal aliasing* takes place in cases where the frame rate is too low and the animation stutters. One effect that occurs due to a limited temporal sampling frequency is the wagon wheel effect: A wagon wheel that is recorded at a limited temporal sampling rate appears to be turning backwards although it is actually turning in a forward direction [AHH08, pp. 118-119]. Here, having an idea of the Critical Flicker Frequency (CFF) of the HVS and the perceptual implications can be a valuable tool to estimate upper limits on frame rates and latencies when displaying smooth animations and when generating images in real-time. Also, when considering animations or interactive rendering, samples are not just entities in space but also in time. Hence, there is considerable research that attempts to exploit this Temporal Coherence (TC).

In this chapter, rather than discussing all methods that have been developed to improve the efficiency of rendering algorithms, the goal is more to explain general concepts and focus closely on relevant perception-driven approaches. An overview of such methods is illustrated in Figure 36. Based on the consideration of the current state-of-the-art relevant for this thesis, four general strategies that improve the efficiency of rendering algorithms can be identified.

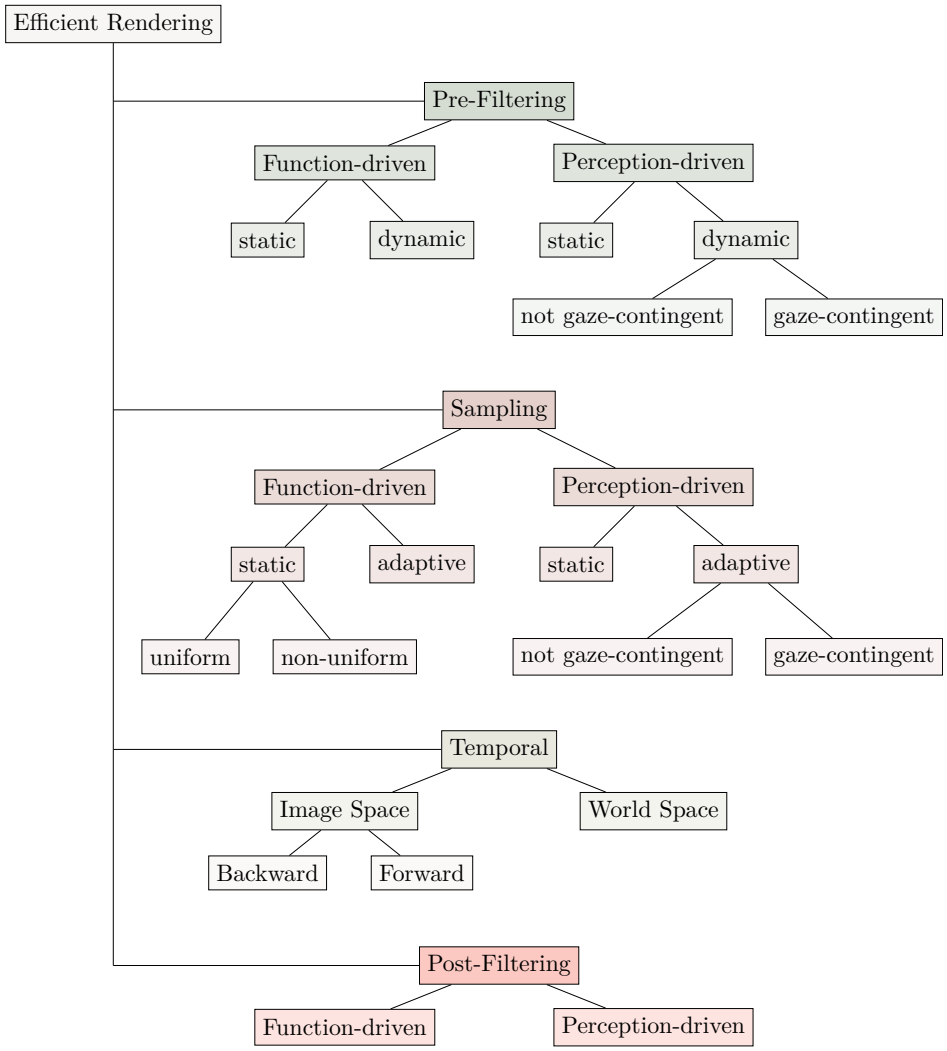


Figure 36: Overview of efficient rendering methods that determines the structure of the state-of-the-art in computer graphics.

1. *Pre-filtering* strategies attempt to provide a (multi-level) representation of the original function, i.e. the 3D scene or object, enabling more efficient and less error-prone sampling. These approaches commonly provide a scene at different Level-of-Details (LoDs).
2. *Sampling* strategies attempt to adapt the sampling rate and positions to generate images that are more perceptually pleasing by putting increased computational effort into regions that matter more to the observer or by transforming perceptually disturbing artifacts into less regular patterns or noise.
3. *Temporal* approaches attempt to reuse samples over time and thus increase the amount of available information per pixel - either by re-projecting samples in image space or by caching mechanisms in world space.
4. *Post-Filtering* approaches, though not improving sampling rate or sampling efficiency, attempt to alleviate and conceal potential artifacts often performed by selective blurring and bilateral filtering mechanisms in image space.

Please note however, that Figure 36 is not meant to be a taxonomy of efficient rendering techniques. One specific implementation of a method might use multiple strategies to produce images of a higher quality. However, the diagram in Figure 36 determines the structure of the state-of-the-art in computer graphics presented in the next sections. Also, this overview enables a classification of the methods and methodologies presented in the following chapters.

Regarding perception driven-approaches, *static* and *dynamic approaches* for pre-filtering and sampling strategies have been developed. The static approaches do not adapt themselves to changing perceptual requirements or the image content. In contrast, dynamic approaches try to distribute computational requirements in areas of the image or 3D scene where a higher quality matters most. Important in the context of dynamic methods is the terminology of *gaze-contingent methods*. Gaze-contingent is a general term for techniques that adapt their behavior, based on where the viewer is looking. Usually, an eye-tracker is used to determine the user's Point-of-Regard (PoR). The acquired information can the be used to update the visualization, the display, or most general, a system's response [Duc07].

In the following sections, we almost exclusively try to focus on those strategies that use perceptual insights, with and without active measurements, to generate more perceptually pleasing images in a shorter time frame. In line with the systems presented in the following chapters, a broad overview is given before a focus is placed on gaze-contingent methods for *pre-filtering* and *sampling* strategies.

CONTRIBUTIONS BY THE AUTHOR This chapter is based on our state-of-the-art report:

> Martin Weier et al. "*Perception-driven Accelerated Rendering.*" In: *Computer Graphics Forum (Proceedings of Eurographics)* 36.2 (Apr. 2017).

In contrast to the published report, a new structure is used to present the related work in the field (Figure 36). In this thesis, the related work is structured based on the fundamental concept of sampling and efficient rendering. Therefore, Section 4.3 and Section 4.4, which cover techniques that exploit *temporal coherence* and *post-processing* in image space, are entirely new. In addition, all chapters have been significantly extended, restructured, and updated to include the latest works – most notably Section 4.2. Here, a more in-depth discussion of the perceptual implications of sampling in Section 4.2.1 and a broader discussion of gaze-contingent rendering techniques in Section 4.2.3, are presented.

4.1 PRE-FILTERING

One way to mitigate aliasing artifacts and enable more efficient rendering is to adapt the sampled function, i.e. the 3D scene. This way, reducing the complexity of a scene accelerates the rendering and reduces artifacts at runtime. A reduction can be achieved by culling invisible objects, adapting object details or by directly employing multiple representations at a different LoDs, for example by reducing the number of polygons. *"In simplification, the goal is often to reduce model size while preserving visual fidelity."* [Lue+03, p. 279] The following section, begins by briefly presenting general approaches before considering the perceptual aspects in more detail.

4.1.1 *General Approaches*

Numerous methods have been developed that attempt to simplify 3D scenes either by processing a scene's polygonal description or by adapting/transforming it into another more simple or compact form. Filtering can happen at a model's surface, such as with bump or normal maps in order to simplify high-frequency surfaces geometric detail [COM98; San+01; WFG02]. Different approaches directly use impostors or layered depth images instead of full polygonal models [Sha+98; Ris07]. However, in this section a focus is placed on methods that are more general and directly work with 3D datasets. To this end, *polygonal simplification*, *point-based* and *voxel-based* methods are discussed.

POLYGONAL SIMPLIFICATION METHODS The simplest form of geometric simplifications techniques are *static* approaches that precompute a small discrete set of models at varying geometric LoDs. During the rendering process, the renderer selects the most appropriate one. Most basic, static simplification methods are vertex clustering or edge-collapsing techniques [Lue+03, ch. 5.1][Lin00], often coupled with specific error metrics such as the well-known Quadric Error Metric [GH97]. As presented in the next sections, adapting these metrics to minimize the perceptual implications of the induced change in appearance is often a key component of the perception-driven methods. However, as discrete changes between the models from the set can become visible, *dynamic* simplification methods have been developed, that attempt to simplify the model continuously at runtime. An advantage of the edge collapse operations is that the operation is reversible as the order of these can be stored [AHH08, p. 562ff]. This also makes it also a valuable tool for dynamic Continuous Level-of-Detail (CLOD) and progressive transmission over networks [Hop96; Tau+98]. Another dynamic LoD approach by Limper et al. [Lim+13] that is well-suited for progressive mesh refinements is based on reordering and quantization of the vertex positions in the mesh. However, in the case of all of the aforementioned approaches, there can be several criteria to select a model or control the simplification process. Potential constraints can be lower bounds on the frame rate or limited bandwidth. *View-dependent* approaches are most frequently found in perception-driven rendering techniques. These allow that the LoD to vary within the model and dependent on the current view. Early examples for these approaches can be found in terrain rendering systems [Duc+97; AH05]. However, also methods for arbitrary input meshes have been developed [Hop97; HSH10].

POINT-BASED METHODS As the name suggests, point-based approaches describe 3D geometry by points in the 3D space. More lightweight representations of such a point set can be derived by removing or averaging nearby points in the set. Traditionally points in 3D are then projected to image space where filtering and reconstruction are performed in order to fill the gaps between the projected points. The QSplat system introduced by Rusinkiewicz and Levoy [RL00] uses a hierarchy of spheres as basic rendering primitives. The hierarchy allows for LoD, backface and frustum culling. Pfister et al. [Pfi+00] introduce the concept of surfels as a term for a surface element in 3D space. These surface elements store positions, normals and textures. Surfels are rendered using an octree-based approach and splatting. In the following years, several more splatting techniques for point-based rendering were developed [SPL04; Bot+05]. A disadvantage of all of the aforementioned approaches is that the rendering processes are reasonably complex. Alternative and often faster are image-based reconstruction techniques such as the Pull-Push interpolation introduced by Marroquim et al. [MKC07]. In the context of this thesis, it has been used as a fast reconstruction method in the rendering framework introduced in Chapter 7. As point-based graphics is a wide and active field of research a general introduction is given in the book by Gross and Pfister [GP07]. Another survey on early point-based rendering techniques was introduced by Sainz and Pajarola [SP04].

VOXEL-BASED METHODS Voxel-based rendering systems realize another type of techniques that use a different representation as polygons. Just as points, voxels are also volumetric entities in space. However, in contrast to point-based approaches, they are generally represented as regular grid structure of different attributes. This regularity makes them well-suited for LoD approaches as coarser representations of a scene can be represented by downsampling the 3D grid to a grid with a lower resolution. Hence, common structures to store voxel dataset are octrees, i.e. hierarchies of grids. Often, these datasets are rendered using ray-based approaches. Building upon traditional grid traversal algorithms [AW87], Laine and Karras [LK11] introduce a very compact sparse voxel representation known as Sparse Voxel Octree (SVO) along with a highly optimized incremental traversal algorithm. Crassin et al. [Cra+09] traverse an octree using a kd-restart algorithm [FS05] in order to omit any need for a stack. Due to the enormous amount of parallel threads on the Graphics Processing Unit (GPU), maintaining a stack is challenging and introduces a potential performance bottleneck. In contrast to point-based approaches, where datasets can be generated by considering (random) points on a 3D surface or are a result of laser range scans of real objects, several articles deal with how to construct voxel representations from polygonal models – a process that is known as *voxelization*. In recent year several voxelization techniques using the GPU hardware were proposed. Systems such as Voxelpipe [Pan11], or the system proposed by Schwarz and Seidel [SS10], perform voxelization using an optimized triangle/box overlap test on the GPU. Other approaches by Dong et al. [Don+04] and Zang et al. [Zha+07] use a GPU accelerated rendering pipeline for performing voxelization. Both approaches render the scene from three sides, combining multiple slices through the model into a final voxel representation. However, this process has a negative impact on performance. Current OpenGL standards allow for writing to a 3D texture or linear video memory directly from the fragment shader. Without prejudging Chapter 5, this thesis introduces a system that is based-upon the voxelization process by Crassin and Green [CG12] using those features. In addition, this chapter proves that SVOs [LK11] are be a valuable tool for view-direction-based ray tracing systems.

INSIGHTS Approaches that pre-filter the function to get a multi-level representation of complex scenes and objects are actively used in graphics applications. For the sake of focusing on perception-driven approaches this section only introduces some of the general concepts. A more in-depth description of the related work in the broad field is given in the renowned books by Luebke et al. [Lue+03] and Akenine-Möller et al. [AHH08]. Given the brief overview of the various fields and important publications now allows us to look at those methods that augment these concepts by exploiting insights from human perception.

4.1.2 *Model-driven Approaches*

Early approaches in the field of perception-driven geometric simplification use rendered images of the models at different LoDs and compare these images using perceptual models in order to guide LoD selection and generation [Red97; LT00; LH01]. This process commonly results in *static* techniques that generate a fixed set of LoDs. An overview of such early LoD methods can be found in the book by Luebke et al. [Lue+03, pp. 264–278]. One example is the terrain rendering system by Scoggins et al. [SMM00]. It transforms terrain data to the frequency domain to investigate a relationship between sampling rate, viewing distance, object projection, and the expected image error caused by LoD approximations. The introduced image metric makes use of visual acuity (Section 3.1.2) and a Contrast Sensitivity Function (CSF) model, in this case the one described by Mannos and Sakrison (Section 3.1.3).

LOW-LEVEL PERCEPTION All of the presented systems so far attempt to measure the perceived quality of the output based on the view, or very simple measures of image contrast and the spatial frequency of the resulting LoD changes. However, they do not focus specifically on textures and effects caused by dynamic lighting as this needs a deeper knowledge of low-level perceptual processes. For that reason, Williams et al. [Wil+03] extends the simple mesh edge-collapsing techniques, discussed in the previous section, in order to estimate the degradation of textures and the induced lighting changes. Their technique creates view-dependent dynamic LoD representations, sensitive to silhouettes, underlying texture content, and illumination. It simplifies regions of imperceptibly low contrast first. Drettakis et al. [Dre+07] and Qu and Meyer [QM08] (Figure 37) further improve on Williams et al. by incorporating visual masking (Section 3.1.6). As this is computationally demanding, Qu and Meyer [QM08] accelerate this process by an off-line pre-processing step that computes an importance map which indicates the visual masking potential of a surface. To this end, they use a model derived from JPEG 2000 (Section 3.1.6) and the Sarnoff Visual Discrimination Metric (VDM) (Section 3.2). As detailed in Section 3.2, different high-level perceptual metrics exist to compare the visual quality of an image to ground truth data (Section 3.2). A couple of these metrics have been used to generate LoDs by rendering and comparing degenerated models. An extensive overview of investigated metrics applied to mesh compression and mesh watermarking is given in the reports from Corsini et al. [Cor+07; Cor+13].

Menzel and Guthe [MG10] present an alternative model that optimizes meshes specifically in low-contrast areas. Their main contribution is a method to move the error computation from image space to vertex space. This avoids costly per-pixel comparison. The idea is to determine the changes in contrast, curvature and lighting at each vertex after a simplification step. Moreover, they include measures of the interaction of spatial frequencies and orientations

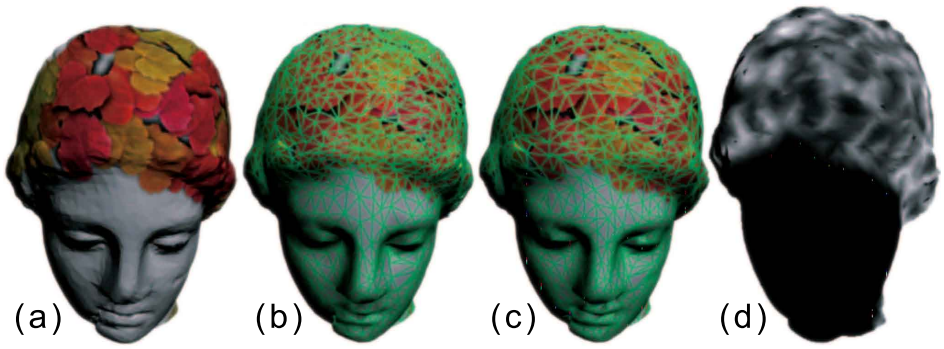


Figure 37: Perception-based mesh simplification by Qu and Meyer [QM08]. For a textured model (a) visual masking is evaluated (d). Compared to traditional simplification (b) including visual masking enables stronger simplification without affecting perceived mesh quality (c). *Image from Qu and Meyer* [QM08]

in order to account for visual masking. A simplification step is only applied if it can be considered imperceptible. Guo et al. [Guo+15] study the visibility of LoD distortions by asking subjects to mark visible ones on a mesh. The derived ground truth is used to evaluate different error metrics. In their study, perception-based metrics outperform purely geometry-based approaches. However, recently, Lavoué et al. [LLV16] concluded that purely image-based metrics including HDR-VDP2 (Section 3.2) perform sub-optimally. Based on these observations, Nader et al. [Nad+16b; Nad+16a] perform an experimental study of the HVS' low-level properties in order to derive a contrast sensitivity and contrast masking function. This allows them to compute a contrast threshold function per geometric face using Barten's CSF (Section 3.1.3). The contrast masking function that is explicitly looking at the visual regularity of surfaces allows for deriving a Just Noticeable Difference (JND) value and guiding the simplification process.

HIGH-LEVEL PERCEPTION AND ATTENTION Besides such measurements based on low- and high-level vision models, geometry can be simplified while attempting to preserve the salient features of the mesh using attention mechanisms. Those parts that probably draw visual attention should be degraded more slowly. An early approach for automatic LoD generation and selection based on attentional models is proposed by Horvitz and Lengyel [Eri97]. The authors evaluate the trade-off between the mesh's best visual quality and the computational savings using a cost-function that is based on mesh degradation and a probability distribution over the attentional focus of the viewer. Lee et al. [LVJ05] make use of a top-down attention model (Section 3.3.2) in order to preserve salient mesh features defining *mesh saliency*. This is based on the observation that *a substantial change in curvature* should be considered to result in high local saliency. For mesh simplification, mesh reduction is steered by evaluating such a geometric saliency. For partial shape matching of meshes, Gal et al. [GC06] define the mesh saliency as a function of geometric features that is determined by clustering regions of high curvature relative to their surroundings. Lavoué [Lav07] presents an extended curvature-based measure for model roughness and shows how mesh saliency can be applied to compute the visual masking potential of the geometry. Later, a study by Kim et al. [Kim+10] confirmed that *mesh saliency* better describes human fixations than random models of eye fixation, validating the importance of the local curvature measure. The approach by Wu et al. [Wu+13] extends the ideas by looking at two more aspects: *Local contrast* and *global rar-*

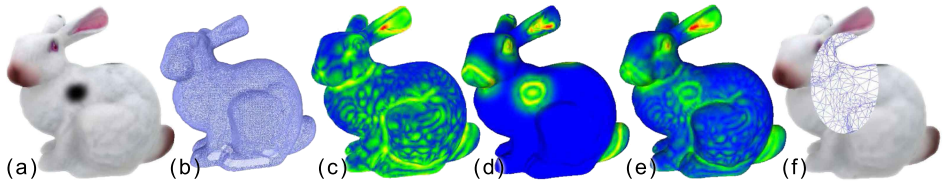


Figure 38: Mesh saliency method by Yang et al. [Yan+16]. The approach takes a textured mesh (a),(b) and measures local geometric entropy (c), color and intensity (d). The features are combined into a final saliency map (e) used to produce a simplified textured model (f). *Image from Yang et al.* [Yan+16]

ity based on the singleton hypothesis. This hypothesis states that the viewer's attention is drawn by stimuli that are locally unique and globally rare (Section 3.3.2). Hence, rare features should be degraded more slowly. To this end, the authors introduce a multi-scale shape descriptor in order to estimate saliency locally, and in a rotationally invariant way. Yang et al. [Yan+16] combine *mesh saliency* with texture contrast resulting in a saliency texture, which is used to simplify textured models (Figure 38).

In contrast to the aforementioned techniques, Ramanarayanan et al. [Ram+07] do not reduce the complexity of the model by geometric means; they reduce the complexity of the materials. Based on their Visual Equivalence Predictor (VEP) metric (Section 3.2), the authors show that the complexity of individual maps and materials can be significantly reduced, without sacrificing the visual appearance. The system by Koulieris et al. [Kou+14b] provides a LoD approach for materials, building upon the ideas of their high-level saliency predictor [Kou+14a]. The first type of information in this model accounts for the fact that an object pops out if it is rotated in a way that violates its expected posture. Other modalities derive a measure of an objects contextual isolation, i.e., is a specific object showing in parts of the scene not typically expected. This allows for continuous adaptation of material quality.

INSIGHTS Geometric techniques that reduce the scene's complexity drastically reduce the workload for geometry processing. Polygonal simplification processes are most common, and most of the more recent related work is incorporating perceptual models in order to enhance the simplification processes. Only a few systems make use of cross-modal effects. Such a system was presented by Grelaud et al. [Gre+09]. They use both audio and graphics to guide LoD selection and jointly adapt auditory and visual quality. Besides the limited knowledge in systems that rely on cross-modal effects, the boundaries of perceptual and attentional models are further blurred by coupling low-level knowledge of the HVS to attentional models. Moreover, improvements of LoD systems for other scene representations (e.g., voxels and points) and high-level scene properties (e.g., materials and lighting), will further improve on reduction rates and quality. [Wei+17]

4.1.3 *Gaze-Contingent Methods*

More than four decades ago Clark suggested that 3D objects can be simplified based on their eccentricity and velocity in the visual field [Cla76]. View-dependent geometric LoD is a typical example of gaze-contingent rendering. Perceptual models and gaze-contingent information can be used to reduce the quality of a scene representation in areas of lower acuity. Such a process

utilizes data from devices such as head and eye trackers as well as inertial measurement units often integrated into modern head-mounted devices [LaV+14; Ste+15].

POLYGONAL SIMPLIFICATION Ohshima et al. [OYT96] employ gaze-aware LoD rendering in a virtual environment. The gaze is used to interact with a discrete set of *static* precomputed presimplified models at different LoDs. Besides a simple model of the visual acuity and eccentricity (Section 3.1.2), the authors take additional perceptual clues from kinetic and binocular vision into account when selecting a model from the set of presimplified models. Interestingly, Ohshima et al. also experimented with saccadic suppression (Section 2.2.3): Rendering is suspended if the gaze movement exceeds velocities of $180°/s$. Furthermore, knowledge of the Depth-of-Field (DoF) is exploited in order to select from the simplified candidates. Unfortunately, a user study to judge the implications of this approach is missing. In contrast, Luebke et al. [Lue+00] simplify geometry progressively based on the gaze. The degree of mesh simplification is controlled by a perceptual model that exploits the limited visual acuity and the contrast sensitivity of the HVS in order to remain mostly visually imperceptible (Section 3.1.7). Reddy [Red97, pp. 105–129] proposes a two-stage approach for generating and selecting LoDs. In the offline stage, each object is analyzed in the spatial as well as the frequency domain to generate simplified model versions with defined maximum spatial frequencies. Also, models are analyzed considering color changes in the CIE LUV color space (Section 3.1.4). In the online stage, a perceptual model (including visual acuity) and a custom CSF are used to select the appropriate LoD based on the projected object rotation, relative size, the user's gaze direction, and other pre-computed object data. Along similar lines, Howlett et al. [HHO04; HHO05] use eye tracking in order to detect salient features that can be improved by better geometric approximations during a mesh simplification process. Murphy and Duchowski [MD01] propose a non-isotropic LoD rendering approach using eye tracking for meshes based on a spatial degradation function derived via a user study. Reddy [Red01] describes a system that recursively subdivides terrain meshes until the projected polygon size reaches an imperceptibility threshold that is coupled to a spatio-temporal CSF-model based on Mannos and Sakrison and a simple visual acuity model utilizing the Cortical Magnification Factor (CMF) (Section 3.1).

BEYOND THE ORDINARY A more recent approach apart from geometric simplification has been proposed by Papadopoulos and Kaufmann [PK13]. They use tracking in front of a large high-resolution display wall to adapt the visualization of gigapixel images to the user's physiological capabilities and visual field. Along similar lines, the next Chapter 5 introduces an approach to accelerate rendering on large high-resolution display walls using a hybrid representation of voxel and polygonal data with an LoD control that adaptively degrades visual quality based on the user's position and visual field. Most recently, Lindeberg [Lin16] uses filtering to simulate DoF in order to conceal artifacts arising from gaze-contingent geometric simplifications by utilizing a tessellation shader in the Unreal Engine. As detailed in Chapter 7, this thesis proposes to use DoF in order to hide undersampling artifacts in image space.

INSIGHTS Systems that adapt a scene's complexity according to the user's gaze have not seen much attention lately. Considering the limited rasterization and ray-casting performance achieved by previous graphics hardware generations, scene simplification techniques usually

led to huge speed-up factors. However, in current pipelines for real-time rendering, shading often dominates rendering costs [Vai+14; HGF14]. As already stated in a previously published state-of-the-art report [Wei+17], novel approaches targeting methods for gaze-contingent geometric simplification, tesselation and a gaze-aware adaptation of other features such as materials are foreseeable in the upcoming years.

4.2 SAMPLING ADAPTATION

Rendering, at its very core, is a sampling process. As such, samples must be generated, evaluated and combined via reconstruction with the aim to produce an output image. The following provides an overview of various perception-driven sampling strategies that attempt to adapt this sampling process in order to meet perceptual requirements. Although tailored rendering approaches exist, for example to adapt textures and draw line primitives [AHH08, p. 124-125], this thesis specifically focuses on sampling arbitrary 3D scenes. In addition, the sampling of higher-order functions such as light sources in a Global Illumination (GI) context is not discussed here. This work restricts itself to methods that attempt to optimize rendering by *screen-based sampling*. Here, modifications of the originally uniform sampling process may happen at (sub-)pixel level as well as at image level in the form of selective rendering. While the former methods are mainly concerned with reducing aliasing in image space, the latter may also take higher-level perceptual properties into account. This makes it possible to sample an image plane, potentially adaptively and progressively, based on knowledge about low-level, high-level and attentional models of the HVS with the goal to increase rendering efficiency and decouple rendering from a fix pixel grid.

4.2.1 *Sub-pixel Approaches*

Although samples can be taken on arbitrary positions for each pixel and the sampling patterns can vary between the pixels, the simplest form of screen-based sampling is to compute a single sample at the center of each pixel. However, doing so, all that is known for the pixel is whether or not a triangle covers the center and the color information of the triangle at precisely this position. As a results, undersampling is likely to occur. Consequently, it is beneficial to compute more sample-per-pixel (spp). This process is referred to as *super-sampling*.

STATIC AND (PSEUDO-)RANDOM SAMPLING The simplest form of super-sampling is Full-Scene Anti-Aliasing (FSAA). Here $n \times n$ samples are computed for each pixel at positions that are laid out in a grid-like fashion (Figure 39). The advantage of FSAA is simplicity and regular sampling patterns benefit from coherent computations and memory accesses. However, super-sampling the image plane is computationally demanding as all positions must be sampled and fully shaded. Besides, regular sampling patterns can lead to artifacts such as Moiré patterns or temporal flickering, artifacts to which the eye is highly sensitive. Therefore, other uniform patterns with slightly improved perceptual and performance properties have been developed. These include Quincunx, Checker, and Rooks patterns (Figure 39). These patterns attempt to alleviate some of the regularity. Naiman [Nai98] showed that the visibility of jaggedness of edges depends on their slope. A peak is found for edges with a slope of one

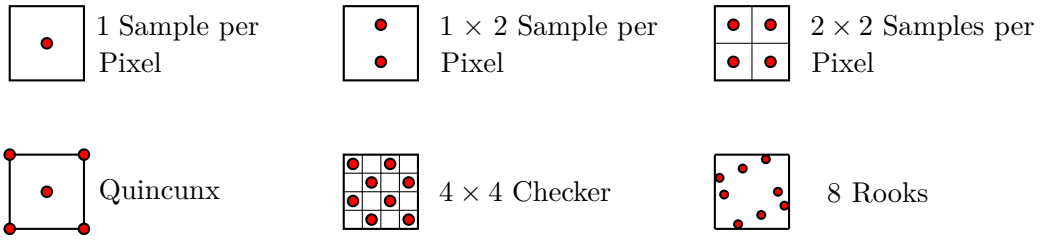


Figure 39: Regular sampling patterns can benefit from the coherency of memory accesses. However, can cause perceptually disturbing artifacts such as Moiré pattern due to their regularity. *Image after Akenine-Möller [AHH08, p. 127]*

(resulting in an edge at an angle of 45°). This is one of the main reasons why Quincux, Checkerboard, and Rooks, despite their regularity, provide an improved perceptual quality over simple FSAA patterns.

Another reason why aliasing artifacts are disturbing is the Vernier acuity. The HVS has the ability to tell if two lines aren't exactly aligned with each other at scales that exceed the visual acuity (Section 2.2.2). Due to those hyperacuity phenomena, jagged edges show, and as long as displays do not allow for achieving resolutions that exceed the Vernier acuity, Anti-Aliasing (AA) approaches are advantageous. However, as polygons can be arbitrarily small, the chances are that a regular sampling pattern can never capture them perfectly anyway. This will cause aliasing in one form or the other. Likewise, due to the regularity of these patterns, they can suffer from the same artifacts that can appear when using FSAA.

In order to counteract the appearance of regular patterns, it is often beneficial to use stochastic processes to select samples. This is called *stochastic sampling* and is driven by the spectral properties of the photosensitive cells' spatial distribution on the retina (Section 2.2.2). It is often better to sample using a random or pseudo-random pattern as images with noise often look better than aliased images [DW85]. Regular aliasing artifacts can be noticed more readily than noise – even though the noise may not represent the underlying scene any more accurately than the aliasing artifacts. On a lower physiological level, processes like lateral inhibition and mechanisms such as the recently discovered positive feedback loop that involves boosting the output of certain light receptors in the retina while damping down others further amplify color and intensity discontinuities caused by the patterns (Section 2.1.1). Also, the direction-dependent processing in the visual cortex is sensitive to regular structures and patterns. The neurons in the visual cortex are selective based on the orientation, the pattern, and the direction of a moving stimulus (Section 2.1.3). On the other hand, humans are able to tolerate surprisingly large amounts of noise in images [Hua65]. Also, it must be noted that the perception of noise in images is depended on the context of the noise. Lucassen et al. [LBR08] show that there are differences in the perception of colored symbols in front of a noisy background and vice versa. Just as the contrast sensitivity depends on the wavelength of the stimulus, the visual perception of noise also depends on the color of neighboring pixels [SK13a]. Interestingly, some studies report that the Stochastic Resonance (SR) induced by the noise can actually increase the readability of letters for the visually impaired [PCR00], or enhance motion discrimination [TDM16].

Unfortunately, a simple random sampling pattern can lead to non-optimal spatial coverage (Figure 40a). Pure random sampling can build clusters or leave large gaps. Therefore, even better patterns are often used in practice, most notable are Poisson-disc patterns (Fig-

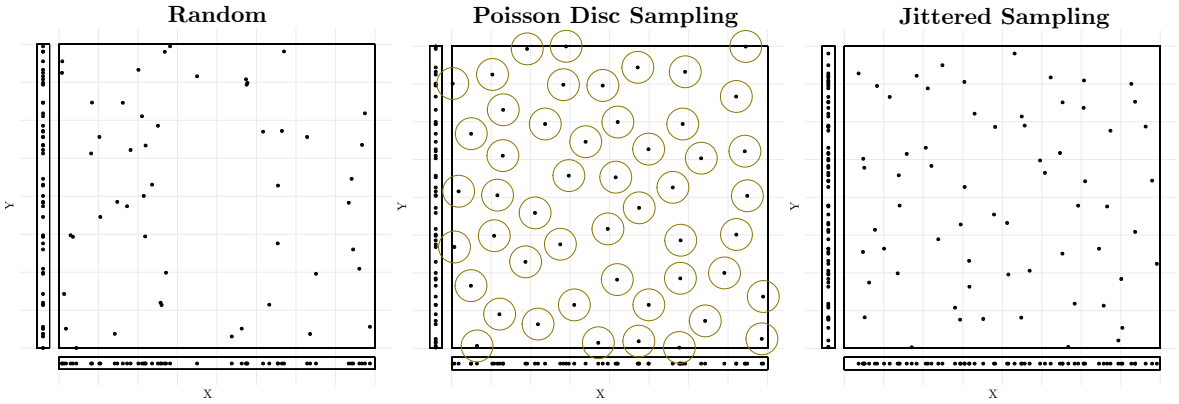


Figure 40: Different random sampling patterns in 2D and their 1D projections.

ure 40b). Yeelot [Yel83] discovered that the photoreceptors on the retina are distributed in a Poisson-disc fashion. Hence, Poisson disk distributions are known to work well in terms of perception, and, when sampling, they most closely resemble the blue noise power spectrum [Lag09]. Likewise, blue noise sampling can be used to model the spatial distribution of retinal cells [Lan+19]. Beside their perceptual properties, their even energy distribution across the spectrum make efficient techniques for blue noise sampling a general field of research for computer graphic applications [HSD13; Ahm+16; GF16] and various other domains [Yan+15]. However, generating a dense Poisson-disc sampling is somewhat computationally involved, especially in multiple dimensions and in a progressive manner. Hence, in order to achieve a better distribution of samples compared to random sampling more efficiently, supporting grid structures are widely used to create less clustered samples. The most basic approach doing so is *jittering*. Here samples are distributed in finer grid-cells build for each pixel. Afterward, the samples are randomly displaced within each cell (Figure 40c). Even better spatial distributions, especially considering the point sets' 1D projection can be obtained using *n-rooks* or techniques such as *multi-jittered sampling*. An introduction in generating those patterns can be found in the book by Suffer [Suf07, pp. 104-107].

Another prevalent type of approaches to generate samples is based on *low discrepancy sequences* (Figure 41). Remember, that good sampling patterns have a well-distributed but not uniform structure – samples do not clutter in certain areas and are not too far apart. The concept of the *discrepancy* gives one measure to describe the quality of such a distribution. The discrepancy of a point set can be computed by artificially dividing the domain into virtual regions. For each virtual region, the count of points inside and the volume of the region are compared. Here, for a good distribution, each given fraction of each volume should have roughly the same fraction of sample points inside [PH04, p. 316]. Different low-discrepancy sequences have been developed. An overview is given in Figure 41. These sequences are commonly used in GI renderers. One advantage is that their construction is computationally efficient. Moreover, they are easy to parameterize and samples can be obtained progressively. More details on generating those sequences are presented in the PBRT [PH04, p. 316ff].

Still, massive oversampling of each pixel is the gold standard to counteract aliasing but leads to highly increased computational costs, as each subpixel produces a shader call. Faster approximating approaches such as Multisampling Anti-Aliasing (MSAA) [Ake93] and its

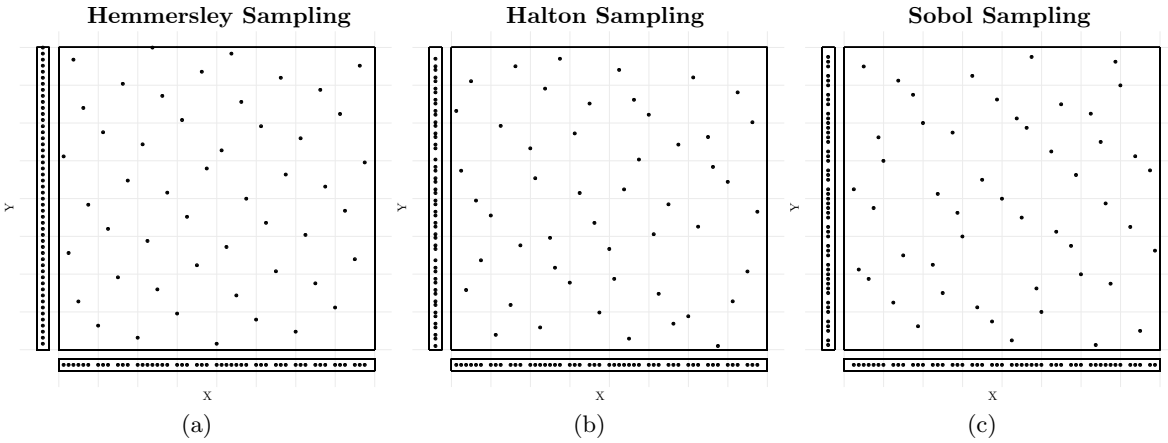


Figure 41: Different low-discrepancy sampling patterns in 2D and their 1D projections.

derivatives [You06; Rel11; Wan+15] decouple shading from coverage, depth and stencil information and can produce one shading color for all subpixel samples (Figure 42). This way, for each pixel only one shading sample for each touching primitive and associated coverage information is computed. Essentially AA happens only at polygon edges, which are potential areas of highly varying contrasts. A filtering process is used to combine the pixel values to a final color. Often a non-uniform sampling grid or centroid sampling [AHH08, p. 128] is used to improve on the coverage estimate. Due to its quality, simplicity, and efficiency, MSAA-based approaches have become an industry standard [Wan+15; Cra+15]. However, in the simplest form, these approaches fail in regions with transparencies, shadows and high-frequency color changes, which can be highly critical for perception. In addition, MSAA approaches have limits when used in combination with methodologies such as *deferred shading*.

Shading for high-quality 3D scenes becomes more and more complex, and visually appealing scenes often contain multiple light sources. However, when using traditional forward rendering approaches, hidden surfaces cause wasted shading operations [HH04]. In contrast to forward rendering, *deferred shading* makes it possible to perform shading computations only on the visible surfaces. This widely used rendering technique decouples visibility computation from shading by rendering different components (depth, normals, albedo, textures, etc.) of the scene into a discrete G-buffer. Afterwards, only the visible surfaces in the G-Buffer become the target for complex shading and lighting operations in a second render pass. However, once a scene has been rasterized into a discrete G-buffer, it becomes impossible to resolve finer geometric details. The process of combining multiple samples for AA happens after accumulation [HH04]. This makes it more challenging to compute coverage information and the number of touching primitives for each pixel.

In order to overcome this limitation and to allow for decoupling shading rate from geometric sampling rate, the technique by Crassin et al. [Cra+15] uses the rasterization pipeline to generate a compact, pre-filtered geometric representation that is stored for each pixel. Wang et al. [Wan+15] show that these *consolidated surface coverage* and a single decoupled depth value can be used for an optimized management for traditional forward rasterization as well. Still, all MSAA methods compute a single or a limited number of shaded samples and are usually tied to a fix resolution of the coverage information. Hence, several other strategies using temporally jittered pixel locations [HA90] and pseudo-random patterns [Jim+11] have

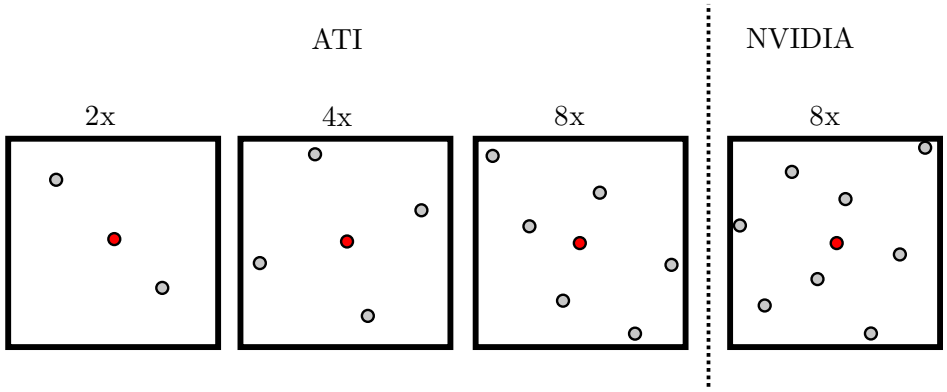


Figure 42: Different MSAA sampling patterns per pixel at varying sampling rates on ATI and NVIDIA hardware. The shading sample is marked as a red dot, the gray samples are positional samples. The latter samples are only used to determine the geometries coverage. This potentially reduces costly shading operations for those samples. Note that these patterns are repeated/replicated for each pixel, but must be considered as parts in a larger pixel grid. *Image after Akenine-Möller [AHH08, p. 129]*

been developed. A general overview of those AA approaches can be found in work by Maule et al. [Mau+12]. Also, a more advanced version of deferred shading, namely *deferred lighting* [Eng09] is possible, due to the processing power of modern GPUs.

Deferred lighting decouples the costly lighting integrations. First, the 3D scene is rendered into a G-buffer. A second pass computes the diffuse and specular irradiance values for each visible surface and stores it to lighting buffers. Finally, in a third render pass, the scene's geometry is rasterized again. This time, lighting information can be read back from the lighting buffers. In this way MSAA becomes possible for the last pass [Eng09]. However, deferred lighting makes it necessary to render the scene geometry twice. Also, handling the lighting buffers, which separate diffuse and specular irradiance values, becomes more complex.

PROGRESSIVE SAMPLING Ideally, sample generation can also be driven progressively using perceptual implications in order to control how to adapt sampling based on perceptual requirements. However, this makes it necessary to (efficiently) resample an image in areas that matter most. Unfortunately, an efficient random (re-)sampling of individual pixels is hardly possible with rasterization, which is always tied to a fixed resolution. Better suited for such perception-driven rendering systems are ray-based methods. An early approach involving such perceptual aspects was introduced by Mitchell et al. [Mit87]. After initially sampling the image plane with n spp, a simple error metric using contrast thresholds for the RGB color values is used to guide a resampling process. Given the n samples, a contrast for each intensity I of each color channel r,g, and b is computed as simple Michelson contrast:

$$C = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}$$

Now, these three contrasts (for r,g,b) are each tested against separate thresholds. Mitchell et al. use $r = 0.4$, $g = 0.3$ and $b = 0.6$ respectively. Supersampling is only performed if one of the contrasts exceeds the thresholds. Please note that this test is meant to take into account the perceptual distribution of the photosensitive cells. *"Because green-sensitive cone cells are far more common in the human retina, sensitivity to green-colored noise is twice as great as*

to red-colored noise and four times greater than to blue colored noise" [Mit87] (Section 3.1.4). Painter and Sloan [PS89] presented hierarchical adaptive stochastic sampling for ray tracing that works in a progressive manner. Confidence and coverage-based stopping criteria are used to control sampling. By using a progressively updated mean values of the samples also the size of the confidence interval can be adapted to stop drawing samples if these are considered to be below a supra-threshold for the user. Based on this approach Meyer et al. [ML92] focus on the change in the acuity of color perception with increasing eccentricities. Here, images are transformed into the frequency domain and a custom LMS color space is used to account for the different densities of the photosensitive cells and the color opponent processing in the HVS (Section 3.1.4). Along similar lines, Bolin and Meyer [BM95] use a ray tracer that inspects images in the frequency domain. It is coupled to an adaptive quadtree in image space and a simple vision model, that controls where new rays are cast by accounting for the perception of colors and frequency of the content within each image block (Section 3.1.4). Later, Bolin and Meyer [BM98] have extended their work by developing a more elaborate adaptive sampling algorithm, based on a simplified model of Sarnoff's VDM (Section 3.2).

A more modern approaches by Jin et al. [Jin+09] propose an adaptive supersampling scheme for efficient ray tracing on many-core architectures such as GPUs. This system uses subpixel tests for geometric attributes, color and contrast gradients between adjacent pixels. When discordance is found for at least one of these measurements, the subpixel is scheduled for further sampling. Here a similar approach to Mitchel [Mit87] is used, testing the individual color channels against separate thresholds (Section 3.1.4). However, these thresholds are scaled by a factor computed from the geometric attributes in order to further increase control over the super-sampling process. Shevtsov et al. [SLR10] propose a SIMD-friendly adaptive sampling scheme for packet-based ray tracing, where a three-pass architecture is described. After an initial sampling step, a discontinuity detection performs a pair-wise computation of gradients based on luminance or per color channel, again similar to Mitchel [Mit87].

Unfortunately, efficient (re-)sampling of individual pixels is hardly possible using rasterization and less interesting due to the raw processing power of modern rasterization pipelines. Only a few systems allow for adaptive supersampling using rasterization. One of those approaches is realized by Perrson [Per07]. As MSAA does only shade one spp along the geometric seams it can, for example, fail for specular reflections on bump mapped surfaces. Such *interior surfaces effects* are solely computed in the pixel shader. Perrson makes use of gradient functions to decide if a pixel needs more shading samples selectively. In this case, those are computed and averaged inside the pixel shader. Siegl et al. [Sie+13] apply multiple render passes. First, only the simple triangles, i.e. those that do not contain challenging appearance properties (e.g. specular highlights), are rendered using MSAA. Afterward, FSAA is used in a second render pass in order to improve the quality for the remaining challenging surfaces. However, at this point, both approaches are solely function-driven and do not consider perceptual implications to further improve image quality. Most recently, a hybrid system called Adaptive Temporal Anti-aliasing (ATAA) [Mar+18] has been proposed that combines the speed of rasterization with the flexibility of ray tracing. After initially rasterizing the frame and by using information from the previous frame, critical discontinuities are detected in image space. Now, to locally increase the sampling densities, rays are cast for those challenging regions. Note, that this approach exploits Temporal Coherence (TC) (Section 4.3). Also, the authors show that post-processing techniques can be applied for image regions that were occluded in the previous frame (Section 4.4).

INSIGHTS What should become apparent from the latest related work, is that the methods for AA in real-time rendering have started to employ many strategies simultaneously [Pet15; Ped16]. This thesis argues that even novel approaches such as ATAA can further be improved by considering perceptual limitations. Likewise, the availability of tailored hardware accelerated ray tracing [Sti18; Ima19] and the increase in sampling flexibility will drive the development of novel AA methodologies. Novel techniques to distribute computational resources will aid progressive sampling techniques.

4.2.2 *Selective Rendering*

Methods from the field of selective rendering take perceptual implications of the generated image into account in order to put more computational effort into important regions of an image. However, in contrast to the methods introduced in the previous section, selective rendering methods look at the "bigger picture". Here, perception-critical *regions* are determined by detecting salient features, such as regions of high-contrast, noise or higher-level features.

LOW-LEVEL PERCEPTION FOR PRODUCTION RENDERING Selective rendering is often used as a flexible rendering method in stochastic ray-based renderers in order to steer the number of sample-per-pixel (spp) or recursion depth. A common goal is to obtain an image that is perceptually indistinguishable from a fully converged, but expensive, rendering solution.

For production rendering, the method by Walter [Wal98, p. 87ff] controls the kernel size in a photon tracing framework by considering luminance and chrominance influences on perception. Guo [Guo98] developed a progressive refinement algorithm for Monte-Carlo rendering that stops refining image blocks based on a CSF model (Section 3.1.3). In the work by Ferwerda et al. [Fer+96], the authors take a closer look at the eye's adaptation process. By performing a psychophysical experiment, they developed a model to display and combine the results of GI simulations at different illumination levels. The work by Myszkowski et al. [Mys98; Hab+01] uses the Daly's Visible Differences Predictor (VDP) (Section 3.2.1) as an image metric to selectively stop rendering in a Monte-Carlo path tracer for GI rendering. Farrugia et al. [FP04] make use of a perceptually inspired metric based on the adaptation of the eye in order to progressively render and stop GI computation earlier when the perceptual quality is sufficient. Yu et al. [Yu+09] analyze the influence of visibility approximations on the perception of GI renderings. They also conduct a study on the perceived realism of scenes rendered with imperfect visibility, (directional) ambient occlusion and another study where renderings using visibility approximations are compared to reference renderings. The authors conclude that using appropriate visibility, approximations can lead to results that are perceived as realistic despite the fact that individual differences between the approximate and reference renderings are visible in a direct comparison. Dachsbacher [Dac11] shows how the analysis of visibility configurations can be used for adapting the sampling process in ray tracing, improving perceptually motivated LoD approaches in real-time rendering and extending visibility classifications in radiosity methods.

HIGH-LEVEL PERCEPTION AND ATTENTION FOR PRODUCTION RENDERING Models of visual attention (Section 3.3) can also be used to improve the quality of GI renderings for animations and dynamic scenes [Mys02]. As rendering quality can be decreased for moving objects or patterns, Myszkowski et al. [Mys02] use temporal reprojection alongside a temporal extension of the VDP (Section 3.2.1) called Animation Quality Metric (AQM), which accounts for motion when computing new samples. Jarabo et al. [Jar+12] take a closer look at the importance of accurate lighting and its effect on perceived realism when rendering crowds. They employ an approximation based on spherical harmonics, which is used to compute a temporal interpolation of the full radiance transfer matrix. The essential factors influencing scene fidelity found by the authors are geometric complexity, the presence or absence of color, the movement of individual crowd entities as well as the movement of the crowd as a whole; both known causes for *crowding* (Section 2.2.4).

The perceptual importance of the final image is often approximated by saliency extracted from previews rendered at lower quality, where the initial image estimate requires at least one spp. For decreased computation times, Longhurst et al. [LDC06] present a method that computes such a preview frame by rasterization. This frame is used to extract saliency including different low-level features such as edges, contrasts, motion, depth, color discrepancy, and scene habituation. The generated saliency map is used to steer the number of samples distributed on each pixel of the image. However, this way the approach fails when a high (re-)sampling weight is needed for phenomena, such as caustics. Such optical phenomena arise when light beams are focused on specific scene elements. Locally focused light can either emerge from the refraction of light beams in transparent objects, or, from bundles of light beams that are reflected back in the scene, e.g. emerging from convex glossy surfaces. Computing such phenomena with rasterization is very challenging, because the global light transport between objects in the scene must be considered. Hence, such effects are missing in the preview frame and cannot be considered for the saliency map. In contrast, Cater et al. [CCW03] and Sundstedt et al. [Sun+05] do not focus on low-level features such as edges and contrast but selectively render task-relevant salient objects and features in high-quality and reduce rendering quality for the remaining parts by adapting the resolution or the number of rays per pixel. In their studies the subjects were not able to distinguish high-fidelity rendering from selective rendering results. The experiments demonstrate the suitability of perceptual rendering if selective attention can be predicted.

One aspect of a saliency computation using attentional models (Section 3.3) is that movement in the background of a scene may substantially influence how humans perceive foreground objects, for example when objects start moving in the midst of a sequence. Yee et al. [YPG01] use a model of visual attention for moving objects in order to accelerate rendering of animations. To this end, they introduce a method to compute a spatiotemporal error tolerance map, based on a velocity-dependent CSF. This CSF is augmented by a top-down model of visual attention (Section 3.3.2) to account for the tracking behavior of the eye when guiding the sampling of a GI renderer. Another system that makes use of attentional models has been developed by Chalmers et al. [CDS06]. The authors investigate several ideas such as importance-based sampling for on-screen-distractors, for example sound-emitting objects. Hasic et al. [HCS10] show the importance of visual tasks and motion for selective rendering, as both attract the viewer's attention. They present various types of movements with varied accelerations in a psychophysical experiment to a group of subjects.

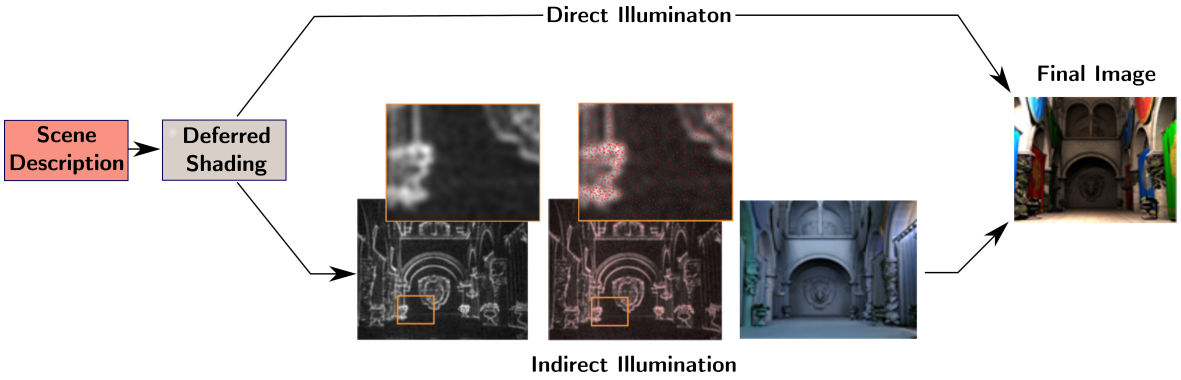


Figure 43: Sampling adaptation method by Galea et al. [GDS14]. A deferred rendering system decouples direct and indirect illumination components. A saliency map (a) is computed and sparsely evaluated (b) to accelerate the computation of the indirect illumination. Inpainting computes a dense representation of indirect lighting that is combined with the direct lighting for the final image. *Image from Galea et al.* [GDS14]

INTERACTIVE RENDERING Besides its extensive use, a disadvantage of rasterization over ray-based approaches is that, for efficiency, rasterization and the corresponding shading pipelines are traditionally tied to a fix resolution. In recent years, several approaches have been introduced that allow for *multi-rate* and *multi-resolution shading*: an enabling technology for perception-driven selective rendering systems using rasterization. Clarberg et al. [Cla+14] propose a modification to current rendering pipelines, which enables varying shading rates on a per-patch basis to reuse shading results within tessellated primitives. He et al. [HGF14] present a system that uses a coarse grid in order to reuse shading samples within grid cells. Vaidyanathan et al. [Vai+14] introduce coarse pixels and tiles, which allow shading samples to be reused in a multi-level grid-like fashion. NVIDIA GPUs support a multi-resolution shading approach, drawing different resolutions within a single pass [Ree15]. The latest NVIDIA GPUs generation allows also for multi-rate shading. NVIDIA calls this extension Variable Rate Shading (VRS) or NVIDIA Adaptive Shading (NAS) [LJ19]. Essentially, pixel shading operations can be applied to blocks of pixels. Here, similar to MSAA, the visibility samples that are computed in the full resolution are used to improve the image quality when interpolating coarser shading samples. Even though, as presented in the next section, these techniques are nowadays widely used for gaze-contingent rendering, they are rarely used for solely perceptual model-based approaches that do no consider active inputs. An example of such a system has been proposed by Galea et al. [GDS14]. They describe a GPU-based selective rendering algorithm for high-quality rasterization. Their sparse sampling approach employs a saliency model in order to evaluate only a set of sparse sample locations which are used to compute an indirect lighting solution that is perceptually equivalent to full sampling. An inpainting algorithm is used to reconstruct a dense representation of the indirect lighting component, which is then combined with direct lighting in order to produce the final image (Figure 43).

Ray tracing systems are often solely used to get the highest image quality for production rendering, even though interactive and real-time frame rates are achievable even for complex scenes [ALK12; PKC15]. Still, ray-based approaches do not yet reach the performance and convenience of rasterization when it comes to real-time rendering. In order to meet performance requirements, a hybrid approach for Head-Mounted Displays (HMDs) was introduced

by Pohl et al. [Poh+15]. Their system deploys rasterization combined with ray tracing on the CPU in order to reduce the sampling rate for areas outside the lens's center. This way they can adapt rendering in HMDs by exploiting the lens astigmatism. Lens astigmatism is a property of an optical system that leads to a decrease in image quality towards outer regions. Also, all of the aforementioned approaches treat the HVS as a single-eyed system, though healthy humans are capable of stereopsis due to binocular vision. Rendering images for both eyes independently doubles the computational effort. The ray tracing approach by Lo et al. [Lo+10] exploits perceptual limits that arise from the brain being able to fuse information from both eyes separately. They show that the resolution could be reduced by a factor of six of one of the images of a stereo pair without being noticed by the viewer. The authors also observed that shadow and disparity cues perform equally well when judging depth. Currently dedicated hardware for ray tracing, such as NVIDIA RTX [Sti18], is becoming available. Ray tracing is integrated into high-performance rendering pipelines [Pha18; Ben19]. Hence, ray tracing could become the primary algorithm to compute visibility and light transport, even for interactive rendering.

MULTI-MODAL INTERACTION Instead of only accounting for visual perception, several selective rendering systems have been introduced that also consider multi-modal aspects. A survey by Hulusić et al. [Hul+12] gives information on the perceptual and cross-modal influences that have to be considered in the course of generating spatialized sound. Harvey et al. [Har+16] investigate the effect of spatialized directional sound on the visual attention of a user towards certain objects contained in the rendered imagery. Hulusić et al. [Hul+09] show that the beat-rate of an audio cue has a substantial impact on viewer perception of a video and video frame rate, allowing for the manipulation of the temporal visual perception. Bonneel et al. [Bon+10] analyze how the auditory and visual LoD influence the perceived quality of audio-visual rendering methods. They show strong interactions between auditory and visual LoDs in the process of material similarity perception.

INSIGHTS In conclusion, the current state-of-the-art shows that perception-based approaches that rely on models of the HVS have traditionally been used for stochastic ray tracing and in GI computation, for example to adapt sampling of path tracing. Due to ever-increasing processing powers, these methods are on the brink of appearing in real-time ray tracing systems as well. Moreover, and in contrast to methods targeting the subpixel features, deferred rendering systems and developments on multi-resolution shading allow efficient selective rendering using rasterization. Those methods will be further improved to enhance the visual quality and performance in consumer level Virtual Reality (VR) and Augmented Reality (AR) devices. Along similar lines, further exploiting other perceptual channels and their cross-modal interaction will continue to improve presence in virtual environments and help to increase the overall performance of modern rendering systems.

4.2.3 *Gaze-contingent Methods*

Active measures of a user's gaze allows to exploit more limitations of visual perception. Adaptation of the sampling and shading quality is an important aspect of gaze-contingent rendering systems. Early works in gaze-contingent rendering primarily observed the general detectabil-

ity and influence of a dynamic quality degradation on visual performance [PP99; PLN01; Nik+04; Dor+06]. Supported by the aforementioned findings, a large body of work focuses on gaze-contingent techniques that exploit the limitations of the human eye by omitting details in the peripheral visual field that are largely imperceptible. The common goal of such techniques is to exploit the spatial fall-off of the visual acuity (Section 2.2.2). This dynamic adaptation of rendering based on the user's retinal capabilities by employing gaze-contingent methods is also known as *foveated rendering*. Often both terminologies, *gaze-contingent* and *foveated rendering*, are used interchangeably. However, strictly speaking, the term foveated rendering places a specific focus on adapting the *sampling* to the retinal capabilities, ideally so that the users do not notice the adaptation. Gaze-contingent rendering, on the other hand, has a broader scope and includes a greater field of techniques, such as methods that adapt the LoD of polygonal representations (Section 4.1.3). In the following, the presented approaches are structured based on the used rendering approach – rasterization, ray tracing or hybrid approaches.

RASTERIZATION-BASED APPROACHES An early rasterization-based system was presented by Guenter et al. [Gue+12]. It simulates the acuity fall-off by rendering three nested layers around the PoR of increasing angular diameter and decreasing resolution. These layers are blended in order to obtain the final image (Figure 44). For the decrease in resolution, a model based on the CMF is employed (Section 3.1.2). This technique achieves impressive shading reductions but also introduces overheads by repeating the rasterization for each nested layer. However, a continuous adaptation of sampling rates and shading complexity over the image plane for rasterization requires efficient *multi-rate* and *multi-resolution shading*. To this end, Vaidyanathan et al. [Vai+14] tested their approach for multi-rate shading in a foveated rendering prototype by using a simplified acuity model (Section 4.2.2). Assuming a fix PoR and a constant radial acuity function, shading is computed at full-resolution in the foveal region and at a lower rate towards the periphery. A practical implementation of multi-rate shading in the Source Engine™ by the Valve Corporation [Vla16] showed that in this way rendering performance can be increased by 10-15 %. Supported by these developments, the work by Stengel et al. [Ste+16] presents a foveated rendering methodology for a deferred shading pipeline (Figure 45). While fully sampling the G-buffer, the actual shaded fragments are selected by a stochastic sampling pattern that is controlled by the user's gaze and the gaze velocity. The authors also point out the importance of contrast and brightness perception by specifically shading samples that have a high saliency or expose high contrasts. For such regions, contrast sensitivity and thus visual acuity remain to be high (Section 2.2.2). Finally, an image reconstruction method based on Pull-Push Interpolation (Section 4.1.1) allows images to be generated that are *perceptually equal* to images rendered with full per-pixel shading but at significantly reduced shading costs – the most computationally demanding part of modern image synthesis algorithms. Later that year, a framework by NVIDIA was introduced by Patney et al. [Pat+16b]. They carefully investigate the impact of several effects induced by quality degeneration in the periphery by distorting images. The result of these preliminary studies led them to develop a foveated renderer with MSAA as well as a saccade-aware Temporal Anti-Aliasing (TAA) [Kar14] strategy in order to improve temporal stability and suppress aliasing artifacts critical to peripheral vision (Figure 46). This system does not attempt to detect and individually create more shaded samples in salient image regions, such as regions with high contrast, but rather provides a post-processing approach in order to boost peripheral contrasts and thus counteracts the losses of contrast sensitivity

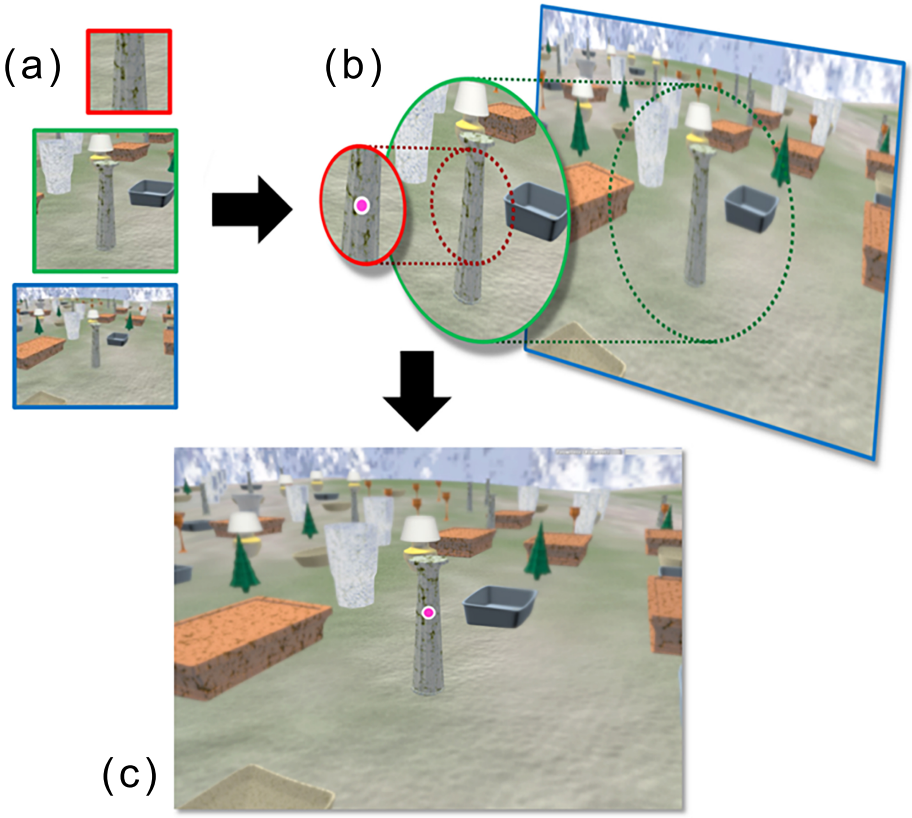


Figure 44: Foveated 3D Graphics by Guenter et al. [Gue+12]. Rasterization is performed at three different resolutions (d) according to acuity fall-off across the visual field. This approach reduces the number of shaded pixels. The results are then blended together (b). The combined image (c) approximates acuity fall-off and is faster to compute than traditional full-resolution rendering. *Image from Guenter et al.* [Gue+12]

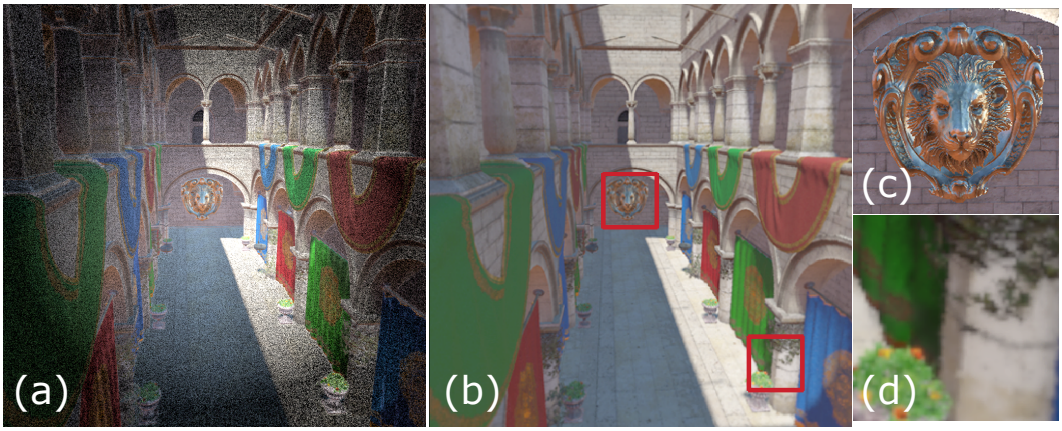


Figure 45: Gaze-contingent adaptive sampling by Stengel et al. [Ste+16]. Incorporating visual cues such as acuity, eye motion, adaptation and contrast, a perceptually-adaptive sampling pattern is computed and used for sparse shading (a). Fast image interpolation (b) achieves the same perceived quality at a fraction of the costs of shading each fragment. The resulting image contains high detail in the foveal region (c) and reduced detail in the periphery (d). *Image from Stengel et al.* [Ste+16]

Figure 46: Image from the foveated renderer by Patney et al. [Pat+16b] This system uses multi-resolution shading with quality degradation as a result in the peripheral visual field in preference to computing efficiency. *Image from Patney et al. [Pat+16b]*

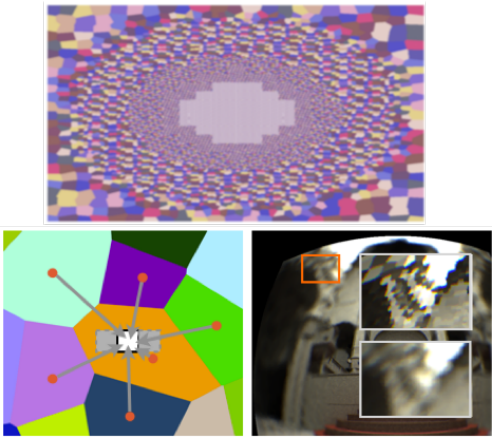


Figure 47: The foveated renderer by Fujita and Harrada [FH14] uses a precomputed Voronoi pattern (top) to reconstruct samples and thus quality degradation in the visual periphery (bottom). *Image from Fujita and Harrada [FH14]*

for peripheral vision. However, the method by Patney et al. is constrained by GPU design, thus it only offers a theoretical saving rather than actual performance (frame rates) benefits [Sun+17].

Meng et al. [Men+18] presented a log-space transform which synthesizes images in a foveated fashion. The system uses TAA in order to increase temporal stability. The log-space transform reduces shading by transforming and retransforming the G-Buffer of a deferred rendering pipeline (Figure 48). Although the results presented are impressive, their method still suffers from temporal artifacts and view-dependent inconsistencies for glossy specular reflections. All of these methods require at least either the visibility computations at the full resolution to be performed which helps to reduce shading calculations, or the scene needs to be rendered multiple times – admittedly at varying resolutions. What also becomes apparent here is that methods either attempt to detect and increase shading fidelity for salient parts in the image (such as Stengel et al. [Ste+16]), or rely on post-processing techniques (such as Patney et al. [Pat+16a] and Meng et al. [Men+18]) in order to conceal artifacts in the peripheral visual field. All methods rely on TC methods to increase the temporal stability in the periphery. TC methods are discussed in the next Section 4.3.

RAY-BASED APPROACHES While rendering to HMDs is mainly based on rasterization due to performance considerations, ray tracing has several advantages when it comes to stereo rendering, wide Fields of View (FoVs), correcting chromatic aberrations and low latency rendering [Hun15]. Especially important for perception-driven sampling strategies is the ray tracing's ability to distribute samples freely on the screen and the inherent possibility of creating high-quality renderings, crucial to achieving a good presence in VR worlds [Toc16, ch. 3.3]. However, ray tracing has been mainly thwarted by its own performance as it is challenging to achieve the same speed as rasterization without specific hardware acceleration, such as NVIDIA RTX [Sti18].

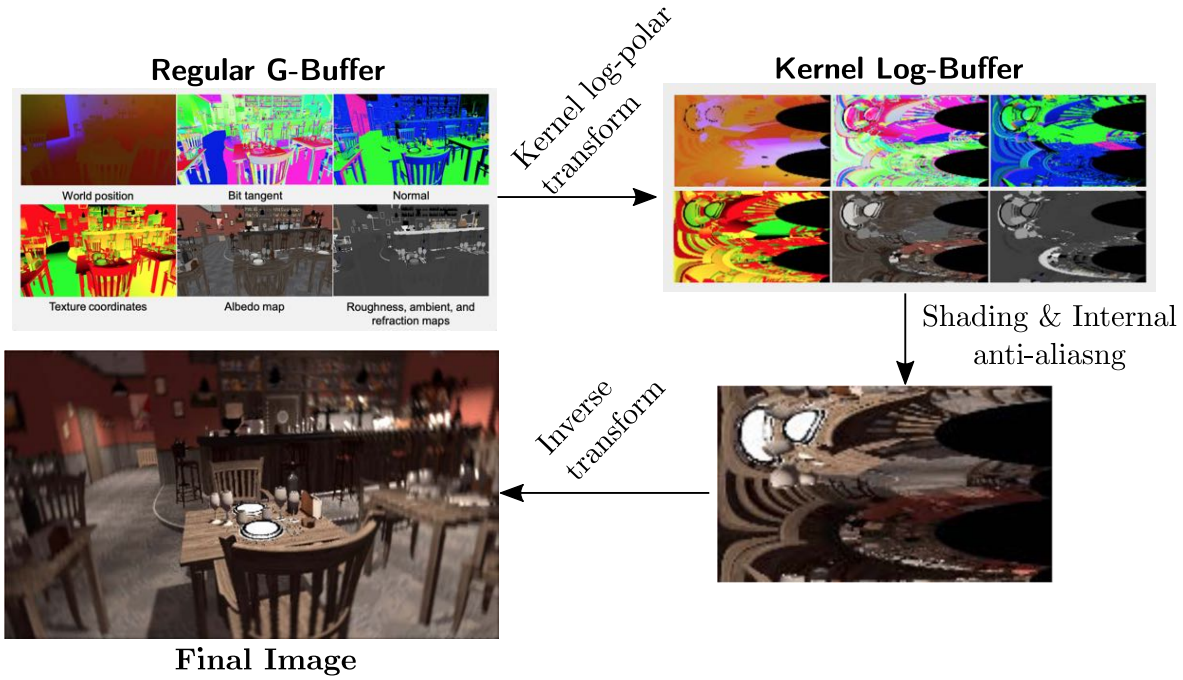


Figure 48: The work by Meng et al. [Men+18] transforms and retransforms a regular G-buffer from a deferred shading pipeline to a kernel log-space to allow for more efficient shading computations. *Image adapted from Meng et al.* [Men+18]

An early system for ray-based volume rendering that adapts sampling using eye tracking was developed by Levoy and Whitaker [LW90]. Here, the number of rays cast through the image plane and the number of samples drawn along each ray are adapted based on the tracking input. In order to reconstruct dense images from the potentially sparsely sampled image plane, this work adapts masks and filtering kernels to the eccentricity. An approach introduced in work by Murphy et al. [MD07] uses a precomputed mesh specifying ray locations for sampling and increasing the sampling density near object silhouettes and regions with high contrast. However, this method cannot accelerate rendering and does not reproject samples from one frame to the next in order to improve image quality. Also, handling the mesh is computationally involved. Another approach to foveated ray tracing was presented in the work by Fujita and Harrada [FH14]. A precomputed sampling pattern together with a k Nearest-Neighbors (kNN) scheme is used to reconstruct the image from sparse samples (Figure 47). However, the proposed system shows artifacts and does not consider the eye's sensitivity to contrasts. In addition, there is no user study. Other recent developments for real-time ray tracing are presented in the system Hierarchical Visibility for Virtual Reality (HVVR) by Hunt et al. [HMN18; Hun17]. The authors present a hierarchical tile-based rendering approach on the CPU in combination with a fast ray caster on the GPU. Although currently not implemented to its full extent nor adequately evaluated regarding perceptual implications, it uses the ideas from Patney et al. [Pat+16b], boosting image contrasts in peripheral vision in its foveated mode.

Foveation methods can also be used to accelerate high-quality rendering on large-high resolution display walls [Rot+15]. In this system, some elements of the method that is presented in Chapter 5 were integrated to guide the sampling process of a path tracer. However, this work by Roth et al. uses a hand-held device in order to select the region that needs the

highest visual acuity as the system is not responsive enough to be used with eye or head tracking. Another system that attempts to use foveated rendering for GI rendering in HMDs is presented in the work by Koskela et al. [Kos+17]. Most notable in this work is that the authors use the inversion method in order to derive a Probability Density Function (PDF) from the visual acuity model by Reddy (Section 3.1.2). This PDF enables to generate acuity-dependent samples in a stochastic fashion. However, their current GI rendering system is far too unresponsive. Hence, users notice the quality adaption. Nonetheless, the accompanying preliminary study shows that users preferred this foveated mode over sampling with random patterns.

All of the approaches so far have only considered the amount foveation, i.e. the loss of resolution for peripheral vision, as a function of the spatial resolution of the different stimuli. However, the visibility of a stimulus is highly affected by its luminance and its contrast. Most recently, Tursun et al. [Tur+19] introduced a foveated rendering system that considers both features in order to derive a more precise predictor of the foveated rendering parameters. To this end, a band-limited contrast is computed in an image pyramid. Next, a custom CSF is described to account for the degradation of visual performance concerning a stimulis' contrast at increasing eccentricities. Here, also a simple model for visual masking is incorporated. Eventually, this process is used to determine an estimator for an eccentricity and image patch-dependent resolution reduction factor. For validity, the model is tested with different rendering techniques and shading models. For foveated ray tracing, it is reported that this method allows for a decrease of 53% of primary rays. In comparison, to stay below the detection threshold with a standard foveated rendering without using the luminance and contrast-aware model, the authors report a reduction of 45%. Potentially higher reduction rates are possible if TC is also taken into consideration.

HYBRID APPROACHES Hybrid approaches attempt to combine rasterization and ray tracing in a single system. Initially, Pohl et al. [Poh+15] presented such a system in order to exploit lens astigmatism in HMDs (Section 4.2.2). Later, Pohl et al. [PZB16] extended their system to also include eye tracking input. Most recently, Friston et al. [FRS19] propose *perceptual rasterization*, another hybrid approach to foveated rendering. Essentially, images are synthesized by warping primitives and their convex hulls, the so-called *primitive-pixel bounds*, in a geometry shader. This warping is performed according to foveation parameters and the requirements of the HMD. In the fragment shader a call is generated for all fragments that reside in the primitive-pixel bounds. Finally, a ray casting step in this shader allows for intersecting the respective warped primitives. Also, this allows for a rolling rasterization in order to update HMDs with rolling displays, such as the Oculus Rift DK2. This way each column of pixels can be updated at a different point in time. While this approach is very promising, implementation is not as straightforward as regular ray tracing. In order to produce appealing images, for secondary effects, the same techniques as performed for rasterization need to be applied. Also, due to the inherent overhead, the method does perform worse than regular rasterization in the full resolution. This is a common challenge for gaze-contingent rendering systems (Chapter 7).

BEYOND THE ORDINARY On the other end, researchers have started to use gaze-contingent approaches for different rendering and data methodologies. One example in this field is the rendering of lightfield data to respective displays as performed by Sun et al. [Sun+17]. While

lightfields commonly support focal cues, they are usually not processed adaptively to the retinal capabilities. In the work by Sun et al., a lightfield is sampled with a function that accounts for eccentricity. This process is implemented using a GPU-based ray tracer. Finally, a 4D Gaussian radial basis function is used to reconstruct a dense dataset from sparse samples. This way, frame times and sampling rates can be reduced significantly.

INSIGHTS Due to vast improvements in eye tracking solutions integrated into modern HMDs, research on gaze-contingent rendering is gaining increasing popularity. Although rasterization makes it inherently difficult to sample individual patterns in accordance with acuity models, advances such as deferred rendering and multi-resolution shading are already showing their potential in order to increase rendering performance. *"The question for the future is: How can locally changing rendering and shading quality make the most effective use of perceptual limits to produce photo-realistic scenes with the required flexibility? "* [Wei+17]. The capability to perform efficient low-latency rendering has been demonstrated for both main rendering strategies, rasterization, and ray tracing [FRS19; Fri+16]. Although rasterization methods are currently faster on GPUs, ray tracing does provide a higher degree of flexibility. Hybrid approaches are promising but introduce additional burdens and limitations. Accordingly, ray-based methods could possibly become the first choice for performance critical real-time VR rendering in head-mounted devices [Hun15; Fri+16; HMN18].

In the course of this thesis, different foveated ray tracing systems for HMDs have been developed. The developments presented in Chapter 6 have used reprojection that exploits TC in order to increase the sampling density and to omit expensive image reconstruction techniques. To this end, the reprojection information is combined into a smoothly refined image allowing for TAA, where parts of the image with a high saliency are adaptively resampled. This results in images that are *perceptually equal* to images rendered with full ray tracing but with a significantly reduced number of traced rays. Chapter 7 presents a more "traditional" foveated rendering framework that uses ray tracing and tightly integrates image reconstruction and gaze-contingent DoF in order to filter rendering artifacts in the peripheral visual field.

4.3 TEMPORAL COHERENCE

While traditionally rendering approaches focus on improving the performance of each rendered image, individually, exploiting TC between subsequent frames is a valuable tool to meet perceptual and performance requirements. According to Scherzer et al. [SYM10] typically more than 90% of points on a surface remain visible from one frame to the next. A color-coded visualization of the number of pixels that can be reused between subsequent frames is presented in Figure 49. Reusing this information by exploiting its TC thus provides great potential for optimizing rendering and sampling techniques. Hence, TC methods have been around for over four decades [SSS74]. This chapter gives an overview of methods that exploit TC in order to improve sampling. Here, the survey by Scherzer et al. [SYM10] on TC methods forms an initial basis for this discussion.

(a) Parthenon (b) Coherence (c) Heroine (d) Coherence (e) Ninja (f) Coherence

Figure 49: Tree different scenes with camera (a) and object motion (c, e). The corresponding images (b, d, and f) show the amount of temporal coherence between subsequent frames. Samples that can be reused by exploiting temporal coherence are colored green. Samples that needed re-computation are colored red. *Image from Nehab et al. [NSI06]*

4.3.1 *General Approaches*

Generally, TC approaches can be categorized into image space and world space methods. Image space methods usually rely on a viewport-sized off-screen buffer. This buffer is used to stored information of the previous frame(s) and accessed in the current frame to aid its visual quality. This is referred to as *backward projection* or *reverse re-projection*. The other type of approach projects samples forward in time based on scene motion from the current frame to the buffer that will become the basis for the computation of the next frame. Consequently, this is referred to as *forward re-projection*.

WORLD SPACE CACHING In addition to the general image space approaches, samples can also be stored in world space. One early world space approach is the Holodeck Ray Cache [WS99]. It caches shaded samples, the associated rays and their hitpoints to a file that can be index using ray beams. These ray beams are computed using a raster on the scene's AABB. Finally, new views are constructed by projecting the cached samples to image space. The *render cache* by Walter et al. [WDP99] also stores shaded samples as points in 3D space. This raw point set is projected to a new perspective. Occlusions are resolved using a simple heuristic on small pixel neighborhoods. Due to its simplicity, more efficient implementations for the GPU have been introduced in the following years [TL05; Vel+06]. The Tapestry system [SS00] by Simmons and Séquin uses a dynamic 3D triangle mesh with vertices corresponding to the sample points. This triangle mesh is based on a dynamically updated and incrementally recentered mesh of a unit sphere that is centered at the viewport. Tole et al. [Tol+02] use a mesh in order to cache irradiance values when computing GI. However, using meshes does not eliminate all the artifacts. Geometric edges, for example, have to be reconstructed using a large number of point samples. Besides, these techniques either require dense sampling or higher order representations such as Voronoi regions or a spherical Delaunay mesh in order to reconstruct a full image. This makes them computationally demanding. Irradiance caching, as introduced by Ward et al. [WRC88] or radiance caching, e.g. by Krivanek et al. [Kri+05], attempt to reuse lighting computations in GI rendering processes. These approaches store the radiance or irradiance values in acceleration structures such as octrees. The system by Dietrich et al. [DS07] stores shaded samples in a hash map. A hash for a hit point on a surface is computed using the primitive's ID and a discretized spatial position using the projected pixel footprint. Shaded samples are combined with old samples to obtain the final pixel color.

World space methods have distinct advantages over image space methods. Having a complete 3D representation of the scene at hand allows for storing samples even for those parts of the scene that are occluded in image space. Moreover, samples can be reused over multiple frames. Having a single buffer in image space to store only the previous frame(s) limits a cache entry's live. However, due to the global nature of world space approaches, storage and processing requirements are not negligible. Moreover, dynamic adaptive acceleration structures in R^3 are more challenging to build efficiently in a progressive manner [JSF17].

IMAGE SPACE CACHING As world space sample caching has high storage and processing requirements, it is often more efficient to directly shoot more rays, render at higher resolutions or use more elaborate sampling or post-processing approaches. Hence, more lightweight solutions that reuse samples in image space have become more appealing and are commonly used for video games [Kar14; Lab18] and to accelerate high-quality rendering [Lab18].

The *real-time reprojection cache* by Nehab et al. [NSI06] describes a method to cache and track surface information through time in image space using a single frame buffer, avoiding complex data structures. The frame buffer stores a *running estimate* of shaded color values of the past frames. By storing projection space coordinates for the previous frame as an attribute to each vertex, the current frame's old image space position of each fragment can efficiently be computed using the old and new model-view and projection matrices. The old image space position can be used to obtain a bilinearly filtered reprojected color from the old frame. Storing additional depth information enables to discard the use of old samples in case of (dis-)occlusions. The work by Sitthi-amorn et al. [Sit+08a] further improves this method by developing a three pass system that does reduce branching and performs better on modern GPUs.

While reprojection caching can be used to integrate arbitrary shader computations over time [Sit+08b], for example to improve the quality of soft shadows [SJW07], its most straightforward use is to temporally integrate shaded color samples in order to increase the sampling rate and to increase image stability. Commonly these techniques are known as TAA methods. Yang et al. [Yan+09] take a close look at the running estimate that is used to integrated samples over time, as this computation is prone to over-blurring [SYM10, p. 11]. Some of the ideas by Yang et al. are extended in Chapter 6 in order to increase the sampling density in the peripheral visual field in the proposed foveated rendering system. In addition, some considerations for designing a running estimate are presented in Appendix A.3.

HYBRID APPROACHES A different type of techniques that exploits TC uses image-based proxies, for example billboards, in order to represent complex geometry in a scene by rendering more simple textured polygons. These textures on the polygons are previously computed samples, usually obtained from image space [Sha+96; Xav+01; Déc+03]. Hence, these methods can also be seen as a form of reverse re-projection applied to individual parts of a scene [SYM10, p. 3]. Alternatively, methods also attempt to partition the entire scene in different layers [LS97] or augment the entire frame with depth information rendering *layered-depth images* [Sha+98]. Image space methods using forward projection, such as the one by Qu et al. [Qu+00], often use image warping to transform the output of the current frame to a new view. However, as this process can leave holes due to (dis-)occlusion, the authors use selective ray casting to computed missing information. As presented in the next section, similar tech-

niques have been used to accelerate stereo rendering or to artificially increase frame rates by inter-frame interpolation. Moreover, the gaze-contingent renderer presented in Chapter 6 uses a coarse mesh to warp the old image to a new view. Other methods attempt to discard the concept of frames completely [Bis+94; Day+05; Fri+16]. In these works, pixels are updated progressively and independent from each other. Reprojection and adaptive reconstruction are used to provide better images and more sensible output. Such frameless rendering approaches will become an interesting field of research for novel VR rendering pipelines.

INSIGHTS Exploiting TC in image space has become a central element in many modern interactive rendering pipelines [Kar14; Jim17]. Due to their complexity and the increasing processing power of modern GPUs, world space methods have not seen much attention lately. TC in image space main advantage is their speed and low storage requirements. Hence, they are well suited for performance critical VR and AR pipelines.

4.3.2 *Perception-driven Approaches*

A considerable body of work uses TC methods in perception-driven rendering pipelines. For example, systems that selectively sample the image plane based on perceptual models or direct measurements, e.g. using eye trackers, likely exhibit noise and temporal flickering in different regions of the image. If this is the case, TC methods can help to increase image quality and provide more stable and steady image sequences. Especially peripheral vision is highly sensitive to motion (Section 3.1.7). Hence, a variety of gaze-contingent rendering systems integrate some form of TAA (Section 4.2.3). However, TC methods have been used in numerous ways in order to aid visual perception. Methods also attempt to accelerate multi-view computations or artificially increase the dynamic range and palette as well as the temporal or spatial resolution.

STEREO RENDERING Up to now, the majority of the presented work treats the HVS as a single-eyed system, though healthy humans are capable of stereopsis due to binocular vision. Rendering images for both eyes independently doubles the computational effort. Several techniques have been introduced to convert 2D images to stereo images automatically [Wei05]. These techniques are commonly used for 2D to 3D movie conversion and use, e.g., depth information that is reconstructed from the monocular video in order to produce a stereoscopic version. However, this thesis considers computer generated images from 3D datasets. Here, depth, as well as other scene information, are readily available. Using this information, several approaches have been introduced that synthesize images by projection or gathering depth samples from several views to reconstruct a new view [McM97; Bow10]. Such approaches are especially well suited for synthesizing artificial stereo pairs. Here, the shift between the view from one eye to the other eye is very limited. Still, any change of the viewport likely causes occlusion to change and may lead to missing information in the synthesized view. Parts of the scene that were not visible in one view cannot be properly reconstructed in the synthesized view. Essentially, determining the pixel color for such parts is always an assumption on what is right. Also, shading of an image is view-dependent to the apparent location of light bouncing off the surface of an object. Highlights on highly glossy surfaces or mirror like reflections

change with the view. Reconstruction such surfaces correctly when synthesizing new views from an other view is very challenging.

For ray tracing, early approaches use TC to accelerate the intersection process using information from one eye in the other [AH93; Bad88]. If more object information, such as object movement is included, the validity of samples and thus the performance can be highly increased [AH95]. For accelerating stereo generation with rasterization, Fu et al. [FBP96] use reprojection in combination with a hole-filling algorithm writing to a modified z-buffer. Later, Didyk et al. [Did+10c] introduce an adaptive warping grid to forward-transform one view into another. The adaptive process leads to areas with similar disparity being warped with a coarse grid, where areas of differing disparity are warped using an adaptively tesselated grid. In addition, the method exploits the fact that it can be beneficial (due to possible occlusions) to change which eye to render first and synthesize one eye or the other consecutively. Marbach [Mar09] takes a closer look at how layered rendering can be used to improve the performance of rasterizing stereo views requiring just a single geometry pass.

INCREASING TEMPORAL RESOLUTION Another field that can greatly benefit from exploiting TC is inter-frame interpolation to increase the temporal resolution of video output, for example to counteract *hold-type blur* [Sch+12, p. 19]. This blur is introduced due to the properties of the display. Opposed to CRT technology, LCDs do not flash the image but present static images that stay on the display until the next display refresh. Moving content presented on such screen can exhibit blur as the eye does integrate images on the retina during eye motions. However, this blur is commonly mixed up with motion blur. Pan et al. [PFD05] showed that only 30% of the perceived blur is a consequence motion blur, while 70% are mostly hold-type blur [Sch+12, p. 19]. This effect reduces image quality [Jan01] and task performance [Did+10b]. Modern TVs already optimize image quality by employing interpolation schemes [Sch+12, p. 19]. Didyk et al. [Did+10b] build upon the observation that the HVS spreads high frequencies of one frame over succeeding blurred frames if a sufficiently high frame rate is reached [TV05][Sch+12, p. 19]. Hence, a lower quality but very efficient warping technique can be used to produce intra-frames – effectively a 40 Hz sequence can be transformed into a 120 Hz output that is, according to an accompanying user study, barely distinguishable from a sequence rendered at 120 Hz. Another technique commonly used for video gaming systems was presented by Andreev [And10]. In this approach, the scene is segmented into static and dynamic elements. Static elements are forward projected using a simple warping approach, and dynamic elements are added on top. Holes in the warped output are filled using pixel patches from their neighborhood. Yang et al. [Yan+11] introduced a method that interpolates a pair of consecutively rendered frames. Here information from both frames is used to compute intermediate frames. A popular image space TC technique to hide rendering latencies is Time Warping [MMB97]. The rendered image is shifted and distorted just before display to compensate for orientation changes. However, due to occlusion it only works for rotations and does not help with translations. Also, the original implementation of Time Warping [MMB97] is synchronous. Warping is executed after new frames have been rendered. However, in order to reduce latencies it is best to perform the warping operation in a separate rendering thread. Nowadays, commercial HMDs make use of techniques such as asynchronous time [BG16] and spacewarp [BHP16] to avoid skipped and repeated frames. With asynchronous spacewarping [BHP16], re-projection is combined with a more advanced warping technique in screen-space that takes the depth buffer into account. This way, this

technique offers more freedoms with respect to camera translations. Still, missing information due to (dis-)occlusions remains an issue. Most recently, Schollmeyer et al. [Sch+17b] propose a hybrid approach to image warping for VR systems that uses a rasterized adaptive grid coupled with a hole and dis-occlusion filling algorithm that employs ray tracing in the pixel shader.

INCREASING THE PALETTE Besides increasing the temporal resolution also the palette can be increased. If colors are presented at high framerates the eye can no longer distinguish between individual frames and colors – the final color is mixed. This property is commonly used in DLP projectors and in LCDs in order to increase the palette [Sch+12, p. 24]. Other effects include using adaptation and afterimages. Bright light sources lead to a short-term receptor bleaching - the retinal image is overexposed (Section 3.1.5). This can be used to simulate and alter the perceived brightness and to enhance the visual experience by conveying a higher dynamic range [RE12; Jac+15; Yu+17].

INCREASING SHARPNESS Methods that exploit TC can also be used to enhance the perception of spatial details. An approach by Didyk et al. [Did+10a] gives the impression of higher resolution content on lower resolution displays. As the receptors of the eye integrate the content of an image along its trajectory, a model of the Smooth Pursuit Eye Motion (SPEM) as well as shuffled and adapted version of still images that are displayed at refresh rates exceeding the CFF, enable to convey the impression of an increased resolution. This work has also been extended to animation sequences by assuming that the eye movement relates to the underlying optical flow [Tem+11]. Another system by Berthouzoz et al. [BF12] uses the perceptual resolution increase by vibrating an entire display. Such techniques can be a valuable tool to drive apparent display resolutions towards and beyond physical limits. Because the HVS has a finite integration time, lower resolution frames can also be fused with high resolution frames in order to produce the sensation of a sharp image. Researcher have investigated techniques that render every other frame to lower resolution in order to reduce the number of shaded pixels. Here, it can be exploited that the HVS is insensitive to both high spatial and temporal frequencies. In order to be imperceptible one of the image should be blurred and in the other image, high frequencies should be amplified. However, techniques such as nonlinearity compensated smooth frame insertion [Che+05] may lead to ghosting artifacts or contrast losses. Most recently, Temporal Resolution Multiplexing [Den+19] tries to tackle these issues using a motion-aware filtering scheme.

INSIGHTS Exploiting the temporal and spatial coherence of the views when rendering in stereo is an obvious field where TC methods can demonstrate their benefits. Perception-driven warping techniques are already showing their potential to hide rendering latencies for performance critical VR and AR headsets. TC methods make it possible to artificially increase the perceived temporal frequency, spatial resolution, dynamic range and the colorfulness of the output. It is astonishing how insight of the human perception can be used to drive displays and rendering outputs beyond their physical limits.

4.4 POST-PROCESSING

Post-processing techniques process the final output of a renderer in order to produce richer and more perceptually pleasing images – usually, output at the final display resolution and possibly including additional information such as depth-buffers. However, post-processing techniques do in no way improve sampling. These approaches hide artifacts in a way that they are less disturbing or ideally concealed.

HIDING ALIASING IN TRADITIONAL RENDERING PIPELINES In the previous years, several techniques have been developed that use post-processing to filter the images in order to reduce aliasing artifacts such as jagged edges. *"The basic idea is to find discontinuities on the image and to blur them in clever ways, in order to smooth the jagged edges."* [Jim+12] While the concept is not new [Blo83; Ove92; IK99] these techniques experience a renaissance. The main reason for this is their performance. Usually, these filters process the image in times not more than 2 ms per frame [Jim+12]. In addition, these filters can be readily used for arbitrary rendering paradigms such as deferred [HH04] or tiled rendering [BOA13]. In the last decade, several of such filtering approaches have been introduced [Jim+11; Lot11; Jim+12]. Most of them are based on the initial idea by Reshetov, introducing Morphological Anti-Aliasing (MLAA) [Res09]. A set of morphological operations is used to classify different edges at discontinuities in the image (Figure 50 (a) and (b)). This classification is used to derive a pixel coverage and in order to compute pre-pixel blending weights. As Reshetov's original approach was implemented for a CPU-based ray tracing framework, several GPU-optimized variants have been developed since [BHD10; Jim+11]. Topological Reconstruction Anti-Aliasing (TMLAA) [AD11] improves upon Reshetov by using some topological information to recover subpixel features, for example for thin objects such as wires and fences that draw to potentially disconnected lines when sampled with an insufficiently high sampling rate. In order to detect edges, TMLAA [AD11] switches to the CIE-LAB color space, that is a more "even" concerning perceptual implications (Section 3.1.4).

Fast Approximate Anti-Aliasing (FXAA) [Lot11] is NVIDIAs version of post-processing AA. Edges are detected by determining changes in the local contrasts between neighboring pixels. To this end, for each 2×2 pixel neighborhood, luminance differences are computed. In order to account for the fact that the HVS is not very sensitive for "blueish" wavelengths (Section 2.2.2), the luminance conversion of the RGB input only takes the red and green channel of the RGB input into account. However, such a detection exclusively on local numerical differences will produce spurious edges that decrease efficiency and image quality during filtering. The image is overblurred and, as blurring is a computationally expensive operation, the filter's run time is increased.

For that reason, Subpixel Morphological Anti-Aliasing (SMAA) [Jim+12] extended the ideas of FXAA by incorporating information on simultaneous contrasts and visual masking. To this end, an adaptive double threshold is performed during the edge detection phase (Figure 50 (c) and (d)). SMAA supports various color spaces. However, the authors usually restrict themselves to using differences in the luma channel only. As the HVS mask contrast edges in the presence of edges with higher contrast in the surrounding, additionally filtering the lower contrast edges decreases performance, downgrades image quality, and temporal stability [Jim+12]. Therefore, the maximum contrast of a set of edges in the surrounding

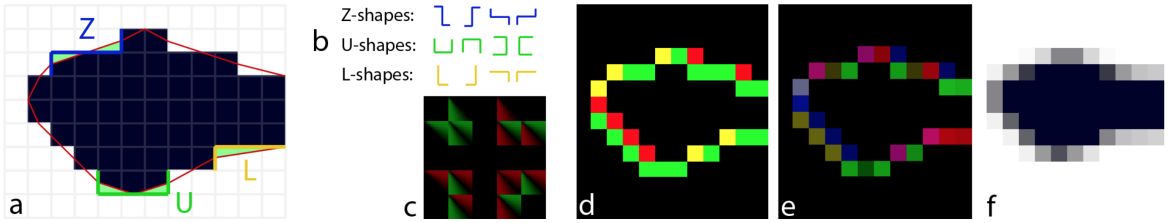


Figure 50: MLAA overview. (a) Input image, with the intended approximation outlined by red lines and the coverage areas shown in green. (b) Predefined patterns in the original algorithm [Res09]. (c) Precomputed areas texture in Jimenez's GPU implementation [Jim+11]. (d) Detected edges. (e) Calculated coverage areas. (f) Final blending. The SMAA algorithm overhauls the whole pipeline by extending (b) and (c) for sharp geometric features and diagonals handling. Local contrast adaptation removes spurious edges in (d). Extended patterns detection and accurate searches improve accuracy in (e). SMAA can handle additional samples in (f) for accurate subpixel features and temporal supersampling. *Image and caption from Jimenez et al. [Jim+12]*

of the detected edge is computed. If the detected edge is above the weighted maximum contrast, it is not further considered. Otherwise, it is not marked for the blurring operation. In general, SMAA provides a high visual quality but FXAA is faster. A more recent technique that improves temporal stability, Conservative Morphological Anti-Aliasing (CMAA) [Str14], positions itself between SMAA and FXAA. Compared to FXAA and low-quality SMAA modes it provides higher image quality and can handle otherwise challenging long edges. Adaptive Approximate Anti-Aliasing (AXAA) improves upon image quality and performance of FXAA by not filtering pixels multiple times, conserving contrast of thin geometry and adaptively setting the search range of edges based on the luma contrast.

Although screen space techniques are widely used in industry [Mit12; Val14] and can improve image quality substantially, they often lack adequate subpixel accuracy and temporal stability. Hence, a variety of the proposed methods specifically target these aspects. Subpixel accuracy can be improved by finding the closest triangle edge per subpixel in a separate render pass [Jim+11; Gör15], by providing G-buffers at a higher resolution [CML11; WJB], or when using MSAA samples for a better gradient and color estimation [IYP09]. Temporal stability can be improved by storing more geometric information per pixel [Per11; Gör15]. All of those approaches are highly computationally demanding and often not better than optimized MSAA variants.

Besides, the introduced screen space techniques fail to antialias bead chains, as they only work for visible edges in image space. *"The fundamental problem [with bead chains] is that the highlight is both very bright and very thin, and any anti-aliasing method based on sampling in image space is likely to miss this feature, even with high supersampling. [Sie+13]"* Adapting these techniques to handle aliasing from shading requires detecting and resolving those artifacts in image space. Detecting such artifacts is a challenge in itself, as they occur in a variety of different patterns and estimating the complexity inside a pixel is computationally challenging and often next to impossible without computing and shading more fragments within a pixel's extent. Removing the resulting artifacts is even more difficult, as too much information has already been lost in image space. This is one reason why rendering pipelines in video games often use a combination of MLAA-based techniques in image space with TAA [Val14], or, if they can afford to do so, MSAA and its successors (Section 4.2.1).

However, depth discontinuities are most prominent regions where reprojection techniques might fail. In addition, accurately representing regions with high local contrasts is crucial for peripheral vision (Section 2.2.2). Hence, an edge detection filter inspired by SMAA in order to detect regions that require resampling is used by the foveated ray tracing framework detailed in Chapter 6.

HIDING ALIASING IN STOCHASTIC RENDERING PIPELINES Other approaches that greatly benefit from post-processing techniques are stochastic GI methods. Here, denoising filters can help in creating more plausible and visually appealing images. As the rendering process is already involved, a large body of work is focusing on offline methods. A comprehensive overview of the state-of-the-art can be found in the survey by Zwicker et al. [Zwi+15]. Such methods can afford to spend time filtering the image. Hence, commonly images are transformed into the frequency or wavelet domain. Also, given high sampling densities, a wide range of statistical analysis is possible. Several methods attempt a regression-based analysis [Bit+16; Moo+15]. Besides, high initial sampling densities and accompanying G-buffers, that separate colors from depth, normals, direct and indirect illumination, allow preserving edges by bilateral filtering [Pet+04]. All this significantly lowers the amount of noise.

In recent years researchers started exploring those techniques to be executed at interactive rendering rates and low sampling densities. For interactive or real-time rendering, it is often not possible to execute complex filters and statistical analysis due to the given limited time budget. However, by separating direct and indirect lighting at the first hit-point and approximations of joint bilateral filtering by À-trous wavelets [Dam+10], Guided Image Filtering [Bau+11] or adaptive manifolds [GO12; BEM15], make efficient implementations possible. Filters can also be combined. Bauszat et at. [Bau+15] create a set of filters from which they select the most appropriate one per pixel and depending on estimated input error and variance. Alternatively, incorporating more geometric information in the edge-stopping function of the filter improves the cross bilateral filter's robustness under input noise. This is critical when these filters are to be used at low-sampling densities. Also, methods profit greatly from exploiting TC, for example the work by Schied et al. [Sch+17a]. With this, the denoising itself is performed by using a hierarchical, image space wavelet filter. An overview of interactive filtering techniques for GI rendering can be found in the book by Schwenk [Sch13], the report by Zwicker et al. [Zwi+15] and the paper by Schied et al. [Sch+17a].

Besides these approaches, currently, the immense impact of machine learning methods can be observed. Techniques such as Deep Learning Super Sampling (DLSS) [Bur18] attempt to hide aliasing artifacts in image space with fast convolution neural networks. Also, there is already a wide range of machine learning techniques that enables to upscale images at a higher-quality compared to conventional approaches [Don+14; Yan+18; Str18]. In order to filter GI, Kalantari et al. [KBS15] propose applying a small neural network to control a cross bilateral filter's per-pixel feature weights. Chaitanya et al. [Cha+17] present a technique for reconstructing image sequences based on machine learning with deep convolutional networks.

DEPTH-OF-FIELD Other techniques that are more in the scope of this thesis, filter images by exploiting the optical properties of the HVS, most notable the eye's DoF. Besides filtering GI the adaptive manifolds introduced by Bauszat et al. [BEM15] allow for an efficient approximation of the DoF effect. The work by Lindeberg [Lin16] uses DoF in order to filter

artifacts resulting from gaze-contingent polygonal simplifications (Section 4.1.3). Chapter 7 details a method to compute DoF in screen space in order to hide undersampling artifacts in a foveated rendering framework. More information on general rendering methods for DoF can be found in work by Barsky et al. [BK08]. The chapter by Demers [Dem04] and the work by McIntosh et al. [MRD12] place its focus on post-processing techniques that compute DoF in image space.

INSIGHTS Post-processing techniques have tremendous potentials for efficient rendering as they move much complexity from the world and sampling space in R^n to image space in R^2. However, post-processing techniques only conceal artifacts introduced by insufficient sampling and temporal inconsistencies are an issue. That's why exploiting TC has had a great impact on filtering, for example GI noise [Sch+17a]. Furthermore, for post-processing techniques, machine learning might be able to show its full potentials. Deep learning techniques already provide remarkable results for filtering and upscaling images. The learned models could become a relevant design criteria for deciding sampling and filtering parameters. Further research in this direction will therefore be carried out. As with more traditional approaches such as bilateral filters, temporal inconsistencies might be one major challenge when designing learners. Also, techniques have not yet exploited perceptual implications to its fullest potentials. Here, processes such as visual masking, contrast and color sensitivity as well as the limited spatial acuity might be helpful when designing novel systems. To this end, ideally, the visual masking potential and response to inputs at different colors as well as contrasts are already integrated inside the learned model.

4.5 CONCLUSION

This chapter has provided an in-depth overview of the wide field of efficient perception-driven rendering techniques as well as general concepts, methods, and trends. Here, this thesis has presented how insights of human perception and the underlying perceptual models can be a valuable tool to enable more efficient rendering algorithms. While recently visual acuity models have proven to be successful in rasterization, a shift of focus can be observed to methods that adapt sampling in real-time ray tracing systems. Due to the increase in geometric processing power of modern GPUs and the fact that shading costs often dominate in modern rendering pipelines, selective ray tracing approaches are on the brink of becoming the first choice in interactive rendering as well. Currently, it can be observed how dedicated ray tracing hardware is becoming available to the masses. Most relevant here is the company NVIDIA that implement hardware acceleration for ray tracing in their latest GPU generations [Sti18]. Having said that, as sampling and rendering is a limiting factor in modern pipelines, it is vital to spend computational power wisely. Increased computational effort is required in order to produce various visual effects ranging from highly-complex direct lighting and shading models to full GI solutions to further improve image fidelity. When rendering for a human observer, it becomes apparent that state-of-the-art algorithms allocate most of the computational effort where it matters most – namely in those regions that are critical to perception. As such, perceptual models enable ray tracing processes to be steered into regions that need more samples, resulting in faster convergence to photorealistic images.

Moreover, it is highly beneficial to consider properties of the HVS in systems that require low-latency rendering in order to reduce nausea. This is critical for immersive VR, believable AR, and mixed-reality applications. Here, rendering quality should ideally be adapted to both the computational power of the rendering system as well as the capabilities of the human perception, thus guaranteeing lower bounds for the refresh rates. In direct comparison, methods that enable gaze-contingent adaptive sampling by using active measurements still function considerably more accurate than their solely model-based counterparts. Likewise, exploiting TC and frameless rendering enables display refreshes, pixel color generation and their transmission to be decoupled. Along with the shift towards methods that support interactive ray tracing, machine learning techniques are already demonstrating some of their potential. Here, novel approaches can be expected that model parts or the entire visual pipeline in order to guide sampling processes, improve reconstruction and provide post-processing filters at a higher quality.

Part III

METHODS AND METHODOLOGIES

In the next chapters, this thesis introduces original contributions to perception-driven rendering pipelines for VR applications. While Chapter 5 presents an approach for large high-resolution display walls that adapts the model quality based on the tracking the user's gaze, Chapter 6 and Chapter 7 detail the author's contributions to the field of gaze-contingent rendering for HMDs with eye tracking.

5 HYBRID SPARSE VOXEL OCTREES

Exploiting Field-of-View and Acuity Limits

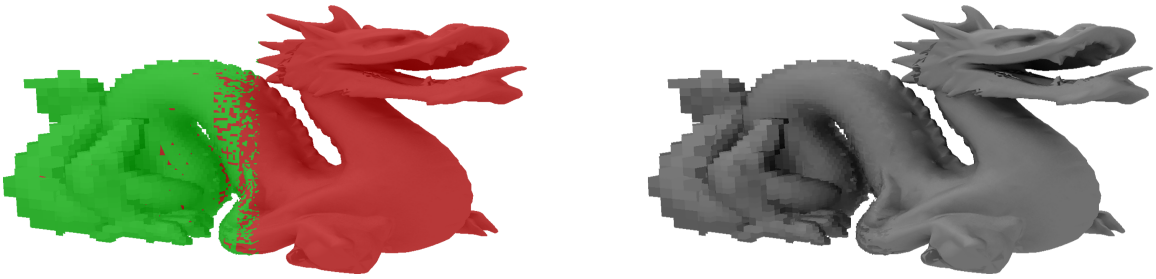


Figure 51: The presented data structure combines voxel data (left, green) and polygonal data (left, red) to synthesize view-adaptive images (right).

As discussed in the previous chapter, one strategy that aims to accelerate rendering and reduce aliasing artifacts is to simplify the original function by pre-filtering in order to provide a version that can be sampled more accurately and efficiently. Hence, for 3D models, various strategies have been developed that enable the reduction of a model's geometric complexity, thus providing representations at various Level-of-Details (LoDs) that are less prone to aliasing artifacts. As we have seen in Section 4.1, a large body of work focuses on geometric simplifications by adapting triangulated meshes. While this works well for most models, primarily when they are described by a regularly tessellated and topologically simple mesh, such methods can fail with high-frequency and irregularly tessellated models. In the latter case, polygonal simplification is error-prone due to the complex structure of the model. Other more tailored solutions take object knowledge into account to provide a higher visual fidelity at a reduced computational complexity. Therefore, specific approaches for special types of models like plants and trees [Deu+02; ZBJ06; Beh+05], buildings [VLA15; PSS16; Bil17] or crowds of humans [Dob+05; MR06; Pen+11] have been developed.

Although specialized solutions provide a high visual fidelity, these cannot be applied to arbitrary scenes and objects. It is difficult to simplify textured models with different surface properties, e.g. no meaningful average can be calculated by combining different materials. Some of those issues are discussed in Section 5.6. Likewise, maintaining code, preprocessing, or modeling of new assets at different LoDs as well as launching rendering stages can be challenging. In order to mitigate some of the problems and to provide a general solution for the simplification of 3D models that is independent of the type of object or scene, *the central idea of the presented approach is to combine a volumetric representation using voxels with a polygonal representation* (Figure 51). This research project has opted to use voxels as volumetric descriptions are less sensitive to a scene's complexity and enable progressive refinement. Once the mesh is voxelized, a coarse representation of the scene can be constructed using a weighted mean of all voxels within a specific space. Volumetric descriptions can then be stored in an octree where consecutive inner levels in the tree describe coarser scene representations.

These representations, so-called Sparse Voxel Octrees (SVOs) [LK11; Cra+09], are well-suited to provide LoD for high-frequency models. However, if these SVOs should provide a visual quality that compares to a polygonal description, a high resolution and thus a considerable amount of memory is required. When arbitrary scenes are voxelized, many voxels need to be created for single triangles, possibly oversampling the geometric domain even though the polygonal representation is more compact and provides higher visual fidelity. For this reason, we set out to explore the potentials when combining both into a hybrid acceleration structure.

In the following we introduce the Hybrid Sparse Voxel Octree (HSVO) that extends upon traditional voxel-only SVOs by augmenting them with triangle references in the leaf nodes. Having voxel and polygonal data in one acceleration structure is an advantage because it minimizes management and storage cost compared to having two separate structures. In addition, having triangle information in the leaf nodes can reduce the size of the octree. The construction of the HSVO can stop for nodes that contain a maximum of n_{split} triangles. Here, often $n_{split} = 2$ is assumed, as two triangles are cheaper to intersect compared to traversing the structure deeper. Also, this is common for non-isolated triangles, i.e. the ones sharing an edge. Assuming that non-isolated triangles form a solid surface inside a voxel's space, they are far less crucial to geometric aliasing problems. Hence, the system can reduce voxelization fidelity in favor of visualization quality or storage requirements and vice versa. Another benefit of the hybrid octree structure is that it enables a convenient, smooth intra-level interpolation, providing a way to blend between layers in the hierarchy. Nonetheless and indeed most notable, it enables faster image generation if parts of the scene exist for which a coarse representation is sufficient.

In order to achieve high performance and to support arbitrary meshes, voxelization and construction of the structure are entirely performed on the Graphics Processing Unit (GPU) using a specialized OpenGL Shading Language (GLSL) shader pipeline. Initially, the octree was intended to reduce temporal aliasing artifacts which are perceptually-disturbing when rendering large outdoor scenes. In order to generate and render these scenes, we make use of our multi-level instantiation framework presented in our prior work [Wei+13]. At the time, is was also necessary to render these and similar highly complex scenes onto a large tiled-display wall. For such systems, the hybrid structure allows the visual quality to be adapted according to the user's visual field and view-direction, thus accelerating rendering by exploiting the user's visual acuity and limited Field of View (FoV). Here, the HSVO allows the geometric complexity for parts that are in the visual peripheral and outside the user's FoV to be reduced. In line with the publications [Wei+14a; Wei+14b; WHS15] this chapter introduces the HSVO structure. However, in comparison to the aforementioned publications, the thesis contains an entirely new evaluation of the method's visual quality, a greatly revised evaluation of the accompanying user study, and a discussion of its potentials in future works. In summary, the following contributions are presented:

- A GPU-based voxelization and octree construction scheme to build a hybrid acceleration structure combining voxel and polygonal information.
- A ray tracing framework based on OpenCL with a LoD selection scheme utilizing the hybrid acceleration structure.
- A validation by performance and quality benchmarks.

- A view direction-based LoD selection scheme for perception-driven rendering accompanied by a user study for perceptual evaluation.

CONTRIBUTIONS BY THE AUTHOR This chapter is based on work published in:

Martin Weier, André Hinkenjann, and Philipp Slusallek. *"A Unified Triangle/Voxel Structure for GPUs and its Applications."* In: *Journal of WSCG*. WSCG 24.No. 1-2 (2015), pp. 83–90. ISSN: 2464-4617.

Martin Weier, Jens Maiero, Thorsten Roth, André Hinkenjann, and Philipp Slusallek. *"Enhancing Rendering Performance with View-Direction-Based Rendering Techniques for Large, High Resolution Multi-Display Systems."* In: *11. Workshop Virtuelle Realität und Augmented Reality der GI-Fachgruppe VR/AR*. Sept. 2014.

Martin Weier, Jens Maiero, Thorsten Roth, André Hinkenjann, and Philipp Slusallek. *Lazy Details for Large High-Resolution Displays*. SIGGRAPH Asia. 2014. Poster.

Martin Weier, André Hinkenjann, Georg Demme, and Philipp Slusallek. *"Generating and Rendering Large Scale Tiled Plant Populations."* In: *JVRB - Journal of Virtual Reality and Broadcasting* 10.1 (2013).

I was the primary investigator for all publications, developed the GPU-based voxelization and octree construction approach as well as the OpenCL-based ray tracing framework. In addition, I developed the view direction-based LoD selection scheme. Contributions by my co-author Thorsten Roth can be found in the virtual ray casting module to determine the user's visible field and the post-processing approach to smooth LoD transitions. Details are provided in Section 5.4.1 and Section 5.4.3. I designed, executed, and evaluated the system with benchmarks, visual quality metrics, and a user study. Here, this thesis contains a new evaluation of the implications for the visual quality in Section 5.3 and a discussion on the method's applicability as well as its limitations in Section 5.6.

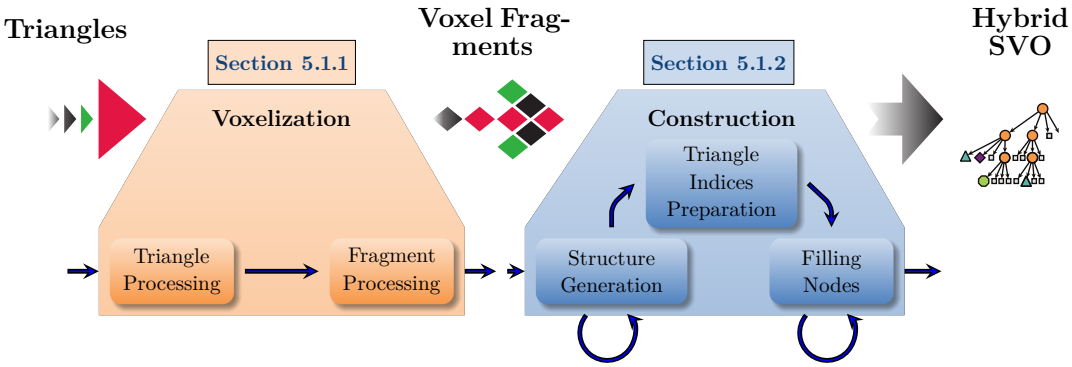


Figure 52: Overview of the GPU-based construction pipeline for the hybrid acceleration structure. A list of triangles is voxelized using OpenGL. Each fragment generated in the fragment shader is extended by a triangle index. After that, shaders are launched that process the voxel fragments in order to generate the octree structure, prepare the triangle indices, and to fill the inner nodes.

5.1 METHOD

An HSVO augments the multi-level voxel information by triangle references that are stored in the leaf nodes. The construction pipeline is illustrated in Figure 52. First, the 3D scene, i.e., a list of triangles, is transformed into a list of unsorted voxel fragments. This is achieved by using the GPU's hardware rasterizer. Essentially, this step performs a 3D rasterization that stores all fragments within the view frustum, their 3D spatial position, and attributes such as colors and normals into a list. This process is detailed in Section 5.1.1. Based on the unsorted list of voxel fragments, different compute shaders are launched, building the structure in top-down and bottom-up processes. An overview of the memory layout of the resulting HSVO and the construction process is presented in Section 5.1.2. Eventually, a ray tracing approach is used to render images. Rendering is performed using traversal routines implemented in OpenCL. A blending process between different levels of the tree can be employed in order to smooth transitions. The traversal methods are presented in Section 5.1.3.

5.1.1 *Voxelization*

Voxelization describes the process of turning a 3D (polygonal) model into a voxel dataset, i.e., into a representation of the model as cells in a 3D grid. This process can be done efficiently on the GPUs using hardware-accelerated rasterization pipelines, such as OpenGL [ED08; CG12].

For constructing the HSVO, voxelization is performed in a similar manner to that presented by Crassin and Green [CG12]. The aim of this process is to rasterize each triangle of the model into the 3D grid in a single rendering pass. Voxelization starts by setting a view port's resolution that matches the voxelized model's target voxel resolution, e.g., 512^2 for a 512^3 voxel resolution. Likewise, the view frustum is set up to match the largest extent of the scene's bounding box to cover all triangles of the model. Now, after disabling depth writes and backface culling, a GLSL pipeline with a custom geometry and fragment shader is launched.

MORTON CODE 8B
RGBA 4B
NORMAL 12B
TRIANGLE INDEX 4B

Table 3: Structure of a voxel fragment entry, including each element's memory size in byte.

However, without any specific hardware support, it is only possible to render a 3D model from a single viewport in a single rendering pass. Triangles that are mostly perpendicular to the main viewing direction are potentially only covered by a few pixels or are not visible at all. Traditionally, rasterization-based voxelization systems have rendered the scene from all three sides to maximize the projected surface of each triangle with respect to the view plane [Don+04; FC00]. However, following the approach by Crassin and Green, the geometry shader is used to project each triangle according to its dominant axis, i.e., where its projected visible surface is maximal. Each triangle is rasterized as it has been viewed from the best possible viewing direction, with respect to the scene's axis-aligned bounding box. This projection is later undone in the fragment shader. This allows all triangles to be processed in a single rendering pass independent of their orientation.

Additionally, to make sure each non-empty voxel is generated, *conservative rasterization* is performed [HAO05]. Conservative rasterization extends each primitive slightly to trigger a call in the fragment shader for all fragments emerging from all triangles touching a voxel's cell. This is necessary since OpenGL samples each pixel during the rasterization only at the pixel's center. The triangles are extended slightly to ensure that each triangle intersecting a pixel covers the pixel's center (Figure 52 *Triangle Processing*). This way, each triangle within the view frustum will create a set of fragments accessible in the fragment shader.

Eventually, using the GLSL and atomic counters, each fragment is written to a chunk of linear video memory (Figure 52 *Fragment Processing*). Each of these voxel fragments stores its discretized position encoded in a Morton code. Morton codes make it possible to perform a fast per-fragment traversal using bit shifts and a rapid comparison with other fragments. Also, the voxel fragments store a color, a normal, and a triangle index. This index is determined per fragment by using the built-in variable `gl_PrimitiveID`. Table 3 gives an overview of the memory layout of the voxel fragments.

5.1.2 *Construction*

The voxelization process results in a list of voxel fragments that represent an intersection of a triangle with a specific voxel of the discretized space. Each voxel fragment stores the voxel's position as a Morton code, a color value, a normal, and the index of the intersecting primitive in the list of triangles. At this point the HSVO structure is constructed. Figure 53 shows the construction pipeline on the GPU. After discussing the memory layout of the structure, the three distinct steps of its construction and the employed shaders are described.

MEMORY LAYOUT OF THE HYBRID SPARSE VOXEL OCTREE The octree itself consists of inner nodes (Figure 54 orange) and empty leaf nodes (Figure 54 light grey). A leaf node can hold the reference to a single triangle along with the voxel information (Figure 54 blue). The configurable parameter n_{split}, determines an upper limit of triangle indices per node for the construction. For nodes that represent up to n_{split} triangles (Figure 54 purple), an indirection to a `triangle index array` is needed to store the respective triangle indices.

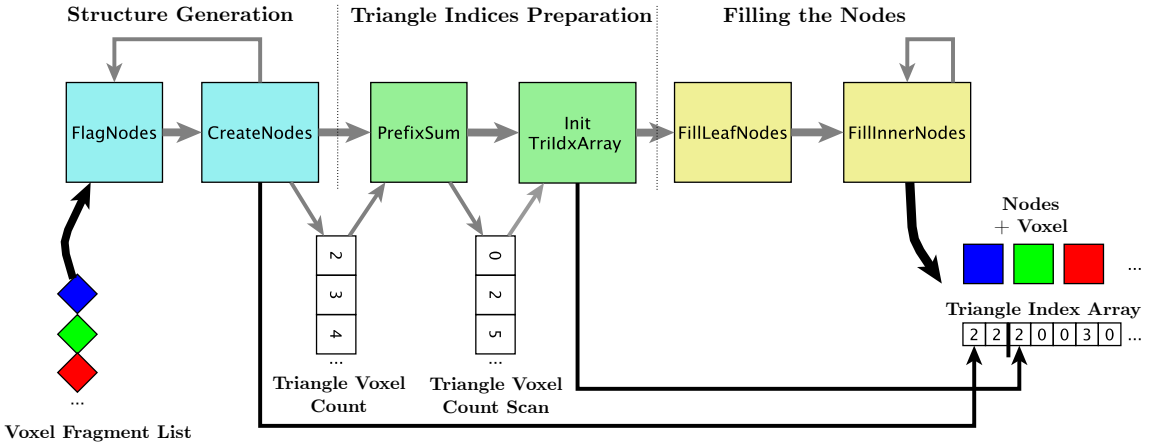


Figure 53: GPU construction pipeline showing the different shader stages (light blue, light green, and light yellow) and their processed and created data.

However, for both node types that hold a single or up to n_{split} triangles, the construction for deeper levels for these nodes can be stopped prematurely. If it contains more than n_{split} triangles, the node must be split, storing deeper levels. At the highest voxelization resolution, nodes can contain more than n_{split} triangles (Figure 54 green). Finer spatial information for further spatial subdivision is not readily available. Hence, these nodes also store all their corresponding triangle indices in the separate `triangle index array`.

Each node of the data structure is encoded in two 32 bit fields (Figure 55). One bit is used to encode whether or not the node is a leaf; another bit is used to mark a node during construction if it needs to be split further. The next 30 bits either encode the index of the first child node, the index of the triangle if it is the only one represented in the voxel, or the index into the triangle index array. The other 32 bits (payload) hold a reference to a voxel array storing the voxel's color, its normal and possibly user-defined fields, such as material parameters (Figure 56).

GENERATING THE TREE'S STRUCTURE After the voxelization process, the construction starts by processing the list of voxel fragments. The aim of this step is to build the general structure of the octree based on these voxel fragment positions. To this end, the tree is traversed repeatedly, top-down, and in parallel, for each fragment in the list of voxel fragments. This way, the tree is built level-by-level. The deepest nodes of the current level are flagged to be split. A second shader creates new nodes based on this information. Initially, this construction starts by adding a single leaf root node to the node list and by calling the `FlagNodes` shader.

`FlagNodes Shader:` The `FlagNodes` shader is launched for each level of the octree. Here, it processes all fragments in the list of voxel fragments from the voxelization process. Each GPU thread traverses the tree according to the voxel fragment's Morton code. Initially, this traversal will reach the root node. However, as construction continues and more nodes are added, deeper levels of the tree are constructed.

Once a leaf node is reached, the algorithm checks whether this node has to be split further. Ultimately, the goal is to determine if a node contains none, a single, n_{split}, or more trian-

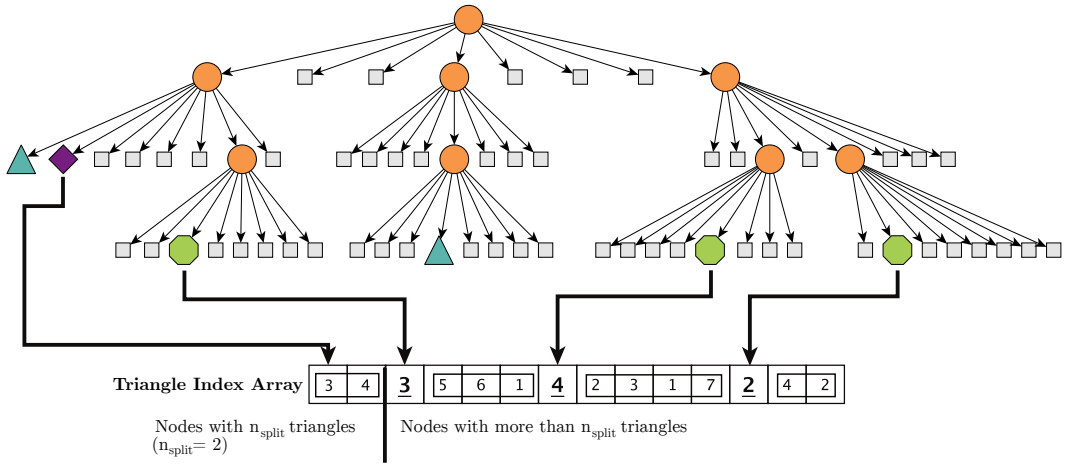


Figure 54: Overview of the hybrid data structure storing triangles and voxels. Inner nodes (orange circles), empty nodes (grey squares), leaf nodes containing a single triangle (light blue triangles), leaf nodes containing up to n_{split} triangles (purple diamonds, here n_{split} is assumed to be two), and leaf nodes containing more than n_{split} triangles (green octagons).

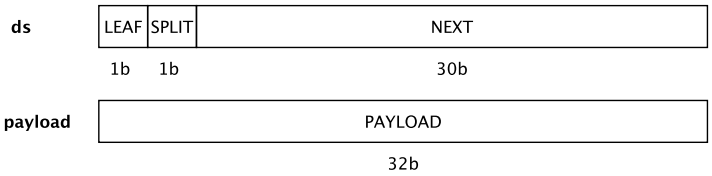


Figure 55: Structure of a single node in the octree using two 32 bit types. While the **ds** (Datastructure) part stores the octree topology, the payload field links each node to a color, normal and triangle indices.

Figure 56: Structure of a single voxel referenced by a node's `payload` field.

gles. This is achieved by caching triangle indices in the nodes using atomic compare-swap operations. Nodes that represent a single triangle can directly cache the triangle index. If $n_{split} = 2$, both triangle indices can also be cached directly in the node's 64 bit of memory. If larger values for n_{split} are desired, the caching must be performed in a separate storage location. However, this additional indirection increases the computation and storage requirements. Hence, it is beneficial to parameterize $n_{split} = 2$. If a node contains more then n_{split} triangles, it needs to be split further and the tree is constructed deeper. This splitting is achieved by setting the node's split bit (Figure 55). After the `FlagNodes` shader has been executed, the `CreateNodes` shader is launched.

`CreateNodes` Shader: The `CreateNodes` shader is then executed for each node of the currently highest *level* in the octree. Each thread checks whether its node at this level has been marked to be split, and, if it has been marked, the leaf node bit is unset and a new empty voxel is created. After this, eight new empty child nodes are created, and the position of the first child node is stored in the **`next`** field of the former leaf. Memory slots are exclusively reserved by each GPU thread in each array for the nodes, namely the voxels and the triangle indices, using atomic-add operations on special atomic datatypes. This ensures that each thread can exclusively write to a distinct array location. However, construction can be stopped prematurely for those nodes that contain less than n_{split} triangles. If a node's split bit has

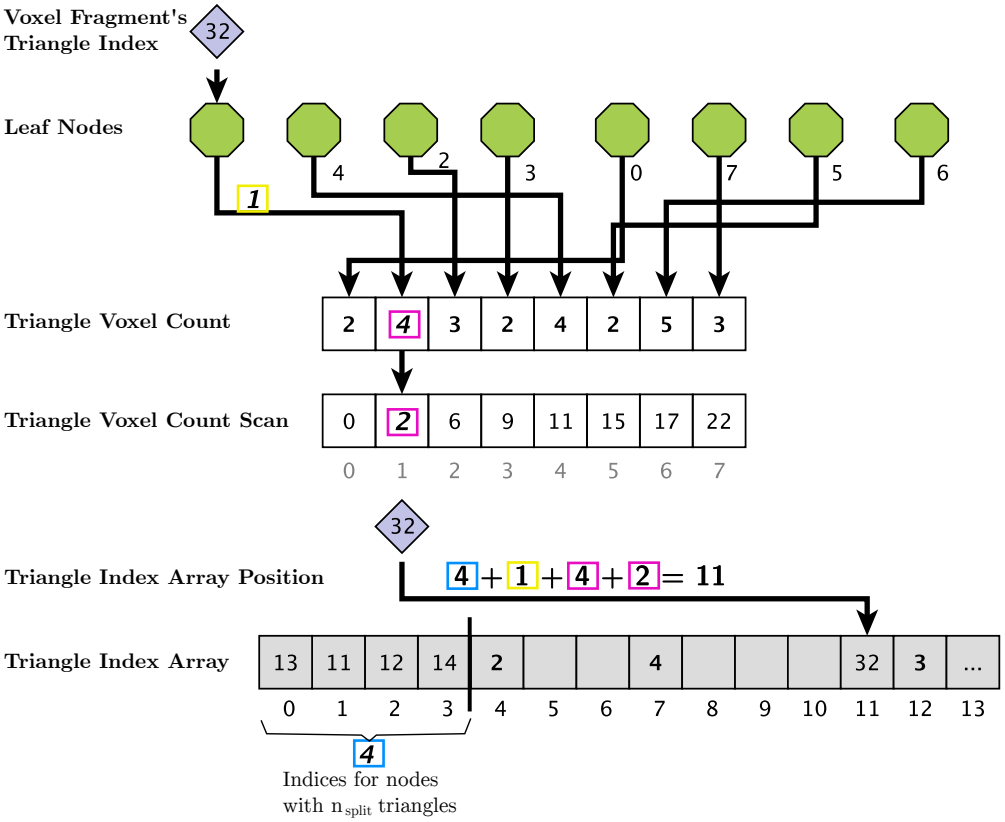


Figure 57: Computation of a fragment's `triangle index` array position. The position of a fragment's triangle index is computed by summing up the number of nodes with two triangles (blue, 4), the index stored in the leaf node's next field (yellow, 1), and the respective array entries in the `triangle voxel count` array and `triangle voxel count scan` array (purple, 4 and 2). Afterward, the value in the `triangle voxel count` array is decremented to point to a new, empty position.

not been set and the construction has not reached the octree's deepest level, the node either contains only one, or represents up tp n_{split} triangles, or is empty.

After the `CreateNodes`, the process starts again by calling the `FlagNodes` shader. This is repeated until the tree's deepest level has been reached. This level represents the scene's highest voxelized resolution. Finally, the `FlagNodes` shader is executed once more, and the tree is traversed the last time. This time the nodes are not marked for splitting as now each fragment represents a single distinct triangle in that voxel. Also, the number of fragments are counted in each leaf node's `payload` field using an atomic add and are stored to a temporary buffer denoted as `triangle voxel count`. This buffer is needed to sort in triangles indices in the further computation steps. The location in the `triangle voxel count` buffer is stored in the `next` field of the node, while a position of a new empty leaf voxel is stored in the `payload` field.

PREPARING TRIANGLE INDICES The last two shader stages build the octree's general structure. However, further computations are needed to link the leaf nodes to references to triangle indices. To this end, the prefix sum of the previously computed `triangle voxel count` array is computed. The results of the scan operation are stored in another temporary

buffer denoted as `triangle voxel count scan`. At this point, another shader is executed for each entry in the temporary `triangle voxel count` array. Each thread takes the value from the `triangle voxel count` array and writes it to the correct location in the `triangle index array`. This copy operation is needed as in this way each leaf node that represents more than n_{split} triangles links to a position in the `triangle voxel count` array that indicates how many consecutive entries are triangle indices and belong to that node (Figure 54). The "correct" location to store this value in the `triangle index array` is the sum of the thread's value from the prefix sum of `triangle voxel count scan` array, the thread's index as well as the number of entries already stored in the `triangle index array` coming from inner nodes that represented n_{split} triangles. These index computations are detailed in Figure 57.

FILLING THE NODES After the tree has been constructed and the `triangle index array` has been prepared, actual data needs to be filled in. In the first step, the `triangle index array` and `triangle voxel count` now make it possible to store all triangle indices of voxel fragments in the leaf nodes. Moreover, information such as voxel colors and normals can be obtained from the initial voxel fragment list and are also written to leaf nodes. Eventually, the latter values need to be combined to fill all inner nodes of the tree with valid normals and colors. This is carried out in a bottom-up process. Both processes are described in the following paragraphs.

`FillLeafNodes Shader:` Having constructed the general structure of the tree, the algorithm continues by filling in the voxel colors, normals, and triangle indices for each node of the tree. The process starts by filling the leaf nodes and by writing each triangle indices to the correct location in the `triangle index` array. The `FillLeafNodes` shader is executed for each voxel fragment in the voxel fragments list. Each thread traverses the tree according to the position of the Morton code of its fragment. Once a leaf node is reached, the fragment's color and normal are averaged using atomic compare-and-swap operations, in a similar fashion as described by Crassin and Green [CG12]. The final step within the `FillLeafNodes` shader is to complete the `triangle index` array. The correct position of a triangle index in that array is computed by taking the sum of the index stored in the leaf nodes' `next` field, together with the value of the `triangle voxel count` array at the index position as well as the number of entries already stored in the `triangle index` array coming from the inner nodes that represented n_{split} triangles. Also, a value from the already computed prefix sum stored in the `triangle voxel count scan` array is added at the index position. Again, Figure 57 clarifies the index computation. The entry in the array of `triangle voxel count` computed by the scan operation is decremented using an atomic add with -1. This ensures that a second thread writes the next triangle index to the next free position in the `triangle index` array. When all threads are finished, the `triangle voxel count` array contains only zeros and the `triangle index` array is filled.

`FillInnerNodes Shader:` The final step of the algorithm is to fill all inner nodes of the HSVO. This last shader is executed multiple times for all nodes level-by-level from the bottom to the top. Each thread averages all colors and normals from its child nodes. Again, this is performed in a similar way to that described by Cyril and Green [CG12]. Each thread checks whether it can write its new summed and averaged value into the voxel's color field by using an atomic compare-and-swap operation in a loop. This loop continues until the new value has been successfully written to memory. For the normals, a simple atomic add on the float

components is used. If normals sum up to a normal with zero length, for example for two opposing faces, the last valid normal is stored. Finally, after the execution has reached the root node, all inner values are filled and the construction is completed. Now, the HSVO can be used for rendering.

5.1.3 *Traversal and Blending*

A custom OpenCL renderer is used to render the data stored in the HSVO. To this end, after the construction, each OpenGL buffer is mapped to OpenCL. These are the buffers containing the nodes, the voxels, the `triangle index` array, and all triangle data as well as the material information of the model.

TRAVERSAL Similar to very early systems, each iteration of the traversal limits the parametric t-span ($[t_{min}, t_{max}]$) of a ray by intersecting it with the three planes subdividing the voxel [AGL89]. With each step down in the tree, these values can be updated iteratively in an efficient manner. The traversal of the voxel structure is implemented using a small stack on the GPU, similar to the work by Laine and Karras [LK11]. Initially, the active parametric t-span of each ray that hits the scene's bounding box is set to cover the extent of this box. The algorithm has three phases:

1. If the current first hit voxel within the active t-span is not empty, the tree is traversed deeper, and the parent node with the current t_{max} is pushed onto a stack. t_{max} is set to point to the end of the active voxel.
2. If the voxel is empty, either the next sibling node of the active parent is processed by setting t_{min} to the beginning of the next node within the t-span, or
3. if the node is not a sibling node of the active parent, nodes are popped from the stack. In the latter case, t_{max} is reset to the position stored on the stack until a hit with the first possible neighboring voxel occurs. From here the process can continue by traversing the tree deeper again.

If the traversal reaches a leaf, its triangles can be intersected – either one, n_{split} or more. Therefore, the algorithm looks at the index stored in the leaf's `next` field. Since the index is encoded using offsets, it can be decided directly if the node references a single, n_{split} or more triangles. The traversal code now determines the closest hit point of the ray and all triangles within that leaf node. Note that now the closest hitpoint becomes the closest hit triangle intersection *within* the boundaries described by the leaf node. Otherwise, the traversal must be continued with the next sibling node.

INTER-LEVEL BLENDING For the LoD selection and to enable a smoother blending between different levels of the HSVO, Ray Differentials [Ige99] are used. Each ray is represented by its origin and a unit vector describing its direction. Besides, the rays store differentials describing pixel offsets on the image plane in x- and y-direction. By using ray differentials, the estimated pixel's footprint in world space on the voxels can be computed by appropriately scaling the differential at each rays' hit point. This footprint can be compared with the size of an individual voxel at level l. If the pixel's footprint is roughly equal to or smaller than the

voxel, traversal can be stopped early as the current voxel resolution is sufficient for the pixel. In addition, a value describing the underestimation $\phi(l, f)$ of the size of the pixel's footprint and the actual size of nodes at level l and $l-1$ can be computed as

$$\phi(l, f) = \frac{2 \cdot v_w(l) - f}{v_w(l)}$$

with $v_w(l)$ being the length of a side of a voxel in world space and f being the estimated length of the pixel's footprint at the ray's hit point. This value ϕ can be used as an interpolation factor between the two subsequent levels in the SVO. Since the tree is traversed using a small stack, the system keeps track of the voxel at level $l-1$ directly and use the interpolation factor during shading and lighting computations for a smooth blending between subsequent levels.

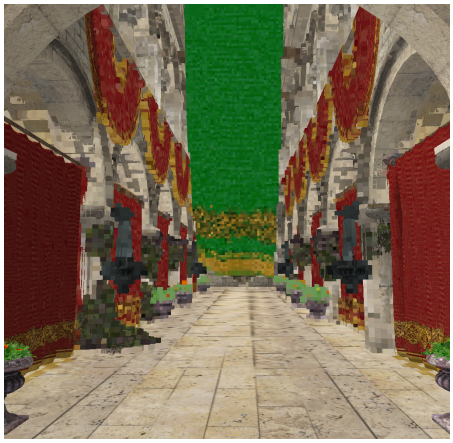
5.2 BENCHMARKS

This section presents the benchmarks for construction and rendering using HSVOs. Benchmarks were performed using an Nvidia GeForce GTX Titan with 6 GiB VRAM on an Intel Core i7 system with 16 GiB RAM. In Figure 58 renderings of the scenes that were used for measurements are shown. Figure 59 presents the construction times of four different test scenes.

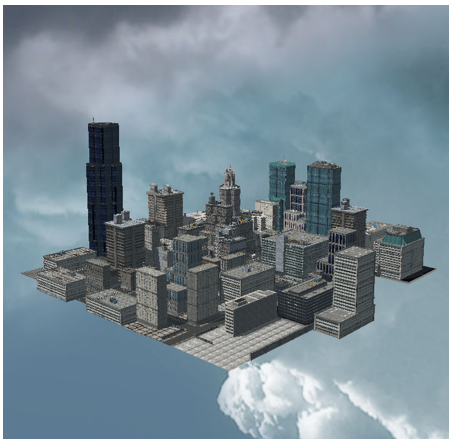
As expected, a rise in the number of triangles increases the run time of the construction. However, the pure triangle count is not the only parameter when it comes to measuring construction times. Highly detailed textures and shaders may further extend the required time to voxelize the model. However, construction is an interactive process (< 20 ms) even for large scenes.

A few other approaches combine voxel- and point-based models with polygonal data – one is FarVoxel [GM05]. There, a voxel-based approximation of the scene is generated using a visibility-aware ray-based sampling of the scene represented by a BSP tree. FarVoxels can be used for out-of-core rendering of very large but static models only – the construction of the tree remains an offline process. Table 5 shows the advantage of the presented method in comparison to a full build of the octree without prematurely stopping the construction for nodes that contain not more than $n_{split} = 2$ triangles. The first impression is that memory savings are not noteworthy. However, the presented numbers also include the size required to store the triangles themselves which in turn largely depends on the scene. The triangle count in the *Sponza* scene is very low. This results in significant memory savings. If only the size of the nodes and the voxel data is considered, the overall saved space amounts to a higher percentage for most scenes. The *Happy Buddha* scene has many very small triangles. For this scene, construction cannot be stopped for most inner nodes which results in only a small memory saving.

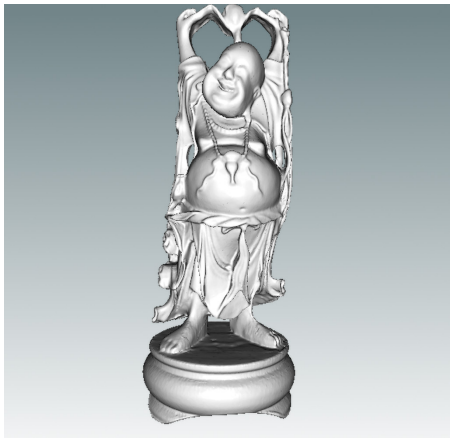
In order to benchmark the rendering performance of the HSVOs, all test scenes have been rendered with a resolution of 1024×1024 using a typical fly-through for about 700 frames. After that, the run times have been averaged. The results in Table 4 list the rendering times from the OpenCL renderer shooting primary rays with Phong lighting, a single point light source, and no texture filtering. Rendering only voxels is fast but lacks visual quality. The traversal of the hybrid structure displaying triangles only provides the highest visual quality



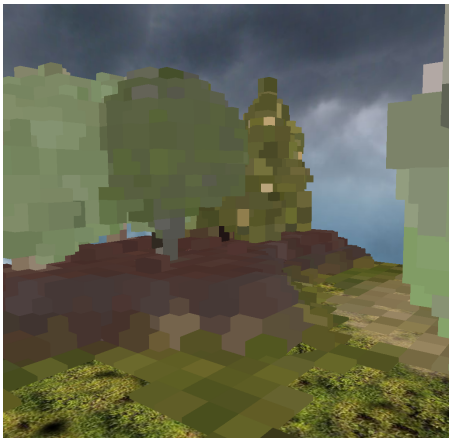
(a) Sponza



(b) Urban Sprawl



(c) Happy Buddha



(d) Forest

Figure 58: Demo scenes at various LoDs used for benchmarking the presented hybrid acceleration structure.

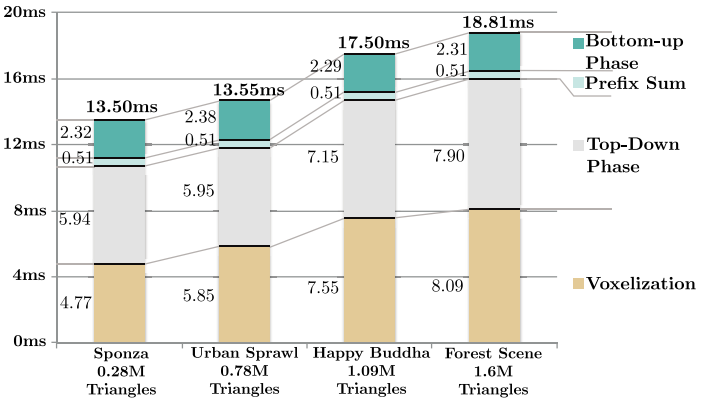


Figure 59: Runtimes for each phase of the construction as well as the overall construction time. Each scene was voxelized with a resolution of 512^3.

Scene	Voxel only	Triangle only	**HSVO**
Sponza	57.3 fps	18.2 fps	20.6 fps
Urban Sprawl	40.3 fps	13.3 fps	23.7 fps
Happy Buddha	63.1 fps	10.1 fps	16.7 fps
Forest Scene	64.2 fps	2.4 fps	12.9 fps

Table 4: FPS of four different scenes rendered with a resolution of 1024×1024 using only primary rays and Phong lighting with simple shadows and a single point light source. Each scene was voxelized with a resolution of 512^3.

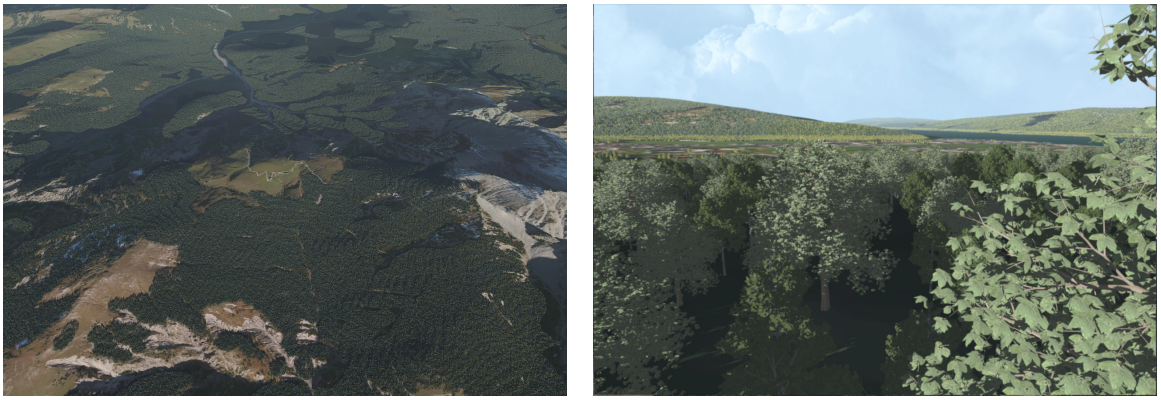
Full Octree Construction ($n_{split} = 1$)						
Scene	Nodes	Triangles	Triangle Index Array	Voxel	**Overall**	
Sponza	42.29	27.66	14.14	46.06	130.15	
Urban Sprawl	18.32	75.19	19.31	20.38	133.21	
Happy Buddha	11.42	103.07	21.94	11.95	148.38	
Forest Scene	30.41	156.25	34.58	33.29	254.53	
Early Construction Termination ($n_{split} = 2$)						
Scene	Nodes	Triangles	Triangle Index Array	Voxel	**Overall**	**Saved**
Sponza	10.89	27.66	12.51	13.97	65.03	**50.03%**
Urban Sprawl	12.37	75.19	18.47	14.77	120.81	**9.31%**
Happy Buddha	10.97	103.07	21.92	11.81	147.77	**0.41%**
Forest Scene	21.27	156.25	34.00	27.2	238.72	**6.21%**

Table 5: Size of the acceleration structure (MB). The upper part of the table shows the acceleration structure size of the test scenes for a tree built for all octree levels. The lower part of the table shows the presented method, where the tree is built only for nodes containing more than two triangles ($n_{split} = 2$) and lists the percentage of saved memory with respect to the full octree.

but is slow and offers no LoD. Therefore, aliasing is prevalent. The hybrid structure can place the emphasis on speed or on quality and offers LoD. Other approaches that combine rasterization and sample-based ray casting to render hybrid data were presented by Reichl et al. [Rei+12]. In their approach, all the polygonal data is subdivided into cubical bricks, essentially performing a voxelization. However, it is mainly used to accelerate rasterization using ray casting methods and not as a general rendering structure. Rendering using the HSVO is possible in real-time with rendering times per frame ranging from 12.9 ms in the *Forest* scene to 23.7 ms in the scene *Sponza*. However, measuring the frame rates for the hybrid approach is non-trivial since they increase significantly if parts of the scene show only the voxel data. For scenes such as *Sponza* showing an atrium where a camera is in effect "inside" the model, only a few camera positions can make use of the voxel data, resulting in only a small increase in speed. In the *Forest* or the *Urban Sprawl* scene, parts of the model are frequently in the distance. The *Forest* scene shows 13 highly-detailed plant models on a small plane. The *Urban Sprawl* model is a medium-sized but highly detailed city model. For these scenes, voxel data is used more frequently resulting in a significant speedup.

5.3 METRIC-DRIVEN EVALUATION

One advantage of the HSVO structure is that it promises to reduce aliasing at lower sampling densities. In this section, we investigate if this is indeed the case. In order to showcase its Anti-Aliasing (AA) abilities, the HSVO is used to render highly-complex vegetated areas with LoD. Here far distant models project to only a few pixels on screen, creating severe spatial aliasing artifacts. The model is created using our *Silva* [Wei+13] framework for generating tiled plant populations. This system creates such populations on an instantiable multi-level hierarchy. A nested hierarchy of kd-trees over Wang tiles with Poisson Disc Distributions is used to



(a) Fly-over (b) Close-up

Figure 60: Rendering of 40 million instantiated tree models using the hybrid acceleration structure.

represent individual plant locations. These tiles allow instanced but aperiodic repetitions to be created that enable large vegetated areas to be covered with "sod lawn"-like tiles. Each scene contains millions of highly-complex plant models reused several times. Although the *Silva* framework is able to generate and render such large populations, the rendered images show severe aliasing artifacts, when using only triangle data. Implementing a LoD system for plant models that is directly adapting the polygonal representation was not considered a viable solution. Here, we believe in the advantages of the HSVO.

The advantage of the hybrid representation over a polygonal simplification is that, within a regular octree structure, an approximation of high-frequency input models (e.g. trees), with different LoDs can be generated independently from the underlying geometric description. Polygonal simplification of such models usually fails due to the complex foliage and branching structure of the trees. One reason is that collapsing as well as merging operations might introduce triangles that do not correspond to the topology of the tree itself. Sample caching strategies in object space that provide LoD are limited to single instances, for example samples cannot be cached in the accelerations structure of a single tree due to the fact that it is reused in the scene several times. Sample caching strategies in world space as presented in Section 4.3 have the drawback of high storage requirements. Therefore, it is beneficial to have pre-filtered voxel data at hand to limit aliasing artifacts or in order to reduce the oversampling needed to create smooth animations and crisp images. Moreover, this speeds up rendering. The presented ray casting system allows billions of instantiated triangles that describe 40 million trees to be rendered at a resolution of 720p with about 5-7 Frames-per-Second (FPS) including direct shadows. Two renderings from the demo scene are presented in Figure 60.

In addition to the performance gain, the visual quality of the results can be estimated. Therefore, a fly-over of the scene was rendered with 1500 frames. The camera path was chosen to contain distant views on a large portion of the terrain, as well as close-up shots that allow the detail of individual plants to be appreciated. An image sequence of the scene is shown in Figure 61 (top). The scene was rendered multiple times with increasing samples-per-pixel (spps), using instantiated HSVOs, and using plain triangle data. Also, the scene was rendered with shadows turned on and turned off and all this at a resolution of 1080*p*. Finally, all conditions were compared to ground truth data computed with 128 spp. An introduction to all metrics used is provided in Section 3.2.1.

The results of the *PSNR-HVS-M* metric are plotted in Figure 61. This metric computes Peak Signal-to-Noise Ratio (PSNR) values that do consider a Contrast Sensitivity Function (CSF) model and visual masking. Figure 61a shows the development of these PSNR values using the HSVO when compared to a ground truth computed with the HSVO voxel description using 128 spp and without shadows. Likewise, Figure 61b illustrates the development of the PSNR values using only triangle data when compared to a ground truth using 128 spp and without shadows. The figures clearly show higher PSNR values when using the HSVO. Image noise levels are reduced significantly. This is also apparent when rendering the scene with shadows as shown in Figure 61e and Figure 61f. Unfortunately, using results with triangle data as ground truth does not work when evaluating the HSVO. Inspecting the results of the triangle and HSVO configurations and the respective image sequences, it becomes apparent that both representations provide slightly different images. A direct comparison is shown in Figure 62. Using voxels, surface materials are averaged into volumetric entities. Moreover, voxels create a more dense representation as they no longer tightly fit the underlying geometry. Similarly, the increased surface area is likely to be more directly illuminated by the light sources. Several of these issues are addressed in Section 5.6. Nonetheless, the PSNR values provide initial insights into the quality of the renderings as far as image noise is concerned.

While PSNR is commonly used to evaluate compression and image artifacts, higher-level video metrics exist; a popular one is Multi-Scale SSIM (MSSSIM). The results of MSSSIM when rendering with shadows is presented in Figure 63 (a,b). However, PSNR and MSSSIM compare image sequences frame-by-frame. Nonetheless, temporal stability is still an issue when estimating the perceived noise levels. One approach that combines several image and video metrics and takes motion into account is Netflix's Video Multi-Method Assessment Fusion (VMAF) [VMA17]. The results of VMAF when rendered with shadows is presented in Figure 63 (c,d).

Both MSSSIM and VMAF show the advantages of using HSVOs regarding image quality. Images rendered with 4 spp using the HSVO achieve quality ratings that compare to $8-16$ spp images rendered with triangles only. Images show significantly fewer artifacts at lower sampling densities and are more temporally stable according to all metrics. However, as all results are computed against two different ground truths, one can argue that the visual quality of the ground truths is not sufficient. However, the difference in the appearance of the two rendering approaches can hardly be solved. Hence, it can be beneficial to estimate noise and artifacts with a no-reference metric. The results of such measurements are presented in Appendix A.1.

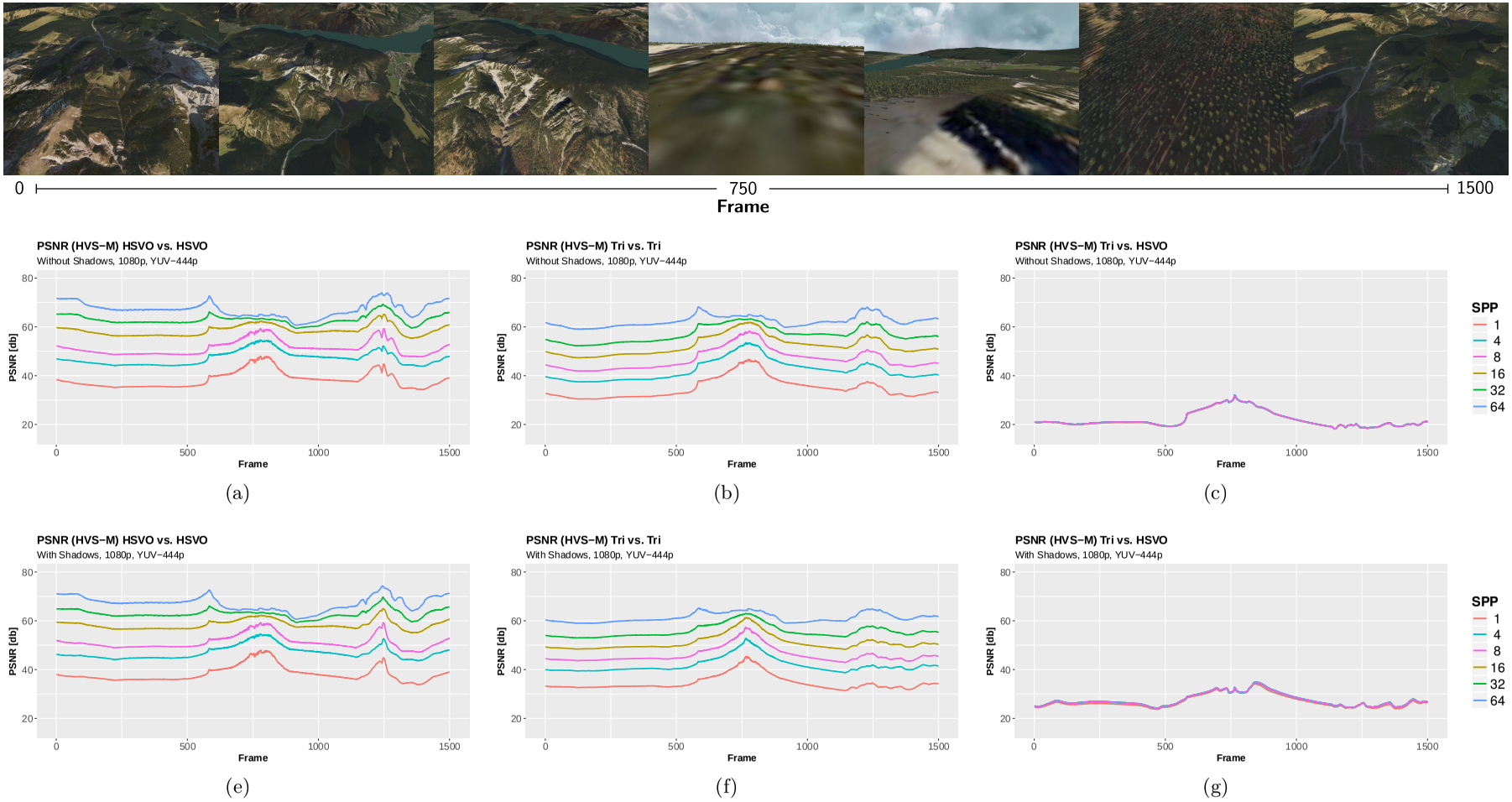


Figure 61: Still-frame sequence of the scene used for evaluation (top of figure) and the PSNR-HVS-M values for various factor combinations both without shadows (a, b, c) and with shadows (d, e, f) with different spp. Renderings using the HSVO or using only triangle data are compared with a ground truth rendered at 1080p@128 spp using the HSVO (a, d) or using only triangle data (b, c, e, f). According to the PSNR values using the HSVO yields a better signal-to-noise ratio compared to using triangle data (a vs. b and d vs. e). However, it is not feasible to compare the HSVO using triangle data as ground truth (c, f).

(a) Triangles only (b) Voxels only

Figure 62: Heavily-zoomed cropped image regions showing the difference in the visual appeal between using the triangle data (a) to using the hybrid voxel data (b). Although both representations provide pleasing images, a metric-based comparison between approaches does not yields comparable results.

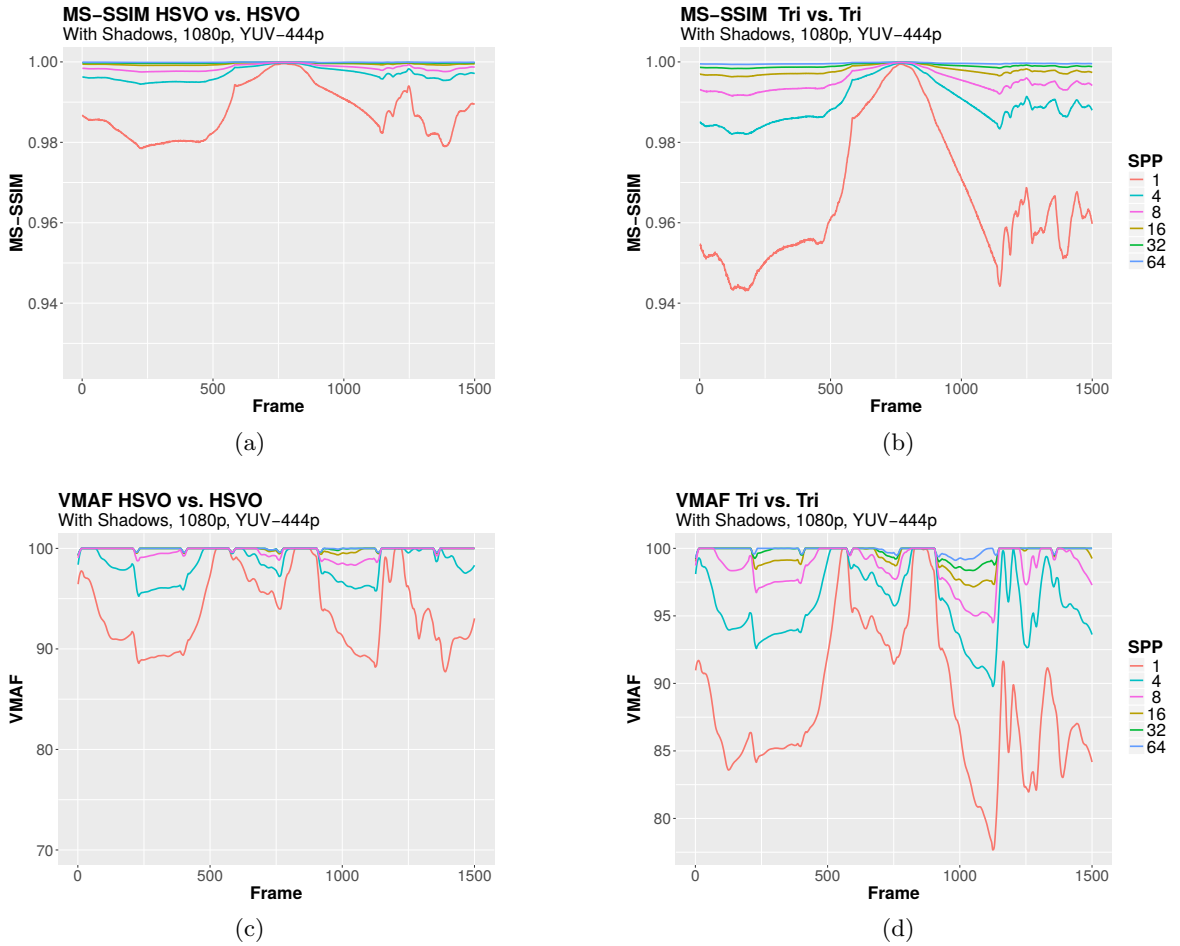


Figure 63: MSSSIM and VMAF values for the evaluation scene rendered with shadows at 1080p@128 spp with different spp. The scene was rendered using the HSVO and compared to a ground truth using HSVO (a and c). Likewise, the scene was rendered using triangles and compared with a ground truth only using triangles (b and d). MSSSIM yields a better structural similarity and VMAF yields results of a better quality for the HSVO at various sample counts compared to using triangle data.

Figure 64: Results of the focus and context strategy using head tracking in front of the large tiled display wall *HORNET*.

5.4 LOD SELECTION BASED ON THE VISUAL FIELD

The traversal routines presented in Section 5.1.3 allow to stop traversal of the structure early, if a voxel roughly projects to a pixel. This form of LoD selection allows for reducing aliasing artifacts while also increasing the rendering performance. However, rendering can also be adapted to other perceptual limitations, such as the user's limited visual field. Central for the evaluation of this approach is a large, high-resolution multi-display system (7×5 monitors) that offers a high pixel density on a large visualization area. This setup enables users to step up to the displays and see a small but highly detailed area. If the users move back a few steps, they do not perceive details at pixel level but will instead get an overview of the whole visualization. However, due to the number of pixels and the amount of data for complex scenes, implementing rendering methods that achieve interactive frame rates on such setups is challenging. This challenge can be tackled using the HSVO with an approach to drive LoD selection based on the user's FoV and acuity limits. The central idea is to parameterize the rendering in a way that the user's central visual field is rendered in high quality. The surrounding is rendered with an adaptive LoD that is chosen depending on the eccentricity in the user's visual field. This is demonstrated in Figure 64.

As users can move around, this allows for rendering highly detailed information when standing close to the multi-display system. Likewise, rendering quality can be adapted to the acuity limits when user's steps back to get the general overview of the whole scene. Here, the Human Visual System (HVS) is unable to resolve all details provided by the displays. The resolution of the displays exceeds the Minimum Angle of Resolution (MAR). To this end, the user's position and view-direction in front of the display wall are tracked using a six degrees of freedom tracking system. Even though eye tracking was not available, this approach mimics the acuity loss for peripheral vision as presented in Section 3.1.2.

Another advantage is that even if users notice a quality degradation in their peripheral visual field, the context of the focused region is preserved. Thus, this approach can also be regarded as a technique for focus and context.

Focus and context approaches are frequently used for 3D datasets in combination with direct volume rendering. Adaptive lenses [LHJ01] or x-ray and cut-planes [KSW06; RLH14] are tools for knowledge transfer, for instance, to visualize unique features of a volume. Other focus and context systems such as the *foveal inset* [Sta+06] use an additional projector to

display high-resolution content on top of a coarser projection. Most similar to our approach is the work by Papadopoulos et al. [PK13] for large tiled display walls. Here, a focus region can be selected to control the visualization of gigapixel imagery. Gigapixel images are challenging because of their size, where a system's bandwidth is the major problem. In the presented system the image generation itself is challenging for highly complex scenes at the required resolution.

The HSVO facilitate rendering coarse representations of a scene for the peripheral visual field and displaying polygonal data within the area with the central visual field. To do so, first, the user's view-direction needs to be obtained to determine what the user sees on the projection or display wall and which regions have to be rendered with all details. The following section thus introduces one approach to generate the Field-of-Sharp-Vision (FSV) based on the tracking input. Afterward, a metric is introduced that can be used to determine what to show on screen and in which resolution. Finally, a post-processing step is proposed to blur the transitions between multiple levels of the representation. All steps are presented in more detail in the next sections.

5.4.1 *Finding the User's Visible Area*

First, to render the specific area the user is looking at in a high LoD, the user's position and orientation in 3D space are determined with an optical tracking system. The FSV is specified manually as an angle describing the horizontal and vertical angular extent of the area to be rendered with all details. Now five rays are created to describe the user's FSV. A central ray defined by the direction the user is looking into and four bounding rays described by angular offsets used to describe the FSV. The intersection of these rays with a virtual model of the display wall is used to compute a leftmost, rightmost, topmost and bottommost intersection point $h_{left}, h_{right}, v_{top}, v_{bottom} \in \mathbb{R}$ all in normalized device coordinates.

5.4.2 *Metric to create the Detail-guide Image*

The intersection of the user's FSV and the display wall results in an elliptical/oval shape e. This shape is defined by the position of the central ray $c \in \mathbb{R}^2$, the user is looking at, as well as the bounds $h_{left}, h_{right}, v_{top}$ and v_{bottom}. These intersections are assumed to reside in an orthonormal coordinate system with the extent of the oval on the x and y axis. Now, these intersections are used to evaluate a metric for each ray that hits the view plane in a point $p \in \mathbb{R}^2$. This metric describes if a region has to be rendered in all its detail or if a coarser representation is sufficient, i.e. the level of the hybrid structure that is to be rendered at a specific location. The idea is to use the distance from the point p to the oval e to compute a greyscale image serving as LoD selection guide. This image is referred to as Detail-guide image (DGI). A DGI is presented in Figure 65b. Black areas in the DGI are to be rendered showing all details, decreasing down to the lowest LoD with increasing brightness.

A DGI is computed as follows: First, to determine if the point p lies within the oval shape it is projected to a unit circle. Now the quadrant of the oval shape e in which p falls is determined and the distances from $h_{left}, h_{right}, v_{top}, v_{bottom}$ to c depending on the quadrant

are computed. These distances are denoted $s = (s_{horiz}, s_{vert}), s \in \mathbb{R}^2$ respectively. Finally, $H(p)$ is computed as

$$H(p) = \begin{cases} 1, & \text{if } \left| (abs(p) - c) \cdot \begin{pmatrix} 1/s_{horiz} \\ 1/s_{vert} \end{pmatrix} \right| \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

If the point is inside the oval, zero is returned as distance, if it is not inside the distance d of the point p to the oval e is computed.

Finding the distance between a point and an oval is not straightforward. The matter is described in work by Eberly [Ebe13]. Each point p_e on an oval/ellipse centered in the coordinate systems origin can be described by

$$\begin{pmatrix} s_x cos(\phi) \\ s_y sin(\phi) \end{pmatrix}, \qquad 0 \leq \phi \leq 2\pi$$

Finding the distance of a point to a point on the oval/ellipse can thus be computed by the function

$$dist(\phi) = \sqrt{(p_x - s_x cos(\phi))^2 + (p_y - s_y cos(\phi))^2} \tag{3}$$

Given an oval shape, the distance from it to a point can be computed for a specific quadrant of the oval using $\phi : \phi \leq \theta$ with $\forall\theta \in [0, \pi/2]$. One solution to minimize this function is to use a nested interval approach to find ϕ. However, such an approach is not robust, is inaccurate and computationally demanding. Therefore, the approach introduced in Eberly [Ebe13] is used. Since for the closest point on an ellipse/oval it holds that the normal of this point must point towards p, equation (3) can be reformulated. This way and a new function $F(t)$ can be derived, for which its unique root gives the minimal distance of the point to the oval.

$$F(t) = \left(\frac{s_x p_x}{t + s_x^2}\right)^2 + \left(\frac{s_y p_y}{t + s_y^2}\right)^2 - 1 = 0, \qquad t \geq -s_x^2 \tag{4}$$

In order to find this unique root, a hybrid approach between bisection and newtons method is used. Please see Eberly [Ebe13] for implementation details. Finally, this distance is normalized according to the normalized device coordinates $d_{norm} = \frac{d}{\sqrt{(2)}}, d_{norm} \in [0, 1]$.

d_{norm} can be computed independently for each intersection of a ray with the view plane and is used during the traversal of the octree to determine when to stop traversing and thus to decide to display the coarse voxel or the fine polygonal representation. If the octree has n levels and it has been traversed to level k, traversal can be stopped if

$$k \geq round(n \cdot (1 - d_{norm}))$$

This way all possible LoDs can be visible on the screen. However, since the distance on the view plane has been normalized, the distance d_{norm} can be intuitively altered to represent a steeper or shallower fall-off of the voxel representation at increasing eccentricities in the rendering of the peripheral area.

5.4.3 *Post-processing the Images*

Although the traversal of the HSVO makes it possible to smooth visible transition between subsequent resolution levels (Section 5.1.3), transitions between LoDs in the peripheral vision are noticeable. The movements of the user can be very abrupt and are decoupled from the topologie of the 3D scene. Such transitions can be greatly disturbing to the user. In order to smooth these transitions, a post-processing step in image space is performed as follows: First, a Gaussian blur with a fixed σ is computed for the rendered image. This fixed σ allows for an efficient computation on the GPU. However, to still account for the fact that blurriness should increase with increasing eccentricities, the blurred image and the original image are blended in the following way:

$$I' = w \cdot I_b + (1 - w) \cdot I_s,\ w = \min\left(1, \frac{d}{f}\right),$$

where I_b is the blurred image, I_s is the sharp image, d is the distance from the ellipse border on the image plane and f is a user-specified fall-off constant, defining the size of the border area around the ellipse in which the interpolation between the two images occurs.

Figure 65a shows a rendering of the hybrid renderer and Figure 65b the used DGI. Note that in Figure 65a only a small area is rendered with all details. In the peripheral visual field, the blurred coarse voxel approximation is used.

5.5 USER STUDY

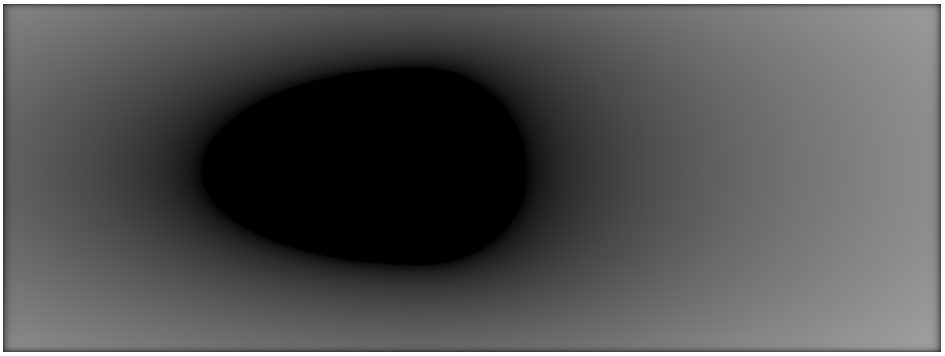
In this section, we present the results of a user study with the aim to measure the FSV and give first insights into perceptual implications. The following research question has been defined to evaluate the approach presented:

- **RQ:** What is an average FSV for which a user can barely distinguish if the image in his peripheral visual field is blurred or sharp?

5.5.1 *Procedure and Apparatus*

The user study was performed on HORNET, a large tiled display wall. HORNET's curved display surface measures approx. 7×3 square meters and consists of 35 Full HD monitors with less than three millimeters bezel each. The total resolution amounts to 72 megapixels. HORNET therefore allows for $1 : 1$ large-scale visualization at a very high resolution – sufficiently high to surpass the resolution of the human eye when standing more than two meters away from the displays. The monitors are driven by three *display PCs* that in turn have three NVIDIA GTX 780 cards with four outputs each. These PCs are fast enough to produce standard (OpenGL + shader, local illumination) graphics in real-time for moderate scene sizes. In order to allow interaction, HORNET is equipped with an optical tracking system consisting of seven tracking cameras with active infrared illumination. The visual field and the user's view-direction can be tracked with a specially prepared bicycle helmet. However, the user studies were performed in a static setup where the participants were not permitted to move and look around.

(a)



(b)

Figure 65: These images show the rendering of the Urban Sprawl scene (a) with a limited FSV and (b) the detail-guide image to control the reduction of visual accuracy.

A total of 14 subjects (9 male/7 female, all with an academic background and experience in Virtual Reality (VR)) participated in the experiment. All were between 18 and 47 years old ($M = 27.75$, $SD = 7.57$) with either no known serious visual impairments or subjects wore corrective glasses or contact lenses. Tracked tests were not performed because humans have difficulties focusing on a single point when they move around (saccadic eye movement) and gaze tracking was not available. Also, compensating for eye movement would need a higher update frequency from the renderer.

After signing an informed consent, each participant was placed in front of the HORNET display wall at a distance of 60 cm to the center display row. A fixation cross was displayed in their central visual field. The participants were told to focus on the center of the cross at all times. For each trial, a single angle was used for the FSV in both the horizontal and vertical direction. At this point, the study started by showing a rendering with all details for two seconds followed by a grey image for two seconds and finally a rendered image with blurred borders using a modified FSV for a further two seconds. In this way, three different Two Alternatives Forced Choice (2AFC) experiments were performed doing threshold testing [Lue+03, p. 284-285]. For each trial, the participants were presented the following question:

- **Q:** Is the visual peripheral sharp or blurred?

The first experiments started with a wide FSV that was then narrowed down. Each time the participants were asked if they considered the visual peripheral sharp or blurred. If the participants reported a discrepancy between the first and second image, the test was repeated with this FSV two further times. One run also was performed showing the completely

sharp rendered image twice. The grey image was displayed in between to make sure that the participants cannot notice a difference due to sudden changes in the peripheral visual field.

For the second experiment, the same test was performed but the other way around. It started with a narrow FSV that was widened in each trial run. During the first runs, the participants could distinguish between the image with all details and the partially adapted image. Each time the participants were asked if they considered the visual peripheral of the second image to be sharp or blurred.

In the final experiment, the combinations were randomly shuffled with FSVs from 0° to 180° in 6° steps. This time no grey image was displayed inbetween configurations. This experiment is closer to a real dynamic scenario since now the visual difference between two images can be perceived more directly. Again participants had to respond if they perceive the image in the periphery sharp or blurred.

5.5.2 *Results*

To answer the research question **RQ**, Figure 66a shows the results of the first experiment. Here, the FSV was gradually narrowed. If the participants reported to have noticed a change in the visual peripheral, the same conditions were repeated two more times. The angular size of the FSV was recorded if the participants reported a change in sharpness for all three consecutive runs. The aim of this procedure was to diminish the influence of not tracking the user's gaze. It is assumed that the subjects can focus on the fixation cross in at least one of the three runs of each trial. Thus, the angular dimensions of the FSV were obtained, for which the lower quality becomes perceivable. This also allows means and quantiles of the data to be computed. Likewise, Figure 66b shows the result if the FSV was gradually widened. Again the procedure was stopped if participants reported that a change in the visible peripheral was not visible for all three runs. In this case, the previous FSV was recorded, i.e. the one where the participants last noticed a change in sharpness.

Finally, Figure 66c shows the results when presenting random FSV. Here, only the angle is plotted when all narrower FSVs were identified to show all details. This way the trials were removed where the participants were doubtful or falsely identified the images as sharp and even narrower configurations were considered blurred. Considering all data and due to the fact that a 2AFC experiment was performed, it can be evaluated plotting a psychometric function. This function for the third experiment is shown in Figure 66d.

5.5.3 *Discussion*

After the user studies, the results in Figure 66a-66c show that on average an FSV of 130.5° is sufficient to create the sensation of a sharp image. It is important to note that probably a narrower FSV can be chosen if more effort is spent on blurring the images at increased eccentricities or, alternatively, data sets at a high voxelization resolution can be provided. In addition to this, in a post-questionnaire, most users reported that they were not concerned if they had a narrower FSV immediately as long as the context was preserved and, more importantly, renderings were updated at higher frame rates. By looking at the psychometric function for the third experiment in Figure 66d, it becomes evident that once the FSV exceeds

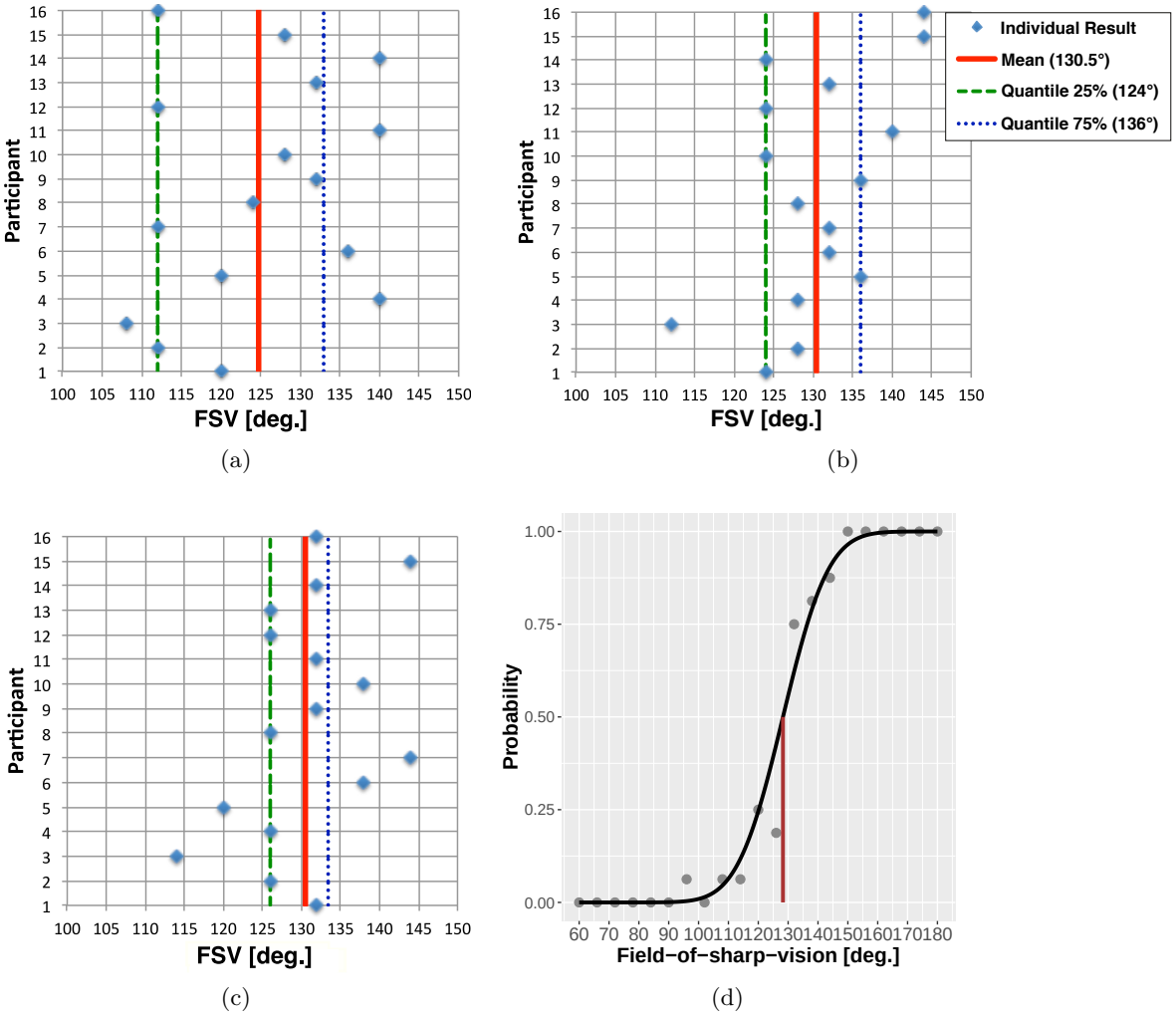


Figure 66: Results of the first experiment (a) where the FSV was gradually narrowed, (b) widened and (c) when testing FSV for randomly selected angles. The psychometric function of the third user experiment (c) shows that once the fields-of-sharp-vision exceeds 128.26°, images are rated with a probability $p > 0.5$ to be sharp.

128.26°, users rate the image to be sharp with a probability $p > 0.5$. This figure as well as Figure 66a-66c show that the spread of the values for the FSV is rather low. The participants had very consistent results.

Taking the results from the user study, performance impact of the presented LoD selection can be measured. The benchmark system was equipped with a Nvidia GTX 680 on an Intel Xeon E5520 system with 16 GiB RAM. In order to display the rendered images on HORNET the SAGE (Scalable Adaptive Graphics Environment) framework [DN02] was used. The DVI output of the render PC was digitized using a DVI grabber installed on a separate machine forwarding the video stream to SAGE. Therefore, only resolutions up to 4k were supported in this scenario. Since a single machine was used for rendering, images could be generated with about $8 - 10$ FPS. Hence, the absolute performance was barely interactive. The run times of the renderer using the mean FSV of 130.5°, determined in the user experiment, are compared to the run times of the renderer when displaying an image with all details showing the full polygonal model. In order to get realistic results, benchmarks were performed using

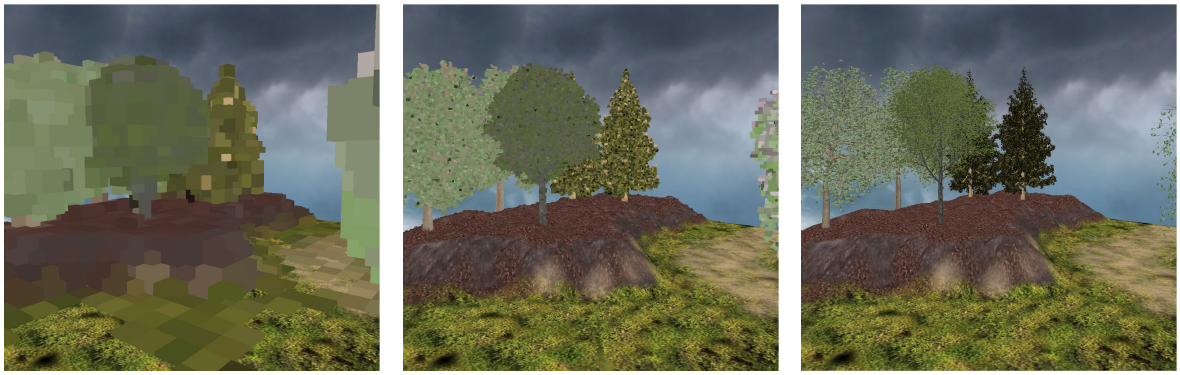


Figure 67: Three renderings of a forest scene showing octree level five (left), eight (middle) and the hybrid voxel polygonal output as rendered by the hybrid approach (right).

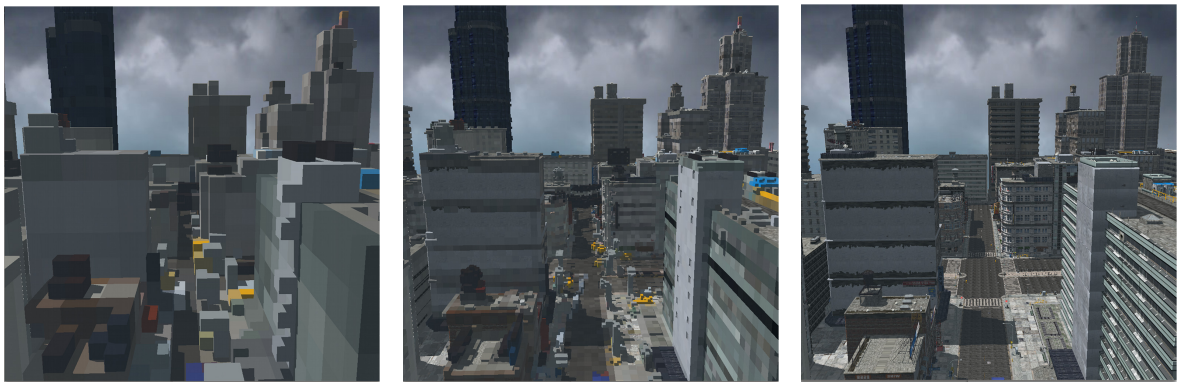


Figure 68: Three renderings of the urban sprawl scene showing octree level six (left), nine (middle) and the hybrid voxel polygonal output as rendered by the hybrid approach (right).

prerecorded tracking data of a representative examination of a dataset on HORNET. For a typical scenario where the users are allowed to walk or interact with the data set freely an average increase of speed can be obtained of up to 32% for the *Sponza* scene with 0.28 million triangles voxelized with a resolution of 512^3, down to 25% for the *Urban Sprawl* scene with 750k triangles voxelized in a resolution of 2048^3. All images were rendered in 4k. Full-HD showed similar results in the performance gain by rendering a narrow FSV.

5.6 APPLICATIONS AND LIMITATIONS

The hybrid structure presented is well-suited to applications that require a general LoD scheme since the regular voxel description enables a representation for arbitrary input meshes to be created. The HSVO can be seen as a multi-level grid, ignoring the fact that the HSVO contains a color and a normal for each grid cell. In this research, the hybrid structure was a logical step to counteract the artifacts present in the large-scale terrain and vegetation rendering system *Silva* [Wei+13] (Section 5.3). Moreover, the voxel representation substantially accelerates rendering. A scene showing several trees rendered at different LoDs is illustrated in Figure 67. Other examples where this LoD structure is beneficial are found in urban scenes. Figure 68 shows a general presentation of a city model rendered at different LoDs. Even though a polygonal simplification of such structures is not as challenging as it would

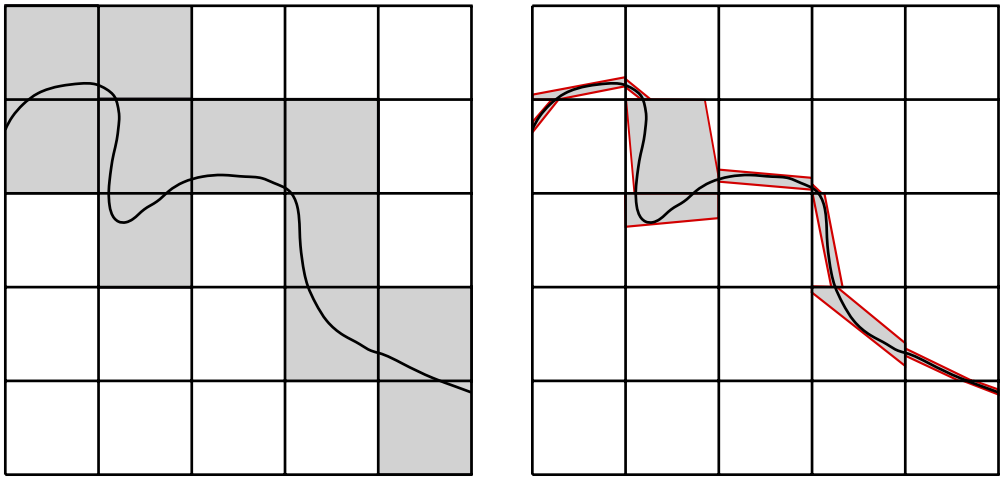
(a) (b)

Figure 69: Rendering of an urban environment (a) using the hybrid octree structure with voxel data in the background and (b) showing a color coding. The red areas are rendered using polygonal data and the green regions rendered using voxels.

be for tree models, renderings of such scenes from a distance have to compensate for high-frequency aliasing. The highly-varying z-depth of the scene generates spatial aliasing which can be reduced by having a pre-filtered voxel structure. Moreover, voxels are independent of the scene's local complexity and possibly large triangles in such a scene further reduce the size of the octree. The hybrid structure allows for smoother transitions and color blending between different levels as well as faster render times for highly-detailed areas in the scene viewed from a distance. A rendering of the model and a color coding of the internals of the structure are presented in Figure 69.

Although the presented LoD scheme can substantially reduce aliasing artifacts and has proven to be applicable for VR systems and perception-driven rendering approaches, a voxel representation has various limitations. In general, voxel representations have high storage requirements and since the presented construction on the GPU is performed in-core, the resolution of the voxelization is limited. Although out-of-core builds are possible in the latest version of the developed voxelization approach, all data still has to be held in GPU memory for rendering. Visually appealing models can easily require gigabytes of storage. Even though compaction methods such as Sparse Voxel DAGs [KSA13; Dad+16] have been developed, voxel representations usually fail to provide a balance between storage requirements and visual quality when used as first-order rendering primitives. However, voxel-based approaches are being successfully used for a fast and approximate simulation of Global Illumination (GI) accumulating secondary light contributions [Cra+11; Pan14; PSS14].

Memory management during construction remains an issue. The number of fragments generated by the voxelizer and the size of the octree as well as the triangle index list are not known in advance. Hence, buffers must either be preallocated with a maximal size or be used in a caching and paging scheme (e.g. [Cra+09]). Admittedly though, determining the size required for buffers is a problem most grid construction algorithms have in common. However, once the voxel fragment list has been generated, the approach presented allows the octree construction to be stopped early when too much memory is required in order to construct



(a) 2D illustration without contour planes

(b) 2D illustration with contour planes

(c) 3D rendering without contour planes

(d) 3D rendering with contour planes

Figure 70: Using contour planes in the hybrid renderer helps to discard empty space at an early point. However, although the model representation is better, using contour planes might produce aliasing artifacts and so incorrectly compute shadows are probably visible.

deeper levels. The system has already been extended to perform an out-of-core voxelization and construction for parts of the scene that have to be voxelized with a higher resolution.

Another issue with voxels arises from their regular structure and cube-like shape. Voxels store single colors and materials but often have to represent spatial entities that emit different colors in various directions. Even a plain triangle can have different textures on each of its sides. Hence, deciding which color to store and how to prefilter the values that reside in a voxel is challenging. Likewise, prefiltering a material description is equally problematic. It is often not possible to find an average material for various samples when the 3D function is prefiltered. Although this is an issue all LoD frameworks have to deal with, it is an especially challenging problem if space is subdivided regularly such as performed by SVOs. In this case, information from entities with different materials must be merged as soon as they reside in the same voxel. On the other hand, for approaches that simplify models using geometric meshes, the weight of each vertex can be altered to prevent the model from merging vertices that ought to be rendered with different materials. Hence, multi-level voxel representations work best for cases in which relatively simple materials are present. Nonetheless, these issues are problematic as becomes apparent in Section 5.3 for quality estimations between the hybrid renderer and a traditional rendering approach. Another issue is that while the structure is adaptive to space, it is not genuinely adaptive to the scene's input geometry. If there is a

highly-complex geometry inside a single leaf voxel, traversing these parts of the scene can have a considerable impact on performance. Merely building a tree to deeper levels by a regular subdivision of these parts is often not sufficient in order to subdivide the model's input geometry. It would be better to either identify these high-resolution parts beforehand and voxelize them separately, or automatically use truly adaptive acceleration structures such as BVHs or kD-Trees for these parts of the scene. However, since a coarser voxel representation is available, the renderer can decide to stop traversing these parts and display the coarse voxel representation in order to stay within constant frame rate limits. Another way to overcome such issues is to use the HSVO as the first layer in a multi-level grid, for example the grid implementation by Perard et al. [PKS17].

A further challenge with voxel structures involves the question of how to encode the empty space in voxels that contain geometry. If a planar surface resides within extent's of the voxel, its cubic shape probably fills considerable space that does not contain any geometry; thus it over-represents the geometric entity. This problem is illustrated in Figure 70a and presented in Figure 70c. One way to counteract this kind of artifacts is to allow voxels to be (semi-)transparent. Unfortunately, rendering (semi-)transparent voxel structures requires at some point a complete direct volume rendering pipeline and will become less efficient. Another way to mitigate such artifacts is to augment the voxel by additional *contour planes*, as illustrated in Figure 70b and presented in Figure 70d. The idea of such contour information is to limit the voxel space by using two planes for each voxel. Introduced by Laine and Karras [LK11], the contour planes are computed based on the average normal of the geometry inside the voxel. During construction, initially, two planes are considered, a lower and upper plane, both orthogonal to that normal. Now, the planes are shifted based on the minimum and maximum projection of each vertex in the voxel on the respective plane. This way these contour planes encapsulate the geometry of the respective node. As the SVO is traversed, each traversed level of the octree continuously limits the space. The only space considered by the intersection is the union of the space encapsulated by all contour planes from the root node down to the node of the current level. While the renderer presented was extended to support such contour planes, it also was extended to compute the contours in another way.

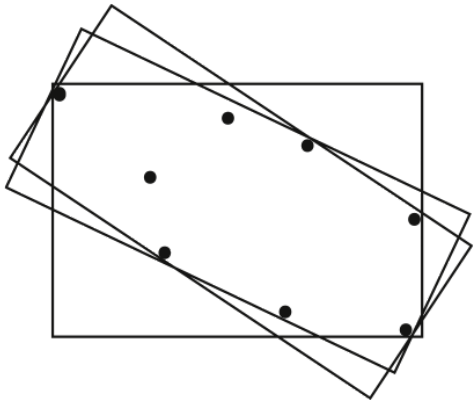


Figure 71: Finding the minimal oriented bounding box that is encapsulating the enclosed set of vertices as an alternative method to generate contour planes.

As averaged normals can wrongly represent the underlying geometry, a high-quality build mode attempts to find the best encapsulating planes by computing the minimal Oriented Bounding Box (OBB) for the geometry inside each node as illustrated in Figure 71. These OBBs are intersected from one level to the next as the SVO is constructed. However, the benchmarks presented and the quality estimation of the hybrid renderer were carried out without contour planes as they do not help to solve another serious issue of voxel representations, namely shadow artifacts.

As the cubic-like voxels may be blocked from the incoming light by other voxels surrounding them and voxels have a problem when representing flat non-axis aligned planar geometry,

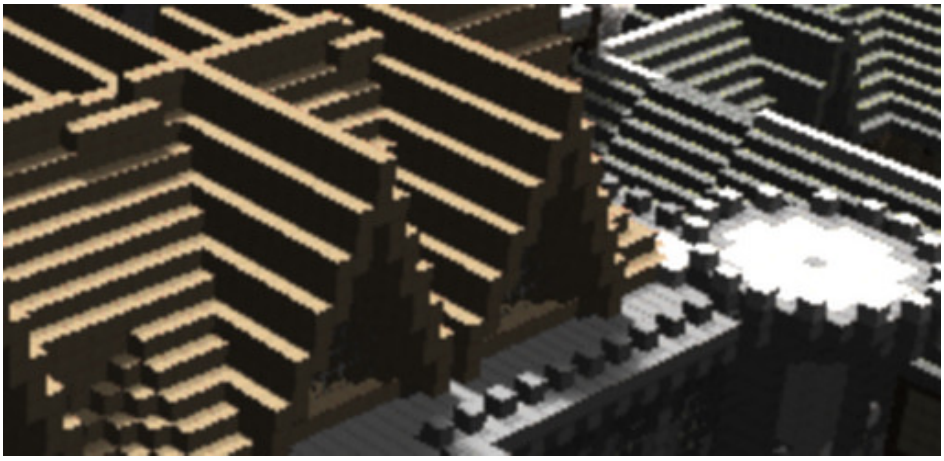


Figure 72: Voxel self-shadowing artifacts that are a severe cause for visually disturbing aliasing artifacts.

voxel representations are prone to self-shadowing. By looking at Figure 72, it should become clear that voxel-representation cause a many of such artifacts. Essentially these produce a very unpleasant high-frequency noise in the scene. Ultimately, shadows are simply another high-frequency signal likely to cause aliasing in the image that should have been removed by the voxel representation in the first place. In a ray tracing framework, the hitpoint with the voxels can be slightly shifted towards the observer or towards the light source for shadow rays in order to prevent them from intersecting direct neighbors. However, this is not the complete solution but merely limits some artifacts. Although it could be assumed that contour planes mitigate the artifacts, they are no much help. No matter how the contour planes are computed, the neighboring contour planes do not often have contact in single points. Hence, highly irregular self-shadowing artifacts are visible. Furthermore, as voxels are no longer cube-shaped, offsetting the ray's hitpoint is more challenging and wrongly computed shadows are more dominantly visible. Thus, it is often beneficial to use other strategies such as filtering shadowing artifacts [RS09]. For the large-scale tree models, also, dynamically switching between two approaches depending on the distance to the camera is possible. Offset shadow rays intersected with the coarse SVO representation are used when computing shadows for trees in the distance (far-field). Here, the loss of image quality due the reduced shadow precision is neglectable. For objects that are close to the camera (in the mid-field and near-field), and a high shadow quality matters, it is advisable to always use the triangle structure when intersecting shadow rays.

However, it should be noted that all of the artifacts mentioned above are less critical when using the perception-driven LoD selection scheme presented here. In this case, the coarse representation is only used for areas that are of minor interest in the visualization or are not visible or noticeable to the user. One limitation is that currently only single users are supported by the presented approach. While the respective FSVs could be combined for multiple users, the resulting larger FSV does increase the computational workload – potentially to a point where using the LoD scheme becomes superfluous. Nevertheless, an improved voxel representation probably helps in reducing the FSV for single and multi-user setups.

5.7 FUTURE WORK

The purpose of this chapter is to open up the limitations for further discussion as these provide the avenues for future work.

While voxel models have high storage requirements, they allow for a progressive refinement making them well-suited for out-of-core rendering. This field is targeted by research. An overview can be found in the work by Crassin et al. [Cra+09]. In this work, volumetric datasets are streamed into GPU memory using a high-level hierarchy of bricks representing subsets of the volumetric entities. Moreover, different approaches have been introduced to compact volumetric representations. Unfortunately, often these approaches only allow representing binary voxelizations or scalar volume datasets (MRT, CT, etc.). A survey on these methods is provided by Balsa et al. [Bal+14]. One approach well-suited for binary voxelizations are Sparse Voxel DAGs [KSA13]. Here, octrees are compressed to more compact Directed Acyclic Graphs (DAGs). Unfortunately, these DAGs lose their compactness once individual colors and normals ought to be stored. Hence, more general schemes like the work by Bas et al. [Dad+16] compresses attributes such as colors.

While indeed more work can be spent on more efficient compression and out-of-core schemes, still, combining appearances (materials and textures) is another challenge to face if voxel representations ought to be used as general rendering primitives. To this end, it is especially important to find efficient representations of direction-dependent information to be stored inside voxels efficiently. However, this will come at increasing costs. Likewise, more effort should be put in proper post-processing and LoD selection schemes in order to reduce regularity and alleviate the block structure in perceptually challenging situations. Due to processes such as lateral inhibition (Section 2.1.1), the block artifacts quickly become noticeable [Koh+05].

There are a lot of possible ideas for future work for the tracking scheme that adapts rendering according to the visual field. However, currently, the achieved rendering times are interactive and only sufficient for setups were only head tracking is used. Currently, there is an ongoing discussion on utilizing a separate dedicated render cluster that is available for HORNET and allowing for faster image generation due to more powerful hardware. Here, our rendering abstraction framework [GWH17] that encapsulates renderers and message passing (MPI) for synchronization using Docker containers could be one valuable tool. Moreover, to further benefit from the FSV, different load balancing strategies ought to be developed to distribute the workload to the different render PCs in the cluster.

Eventually, with newer hardware generations, HSVOs will be fast enough to be used in gaze-tracked setups where the system has to cope with fast eye movements. As gaze-tracking was not available for HORNET and rendering times were only just interactive ($8 - 10$ FPS) it has to be left for future work how the required FSV changes in dynamic or gaze-tracked scenarios. Here, post-processing filters and better optimized bilateral blurring could prevent transition artifacts from becoming perceivable. Also, even though a peripheral LoD change might be visible in some scenarios, it usually should not bother the user to fulfill a specific task – probably at equal speed and precision. Lastly, given the interactive construction times that can be achieved already, it should be possible to use HSVO for dynamic scenes on upcoming GPU generations as well.

5.8 CONCLUSION

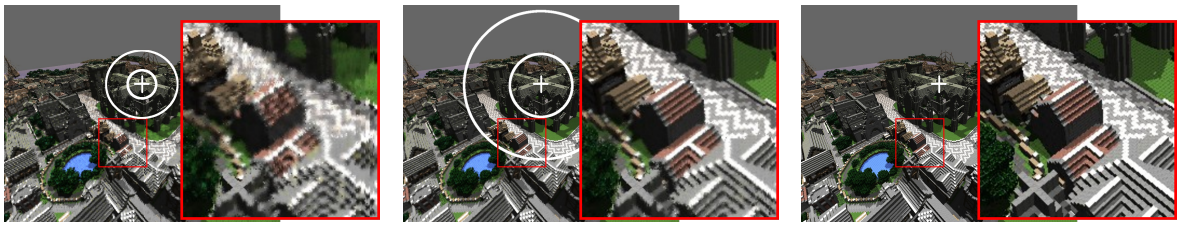
This chapter has presented an approach to building a hybrid acceleration structure storing voxels for inner nodes, stopping construction of deeper levels if the number of primitives within a node is not greater than n_{split}, and storing the full triangle list for each leaf node that represents the finest voxelized level. This way, an LoD description of the input geometry is generated so that sampling can be performed more efficiently. In addition, the presented hybrid acceleration structure makes a substantial reduction in aliasing artifacts possible. Nonetheless, the inherent issues and presented limitations, especially the high memory requirements and the challenges with direction-dependency and the shadowing artifacts must be considered. In the last years, traditional LoD approaches to cope with aliasing artifacts are becoming less prominent for use-cases such as video games, where it is possible to hand-tune polygonal information in order to meet perceptual and performance requirements. The current GPU generations are far less sensitive in dealing with massive triangular meshes - triangle throughputs and fill rates have substantially increased over the last decades. In 2016 a newly released NVIDIA had a peak pixel fillrate of 111 Gpixels/s and a peak rasterization rate of 6.9 Gtris/s. Ten years earlier, the high-end NVIDIA GTX 7900 only achieved 10.4 Gpixels/s with a peak rasterization rate of about 0.25 GTris/s in ideal conditions.

In general, if render times are not an issue, it is usually better to take more samples to counteract artifacts. For specific scenarios that require interactivity such as video games, for example when rendering outdoor scenes and vegetation, specialized LoD schemes have proven to be of a higher value. Industry can afford to maintain approaches for all the various aspects of realistic scenes. Nonetheless, most AA is handled by techniques such as Multisampling Anti-Aliasing (MSAA) or Temporal Anti-Aliasing (TAA) that are industry standard [Tat+16]. Despite this, the presented LoD approach is well-suited for lab setups and specialized VR installations where arbitrary models need to be visualized and hand-tuning or specialized rendering approaches are not available. This is the case for high-quality visualizations of large outdoor scenes which must be generated quickly for mostly unknown datasets. For such installations, a perception-driven LoD selection mechanism has been introduced here to enhance the rendering performance for large tiled display walls. Even though the experimentally found FSV of 130.5° would appear to be wide, it is sufficient for typical scenarios to significantly increase the application's speed. Using the presented display wall, this FSV of 130.5° means that standing more than two meters away from the wall, the image has to be rendered with the full resolution, i.e. the FSV covers the entire wall to match the visual acuity of 1°. For increased distances, reducing the display's resolution is possible as rendering in 1080p on each display exceeds the visual acuity limits at these and greater distances. In practice, one observation is that people using display walls to discuss data sets stand very close to them. Here the system presented in this chapter is especially beneficial. Also, the size of the FSV can probably be further reduced by applying more elaborate post-processing methods such as an edge-aware bilateral blur to hide transitions between the voxels and polygonal representation.

Although the view-directed LoD selection approach has proven to be a valuable tool for large high-resolution display walls and projection systems, it is challenging to achieve interactive render times. Hence, the approach presented here can not be used for gaze-contingent rendering. To this end, it is often beneficial to adapt the sampling itself and so meet the low-latency requirements to cope with fast eye movements.

6 FOVEATED RAY TRACING

Exploiting the Limitations of the Sensor



(a) Small foveal region with (r_0=5°, r_1=10°, p_{min}=0.01)

(b) Medium foveal region with (r_0=10°, r_1=20°, p_{min}=0.05)

(c) Full Renderer

Figure 73: Images generated with the foveated renderer showing the effect of different configurations for the foveal region, including an image that was rendered by ray tracing every pixel.

In previous years, advances in display technologies and mass production have led to the introduction of a range of high-quality Head-Mounted Displays (HMDs) with a wide Field of View (FoV), which have attracted the interest of researchers in the field of virtual and augmented reality. At the same time, pixel densities have dramatically increased over the last two decades (e.g. *Forte VFX 3D* at 263×480×2 in 1998 and *StarVR* at 2560×1440×2 in 2016, Appendix A.4). One of the key challenges when rendering to HMDs is achieving low latencies while filling the pixels. This is crucial to increase presence and reduce fatigue. The combination of high pixel densities and refresh rates are a major challenge when bringing image synthesis algorithms to HMDs and yet as far as resolution is concerned vast improvements are required. Providing the highest possible visual quality at retinal resolution would require at least 432M pixels for the full dynamic field of view of the human eye (Appendix A.4). Rendering at such resolutions with the necessary refresh rates ($\geq$ 75 Hz) is far beyond the reach of current and foreseeable hardware and software solutions.

One possible solution for increasing rendering efficiency is by a gaze-contingent adaption of the visual quality in rendering systems based on the decreasing visual acuity with increasing eccentricities of the visual field. Knowing the user's gaze and his Point-of-Regard (PoR) on the image plane allows the sampling to be adapted to the retinal limitations. As presented in Section 4.2.3, several of such *foveated rendering* methods have been introduced in recent years. This is made possible by the availability of reasonably-priced, accurate eye tracking devices. The integration of these devices in HMDs has further increased the popularity as the fixed relative positioning between eye and screen effectively reduces calibration costs and improves accuracy. These necessary accuracies when determining the PoR are usually not achievable when users are allowed to move in front of large-scale Virtual Reality (VR) installations, such as the large high-resolution display wall mentioned in the previous chapter. Likewise, the detailed method to adapt the visual quality using Level-of-Detail (LoD) and voxels has

proven to be not sufficiently efficient to be used inside an HMD. Especially considering the fact that its main purpose is to visualize highly complex geometry. For low-latency rendering of commodity 3D scenes, it is better to adapt sampling directly.

This chapter of the thesis describes the development, a foveated rendering system based on ray tracing that is capable of rendering high-quality images quickly enough for modern HMDs (Figure 73). The sample density of the rendering process is reduced by adapting the ray generation to visual acuity. In contrast to the methods available when this research began, a focus is placed on ray tracing as the primary rendering method as it has several distinct advantages when rendering to HMDs (Section 4.2.3). One of the key challenges for foveated rendering methods is to reconstruct images at a high quality in order to limit the detection of visual artifacts. To this end, the presented approach is coupled with a reprojection scheme to increase temporal stability. The reprojection does combine samples thus creating a smoothly refined image. Parts of the image that expose high contrasts are resampled, as those are likely to cause artifacts that remain perceivable due to the eye's contrast sensitivity (Section 3.1.3). Using the approach presented, missing information from the sampling process can be reconstructed either using a *support image* that is guaranteed to sample the full scene using a lower uniform resolution or by using information from reprojected frames to improve the quality of the reconstructed final image.

The benchmarks demonstrate the high performance of the presented implementation when compared to standard ray tracing. In order to determine the methods perceptual quality, this chapter also presents the results of a accompanying user study which employs an Oculus Rift DK2 equipped with an eye tracker. This made it possible to substantiate the high visual quality by the approach presented. Without overly prejudging the presented results, *visual tunneling* (Section 2.2.5) and retinal velocity (Section 3.1.2) had interesting effects in the user study when the visual quality needs to be judged. This demonstrates how mental workload and other perceptual properties can be used to further optimize foveated rendering systems. Fewer samples can be generated in the periphery when users concentrate on a specific part of the scene or need to accomplish a task in the virtual world.

In summary, this part of the thesis presents a gaze-contingent rendering system for HMDs including the following contributions:

- A high-performance, adaptive sampling approach for ray tracing driven by eye tracking and limitations of human perception.
- A reprojecting and merging process using a coarse approximation of the scene geometry to support reconstruction of the final image from sparse samples.
- An estimation of the tracking precision and fixation accuracy, supported by the evaluation of eccentricity-based quality ratings.
- A user-study showing that the method only has minimal impact on the perceived quality when regarding foveal region limits. The study also reveals a great potential for deploying visual attention to further optimize foveated rendering techniques.
- An analysis of the connection between subjective perceived quality and fixation accuracy, providing possible evidence of the presence of visual tunneling effects and the magnitude of their influence on the user's perception.

CONTRIBUTIONS BY THE AUTHOR This chapter is based on work published in the paper:

> Martin Weier, Thorsten Roth, Ernst Kruijff, André Hinkenjann, Arsène Pérard-Gayot, Philipp Slusallek, and Yongmin Li. *"Foveated Real-Time Ray Tracing for Head-Mounted Displays."* In: *Computer Graphics Forum (Proceedings of Pacific Graphics '16)*. Oct. 2016.

I was the primary investigator for this paper and developed the GPU-based rendering, reprojection, and merging pipeline for foveated rendering as presented in the following sections. My co-author Thorsten Roth supported this work by developing the ray generation kernels for sparse sampling and the post-processing filter when using stochastic rendering methods. As the latter components are not the author's original work, only a few details are provided in this thesis in Section 6.1.1 and Section 6.1.5. The ray tracing core itself was created in collaboration with Arsène Pérard-Gayot. It was later also used for comparison purposes in our paper:

> Arsène Pérard-Gayot, Martin Weier, Richard Membarth, Philipp Slusallek, Roland Leißa, and Sebastian Hack. *"RaTrace: Simple and Efficient Abstractions for BVH Ray Traversal Algorithms."* In: *Proceedings of the 16th International Conference on Generative Programming: Concepts & Experiences (GPCE)*. ACM. Vancouver, BC, Canada, Oct. 2017, pp. 157–168.

The design and evaluation of the user study, presented in Section 6.3, was done in collaboration with my colleagues Thorsten Roth and Ernst Kruijff. Noteworthy in this context is the evaluation of the eye tracking data. My colleague Thorsten came up with the initial ideas and details are provided in our paper:

> Thorsten Roth, Martin Weier, André Hinkenjann, Yongmin Li, and Philipp Slusallek. *"A Quality-Centered Analysis of Eye Tracking Data in Foveated Rendering."* In: *Journal of Eye Movement Research (JEMR)* 10.5 (2017).

While some of this work can be found in this thesis, I explicitly marked the relevant places. Here, I also added a new evaluation of the retinal velocity. This evaluation provides a further explanation to our findings on visual tunneling effects that has so far not been discussed. Details on this matter are provided in Section 6.3.3.

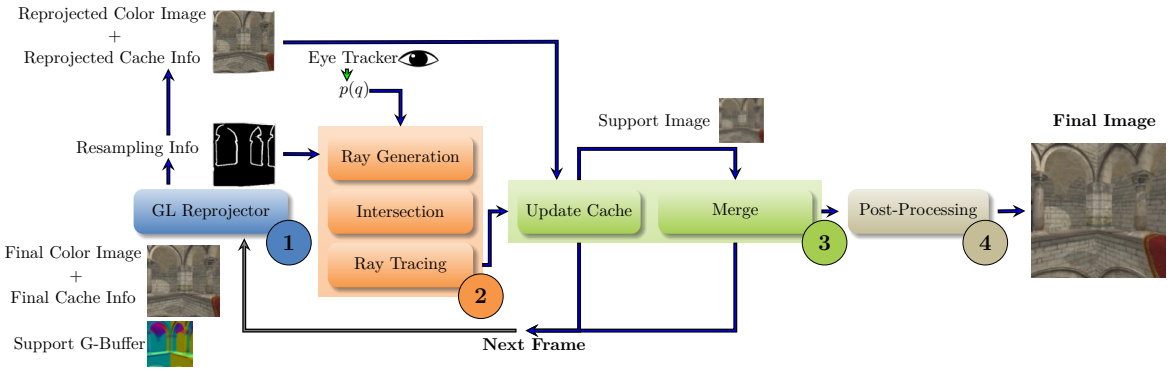


Figure 74: Building blocks of the reprojection pipeline. Old *color* and auxiliary *cache info* are reprojected using a *GL Reprojector* (Block 1). New rays are generated based on the user's gaze and possible errors introduced by the reprojection are marked in a *resampling info* buffer. The ray traced pixel values (Block 2) are blended using a temporal caching and merging scheme (Block 3). An optional *Post-Processing* is used to smooth artifacts arising from stochastic sampling processes such as ambient occlusion (Block 4).

6.1 METHOD

In this section, the building blocks of the foveated rendering system are described. An overview of the entire pipeline is presented in Figure 74. The system's core is a fast ray tracer based on NVIDIA CUDA using an SBVH acceleration structure [SFD09] (Figure 74, Block 2). It generates a sampling pattern from three parameters describing a foveal region to account for the user's visual acuity and gaze (Section 6.1.1). However, this foveated sampling process results in a sparse image. This means that with increasing eccentricities there are proportionally larger gaps between sampled pixels. Presenting such an image to the user would not meet the perceptual requirements, as the gaps would result in the sparse image's brightness being vastly different from the fully sampled image. Also, strong temporal flickering would be present due to the stochastic sampling process, responsible for generating rays. Thus, it is necessary to provide a method that improves image quality outside the foveal region, generating a smooth image from the sparse samples. This method has to meet specific requirements: While performance has to be high enough to stay within VSync limits (13.3 ms at 75 Hz), image quality has to suffice human perception.

In order to maintain a high image quality but still meet performance requirements the central idea of the presented approach is to exploit Temporal Coherence (TC) between subsequent frames to increase the number of available samples. A frame rendered a timestep $t-1$ is projected to a new frame at timestep t using a coarse mesh that is generated from the user's view. Artifacts arising from the reprojection and the use of the coarse mesh as well as depth and brightness discontinuities are detected. This information is used to guide the sampling process as resolving those artifacts is critical for the perception in the peripheral visual field. More samples are generated based on the user's gaze in combination with a temporal caching and merging scheme. This allows for reusing samples across time and space.

Buffer Name	Resolution	Components	Description
`Reprojected Image`	Native HMD Resolution	RGBA32F	Reprojected color image that is used to compute the current frame.
`Reprojected Cache Info`	Native HMD Resolution	RGBA32F	Reprojected weights for temporal integration. The weights are used when combining new samples and the reprojected image.
`Final Image`	Native HMD Resolution	RGBA32F	Last frame's color information.
`Final Cache Info`	Native HMD Resolution	RGBA32F	Last frame's weights for temporal integration.
`Resampling Info`	Native HMD Resolution	R8	Auxiliary buffer to mark parts that benefit from additional sampling.
`Support Image`	Reduced Resolution	RGBA32F	Low-resolution color buffer to fill in values that cannot be reconstructed from the last frame or the new samples.
`Support G-Buffer`	Reduced Resolution	RGBA32F	Low-resolution G-Buffer used to reconstruct coarse geometry to reproject the next frame.

Table 6: Overview of the buffers that are used in the rendering pipeline. The last three buffers increase the storage overhead compared to traditional temporal anti-aliasing approaches [Yan+09]. However, these operate at low resolutions or can be stored in a single channel.

The rendering pipeline consists of four steps (Figure 74):

1. The calculation of a new frame starts with the processing of an old frame (Figure 74, Block 1). The *final color image* of the old frame with additional information is reprojected using a low resolution *support G-buffer*. The latter is used to generate a warping geometry to turn the *final color* and *cache info* into textured geometry. The textured geometry is now reprojected into the next frame using OpenGL. In addition, this step detects regions in the image where this reprojection failed or which have high contrasts and as a result a high saliency in peripheral vision. These areas are marked in an auxiliary buffer called *resampling info*.
2. Eye tracker inputs with a foveated sampling scheme and the *resampling info* are used to calculate new samples using ray tracing (Figure 74, Block 2).
3. Finally, the information is combined and a final image is (re-)constructed (Figure 74, Block 3). Here, missing samples can either be reconstructed from the reprojected previous frame or from the low-resolution color and G-buffers (*support image* and *support G-buffer*). These are fully updated per frame. However, the resolution of these buffers is only a fraction of the target resolution required for the HMD. The *support image* contains a regular low-resolution color image, while the *support G-buffer* contains the geometric normals and depth values that are in turn used to reconstruct the coarse geometry for the next frame.
4. An optional post-processing step (Figure 74, Block 4) can further improve image quality when stochastic sampling processes are used.

An overview of the buffers used and their descriptions can be found in Table 6. The individual steps of the pipeline are described in more detail in the following sections.

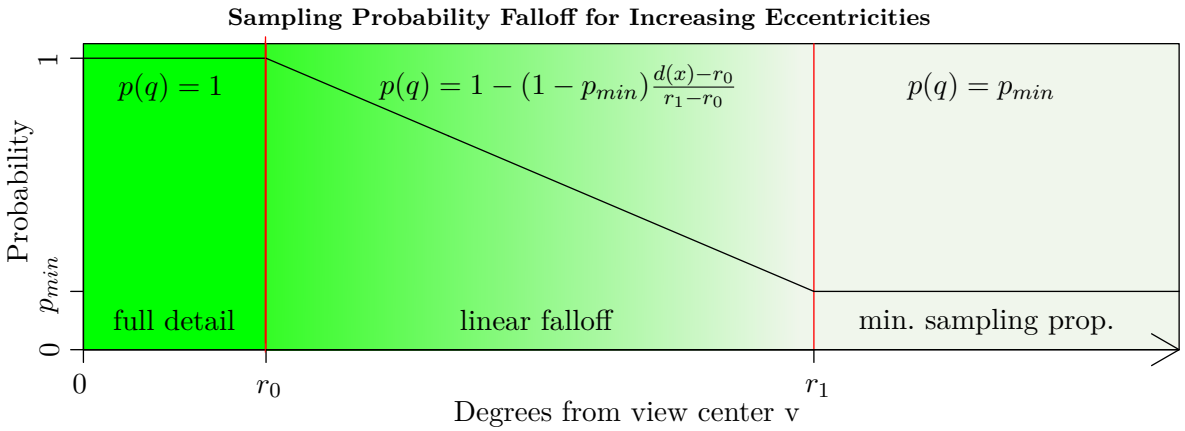


Figure 75: Probability $p(q)$ for sampling a specific pixel q based on its eccentricity and user-defined parameters r_0, r_1, and p_{min}. Despite the hyperbolic fall-off of cones towards outer regions a linear fall-off is employed to improve motion perception in the periphery and to reduce spatial and temporal aliasing artifacts in these areas.

6.1.1 *Ray Generation and Ray Tracing*

Initially, a ray generation kernel is launched which selects the pixels to render regarding the *visual acuity*. As presented in Section 3.1.2, visual acuity is subject to a hyperbolic fall-off with increasing eccentricity. However, instead of relying on the models by Guenter [Gue+12] or Reedy [Red01], for this work it was decided to use a model that allows for a more intuitive adaptation of the visual field with a steep linear fall-off. Although this model yields a slightly higher sampling rate in comparison to models with a hyperbolic fall-off (Section 3.1.2), all acuity models are user-dependent anyhow and must be carefully tuned to each subject. Also, the introduced linear model guarantees a minimum sampling density for the peripheral visual field. Due to the higher sampling rate, this limit reduces spatial as well as temporal aliasing artifacts in these areas that are critical due to the high flickering sensitivity at larger angular distances to the center of vision. The spatial *visual acuity model* used is illustrated in Figure 75 and compared to other models in Section 3.1.2.

In order to achieve a linear behavior, the ray generation is based on two user-defined *eccentricity* thresholds: An inner threshold r_0 and an outer threshold r_1, both given in degrees of the visual field. Here, r_0 determines the size of the foveal region (i.e. the area rendered in full detail). The angular eccentricity r_1 together with the minimum sampling probability p_{min} determines the probability fall-off beyond r_0. These three values represent a user-controllable triplet, referred to as a Foveal Region Configuration (FRC). Pixels with a larger eccentricity than r_1 are only sampled with a probability of p_{min}. Sampling probabilities $p(q)$ are computed based on the eccentricity of each pixel using the function presented in Figure 75. Generally, pixels are only selected for sampling if $\xi_q < p(q)$ with $\xi_q \in [0, 1]$ being a uniformly distributed random number. In addition, the sampler schedules all pixels for sampling that are required for the uniform lower resolution *support image* or that are marked in the *resampling info* generated by the reprojection (Figure 74). Based on the selected samples, the GPU kernel stores the respective primary rays in an array using atomic index increments.

Next, the generated rays are intersected with the scene geometry resulting in a list of hit points and their respective pixel indices. The kernel introduced by Aila and Laine [ALK12] is employed for ray traversal.

After the intersections, a kernel computes shaded pixel colors for the rays' hit points. Shading currently supports Phong lighting with mipmapping, ambient occlusion, point, and area light sources. The irradiance from the area light sources and the ambient occlusion value are stored in a separate *light buffer*. This way the running estimate, used to combine samples temporally, can be adapted to different rates. This enables a separation of color and irradiance information allowing to reduce noise introduced by stochastic processes, for example when sampling area light sources or ambient occlusion.

6.1.2 *Reprojection*

After computing an array of newly shaded samples along with the pixel indices for the current frame, the reconstruction relies on information from the previous frame. As computing more samples is expensive, exploiting TC helps to increase performance while it can also improve image quality. In order to increase the available samples spatially, *forward-reprojection* is performed. However, while it is theoretically possible to understand each sample as own unit in space, reprojection each of those samples individually is costly. Even more challenging is that this way it is hard to determine and resolve occlusions and disocclusions correctly. Hence, the method forward-reprojects samples using the rough scene geometry. As presented in Section 4.3, possible approaches for such types of forward-reprojection have been introduced by Simmons and Séquin [SS00] and Tole et al. [Tol+02]. These construct and update an irregular mesh with potentially as many vertices as pixels at the highest quality level. However, such approaches are costly and ray tracing can be fast (up to 200 Mrays/sec for traversal [ALK12]). Usually, computationally expensive methods are not worth the effort.

In order to maintain high refresh-rates, this work uses a reprojection strategy based on a coarse and uniform mesh that can be handled most efficiently. Although this process is likely to produce errors in certain image areas, they can be resolved more efficiently by computing additional samples instead of constructing a more precise mesh representation. The reprojection process (Figure 74, Block 1) transforms the *final color image* along with the *final cache info* to a new *reprojected color image* and *reprojected cache info* buffer. The *cache info* buffers are used to keep a state in the buffer that maintains how samples should be combined temporally. As described in Section 6.1.4, this buffer is a `float4` texture that is reprojected along with the color information. For each frame, a uniform mesh is generated from the *support G-Buffer* by creating and displacing a uniform grid of vertices matching the *support G-Buffer's* resolution.

In order to reconstruct the scene geometry from the depth information stored in the *support G-buffer*, the vertices are adjusted to these image space depth values with a geometry shader. Afterward, an "unprojecting" step is performed using the model, view, and projection (MVP) matrices of the previous frame. These were also used to create the *support G-buffer* and yield the vertex positions in world space representing the surface of the visible scene geometry from the previous frame. In the next step, the vertices are forward re-projected to the new frame using the current *MVP* matrices. This mesh is rasterized at the full rendering resolution,

textured with the last frame's *final color image* and *final cache info*, finally yielding the reprojected version of the previous frame.

Due to the changed perspective, each pixel's footprint may cover a couple of texels of the previous frame's *final color image*. Therefore, simple bilinear filtering of the texture is not sufficient and special care has to be taken to filter this texture during rendering. As computing a mipmap hierarchy along with anisotropic filtering for the reprojected texture per frame based on the new pixel's footprint is too expensive, another texture filtering method has to be used. This filtering method randomly samples the pixel footprint multiple times using a normal distribution inside the fragment shader in order to compute the final reprojected color.

6.1.3 *Handling Reprojection Errors*

The uniform mesh employed for reprojection is not a perfect representation of the actual scene geometry. This may lead to perceptible errors. Possible causes for these errors include geometry newly entering the view frustum, disocclusions, and undersampling [MMB97]. If a part of the scene has not been inside the view frustum in the previous frame and sampling has not been triggered by the visual acuity model, missing pixels are reconstructed from the coarse *support image* (Section 6.1.4). Disocclusions and undersampling can both cause strong visual artifacts to appear in the image (Figure 73a). This is caused by incorrectly or incompletely reprojected information. Therefore, an additional render step attempts to detect and create additional samples for areas with such artifacts, consequently improving perceived image quality.

First, to detect regions that need further sampling, the scene is rendered using the coarse resolution matching the *support image* and *support G-Buffer* using the reprojection procedure described in Section 6.1.2. If there is a depth or luminance difference between a pixel and its direct neighborhood in the reprojected image that is larger than a user-defined threshold ϵ_{depth} or ϵ_{lum}, a pixel is marked for resampling in the *resampling info*. This process resembles edge-detection steps in post-processing methods like Subpixel Morphological Anti-Aliasing (SMAA) (Section 4.4) and schedules critical regions for resampling. Interestingly, this entire process shows similarities to the recently proposed hybrid Adaptive Temporal Anti-aliasing (ATAA) (Section 4.2.1).

Depending on the chosen value for ϵ, geometry that does not resemble the scene may be used for reprojection anyway, e.g., in case of relatively flat objects in front of a wall. If such geometry is looked at frontally in frame $t-1$, moving the camera to frame t can result in undersampling artifacts because the possibly wrongly closed geometry is reconstructed, connecting the object to the wall. These objects might expose depth and luminance distances well below the respective ϵ-thresholds, while the closed geometry resulting from the reprojection process is actually wrong [MMB97]. Such surfaces occur along the user's viewing direction, i.e., when the angle between the surface normal and the observer is close to 90°. These corner cases are detected with an additional test looking at the surface geometry. From the previous frame's geometric normal $\vec{n}$ and camera orientation $\vec{d}$, we compute $\text{edge}_t = \max\{\vec{n} \cdot \vec{d}, 0\}$. If $\text{edge}_t < \epsilon$, the pixel is marked for ray generation. Partial derivatives of texture coordinates would be another measure to detect regions that need sampling. They yield information about an observer's angle towards a potentially undersampled surface. However, in this work no no-

ticeable visual enhancements by using this information could be found as head movements are limited when wearing an HMD. In case of complex geometry, a considerable part of the image may be covered by possibly undetected and thus undersampled edges (Figure 73a). This necessitates a measure for sample quality accounting for a sample's age, as presented in the next section.

6.1.4 *Cache Update and Merging*

At this point, the current image only consists of the previous frame's *reprojected color image* (Figure 74). Newly shaded samples from the ray tracing process have to be combined with this cache image using a temporal blending method. This accumulation process should be designed in a way that reduces the weight of older samples, as simply accumulating samples with equal weights does not make sense for two reasons: First, due to the sparse sampling process, each pixel may have been sampled last at a different point in time. Second, assigning a high weight to old samples leads to visual artifacts like smearing on edges. However, at the same time just using the new sample without considering cached values can lead to disturbing temporal noise, especially because of the human eye's high peripheral flickering sensitivity (Section 3.1.7). Hence, samples are usually temporally combined using a *running estimate*. Some considerations on such a running estimate are presented in Appendix A.3.

In contrast to the state-of-the-art, not each pixel is sampled in every frame but the sampling rates are adapted based on visual acuity. Hence, samples might be reprojected multiple times before they are updated with a newly computed sample. Relying on reprojected values only, most probably decrease the image quality. Hence, a smooth temporal blending process taking into account a limit of the samples' age and its last update is applied. While such a process reduces temporal flickering, large-scale contrast for the visual periphery is preserved. This leads to more stable images in the visual peripheral.

However, to begin with, not all samples are subject of temporal integration. A sample's color is directly written to the output if it belongs to the foveal region, is otherwise part of the resampling process (marked in the *resampling info*), or is written to a part of the image that did not contain any reprojected color due to disocclusion or movement. Also, this pipeline stage does extract the *support image* and *support G-buffer* without considering temporal integration (Figure 74). For all other samples, temporal integration might be employed. Following the renowned considerations by Yang et al. [Yan+09] for Temporal Anti-Aliasing (TAA), possible (dis-)occlusions caused by a change in perspective are determined by looking at the depth difference between $c_{t-1}[p]$ and $s_t[p]$. Here, $c_t[p]$ does refer to the cache value at pixel p and frame t, while $s_t[p]$ is the newly computed sample for pixel p and time t. If the depth difference of those samples is above a threshold ϵ, the reprojection contains a potential occluder or parts have become disoccluded, as the ray has hit a part of the scene different from the cache. In this case, only the newly generated sample are considered without integrating anything from the cache. If the depth difference is below the threshold, the cached color values at pixel p can be combined with the new sample s. The final pixel value $f_t[p]$ is now computed using a blending value α_t[p]:

$$f_t[p] \leftarrow \alpha_t[p] \cdot s_t[p] + (1 - \alpha_t[p]) \cdot s_{t-1}[p] \tag{5}$$

Finally, to account for the sample's age and limited update rates, $\alpha_t[p]$ has to be adjusted according to the number of samples accumulated at pixel position p as well as the most recent update-time of this pixel. To do this, $\alpha'_t[p]$ is computed as:

$$\alpha'_t[p] \leftarrow \frac{1}{N_{t-1}[p]+1} \qquad N_t[p] = \left(\alpha_t[p]^2 + \frac{(1-\alpha_t[p])^2}{N_{t-1}[p]}\right)^{-1},$$

where $N_t[p]$ represents the number of samples that have been accumulated at pixel p and frame t. This process resembles the method of Yang et al. (Appendix A.3). To avoid infinite accumulation of samples, k is the minimum possible weight for the new sample to finally compute $\alpha_t[p] \leftarrow \min\{\alpha_t[p]', k\}$. However, it is proposed that it is best to adapt k dynamically based on a sample's age.

If a pixel has been sampled a couple of frames ago, it has undergone the potentially imprecise reprojection process multiple times, especially since the camera is constantly moving when head tracking is active. If the timespan between the previous update and the current time is high, it is better to account for the current sample with a higher weight. Therefore, instead of a fixed k, the following exponential function is used.

$$k_t(\Delta t) \leftarrow \min\left\{\exp\left(x_0 + \frac{\Delta t - 1}{t_{max} - 1}(x_1 - x_0)\right), k_{max}\right\}, \quad x_0 = \ln k_{min} \text{ and } x_1 = \ln k_{max}$$

This function can be parameterized based on a fixed interval $[k_{min}, k_{max}]$ and a maximum timespan t_{max} for which it accumulates samples. The value $\Delta t = t - t_{touched}$ is the difference of the current frame index and the frame index a value has last been touched and updated in the cache. Here, t_{max} is the user-specified maximum number of frames between two samples. Computing k this way, choices have to be made. The value t_{max} should be chosen according to the refresh rate and in a way that resolves possible artifacts as early as possible by giving the new sample a higher weight. At the same time, weighting older samples relatively high guarantees a smooth temporal transition and reduces flickering. An additional *cache information* buffer is used to keep track of $\alpha_t[p]$, $N_t[p]$ and $t_{touched}$. These are stored per pixel along with the *color image*. However, to have the estimates available in the next frame it is necessary to reproject this buffer to the new perspective the same way as it is performed for the color values described in Section 6.1.2. The image now contains all newly computed samples and the reprojected samples from the last frame. Still, pixel values might be missing. This is the case for parts of the image that have not been in the view frustum for frame $t-1$ and have not been updated by a newly computed sample. Eventually, a separate *merge* kernel (Figure 74, Block 3) is launched to fill in all missing pixels with samples obtained from the low-resolution but completely updated *support image*.

6.1.5 *Post-Processing*

As images are updated with varying rates due to the use of the visual acuity model that sparsely samples the image plane, stochastic sampling processes might lead to varying convergence behaviors in the scene. These artifacts are likely to be observed by a user. Hence, a post-processing filter (Figure 74, Block 4) is proposed that does not only integrate samples temporally but provides a path to integrate irradiance values spatially. For each pixel q in a region that did not integrate information temporally, the nearest reconstructed (i.e., non-resampled) neighbor along the horizontal and vertical axis on the image plane is searched. The

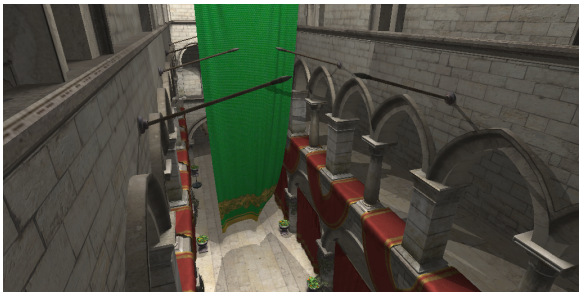
distance to this neighbor is then used to create a search window which is randomly sampled n times. This process selects the closest reconstructed pixel r found during the sampling step and applies this pixel's irradiance to the noisy pixel q. However, as the proposed system is not yet fast enough to account for area lights, ambient occlusion or even Global Illumination (GI) in HMDs – processes that need stochastic sampling – and the filter is not the authors original work, no further discussion is presented here. However, more details of this filter are provided in the previously published paper [Wei+16].

6.2 BENCHMARKS

The hardware configuration for the performance benchmarks consisted of an Intel Core i7-3820 CPU, 64 GiB of RAM and an NVIDIA GeForce Titan X driving an Oculus Rift DK2. Using the Oculus SDK, the FoV was determined for a single eye and in turn used to compute the projection matrix. The rendering resolution was set to 1182×1464 per eye. Table 7 lists the benchmark results of fly-throughs with 1000 frames each. It was decided to use the parameters ($r_0 = 10°, r_1 = 20°, p_{min} = 0.05$) to configure the user's foveal region. As shown in Section 6.3, users were mostly unable to detect any visual differences in images from the full renderer for this FRC. For the benchmark process, the foveal region was statically positioned at the image center. A resolution of 256×318 was selected for the *support image* and *support G-buffer*. This resolution was chosen empirically as it provided a good balance between speed and quality for the used HMD and scenes. The four following test scenes were selected: *Sibenik*, *Sponza*, *Rungholt*, *Urban Sprawl* (see Figure 76). Each of the scenes was rendered with one point light source, an area light source with 8 sample-per-pixel (spp), and ambient occlusion using 16 spp.

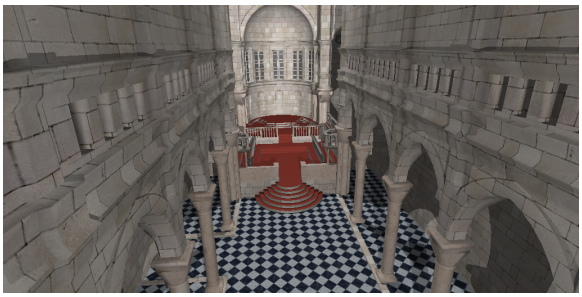
Table 7 shows the speed-up of the foveated ray tracer compared to a full ray tracer. It scales well with increasing ray workloads, as this reduces the number of rays. Hence, the smallest speed-up of 1.46 is achieved when rendering the scene *Urban Sprawl* with a single point light, while the maximum speed-up of 4.18 is achieved for *Sponza* with ambient occlusion. This table also gives the time required for reprojection, cache update, merging, and the optional post-processing step. These run-times are nearly constant for all scenarios, as they are mainly dependent on the rendering and *support buffer* resolutions. The coarse reprojection and resampling can theoretically replace the asynchronous time warp functions performed by the Oculus SDK, yet they only add a minimum latency to the overall rendering process due to their asynchronous nature (Section 4.3.2). The influence of the FRC on the rendering performance measured in Frames-per-Second (FPS) for *Sponza* is illustrated in Figure 77. For a higher number of rays with a large FRC, the coherency of rays increases. Thus, it becomes clear that rendering performance does not decrease linearly with an increase in the number of rays traced per frame.

Even though rasterization is inherently different from ray tracing, several benchmark results can be provided in order to make a comparison to other state-of-the-art approaches. In "Foveated 3D graphics" by [Gue+12] the image is rasterized in three layers with different resolutions. This yields a speed-up of 6.2 with only 7% of the pixels being rendered as the images are strongly undersampled. To still achieve an acceptable visual quality, this approach needs to rely on specific Anti-Aliasing (AA) methods, which in turn limits its applicability [Ste+16]. Moreover, numbers are only reported for a single scene. By using NVIDIAs Multi-

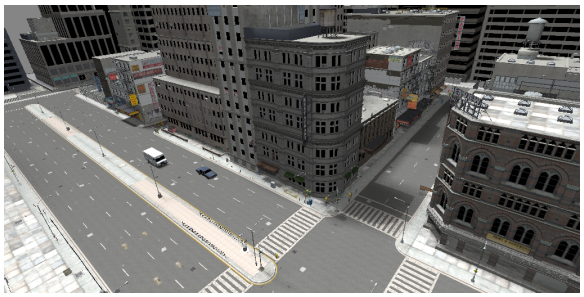


(a) Sponza

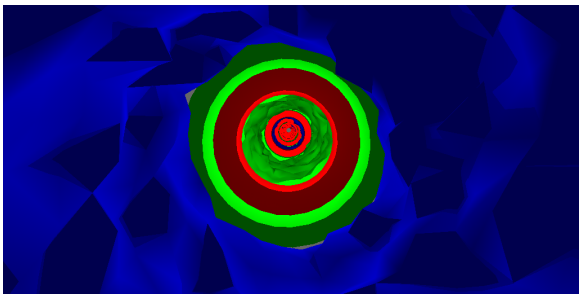
(b) Rungholt



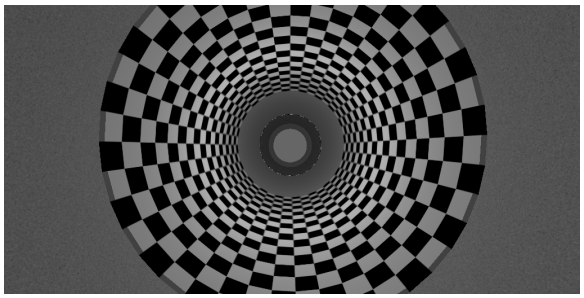
(c) Sibenik



(d) Urban Sprawl



(e) Tunnel_Geom



(f) Tunnel_Maps

Figure 76: Scenes used for benchmarks and user studies of the implementations.

Resolution Shading [Ree15] that does not allow for a foveated adaptation but adaptive shading to match lens distorions of HMDs, a speed-up of 1.3 to 2 is reported depending on the configuration. Stengel et al. [Ste+16] report a speed-up of 1.34 on average, with the number of shaded pixels being decreased by 65% for a resolution of 1280×1440 pixels and 83% for twice as many pixels. Our method has shown a reduction of sampled pixels by 79% on average for all benchmark scenes, with an average speed-up of 2.55. The ray-based approach by Fujita and Harada [FH14] shows similar frame rates compared to the approach presented in this thesis, even though they use different and more GPUs rendering at a lower resolution. The performance of the method here could be further improved by generating rays that directly match the image distortion of the HMD, making it possible to achieve even higher resolutions while maintaining refresh rates.

	Scene Type	# of rays		Re-	Ray Tracing		Cache	Merge	Post	Total Time		Speed-up
		full	ours	project	full	ours	Update			full	ours	Factor
Sibenik	point light	3.46M	0.81M	1.41	10.74	3.85	0.64	0.84	0.00	10.74	6.74	1.59
75K Triangles	area light (8 spp)	15.57M	3.64M	1.51	47.96	14.80	0.67	0.85	0.33	47.96	18.36	2.61
Figure. Figure 76c	ao (16 spp)	29.42M	6.88M	1.45	94.64	27.96	0.64	0.83	0.32	94.64	31.39	3.02
Sponza	point light	3.46M	0.64M	1.41	13.93	4.43	0.58	0.84	0.00	13.93	7.26	1.92
154K Triangles	area light (8 spp)	15.54M	2.89M	1.44	81.61	20.07	0.58	0.83	0.24	81.61	23.17	3.52
Figure Figure 76a	ao (16 spp)	29.36M	5.45M	1.43	179.01	39.79	0.58	0.83	0.24	179.01	42.87	4.18
Rungholt	point light	2.90M	0.58M	1.38	9.51	3.54	0.59	0.83	0.00	9.51	6.34	1.50
6704K Triangles	area light (8 spp)	11.10M	2.25M	1.38	34.20	9.98	0.59	0.83	0.19	34.20	12.97	2.64
Figure Figure 76b	ao (16 spp)	20.47M	4.17M	1.39	168.03	51.15	0.59	0.83	0.19	168.03	54.14	3.10
Urban Sprawl	point light	3.06M	0.64M	1.37	8.97	3.33	0.60	0.83	0.00	8.97	6.12	1.46
773K Triangles	area light (8 spp)	12.34M	2.60M	1.36	29.67	8.68	0.60	0.83	0.28	29.67	11.75	2.52
Figure Figure 76d	ao (16 spp)	22.95M	4.85M	1.39	110.11	30.02	0.60	0.83	0.29	110.11	33.13	3.32

Table 7: Times in ms for each stage of the pipeline in comparison to a full renderer showing the speed-up of our approach. Times and speed-ups are computed for a medium sized foveal region with ($r_0 = 10°$, $r_1 = 20°$, $p_{min} = 0.05$) for a single eye with a resolution of 1182×1464 and no oversampling on an NVIDIA GeForce Titan X. For the chosen foveal region, users were mostly unable to detect any visual difference to full rendering in the user study.

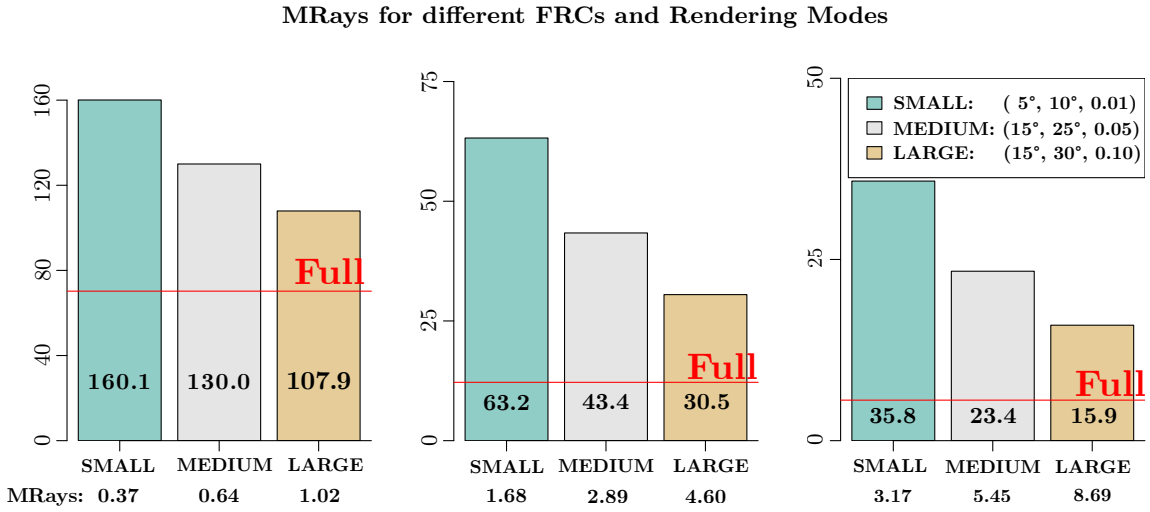


Figure 77: Influence of configuration of the foveal regions on FPS for the scene Sponza, (a) rendered with a point light source, (b) an area light source with 8 spp and (c) with ambient occlusion with 16 spp. MRays denotes the mean number of megarays per frame. The scene was rendered at a resolution of 1182×1464 using an NVIDIA GeForce Titan X.

6.3 USER STUDY

The user study addressed the perceived visual quality of the presented method. It was driven by the following research questions:

- **RQ1:** Can subjects differentiate between scenes with varying graphical content, rendered with and without the foveated rendering method?
- **RQ2:** Do modifications of the foveal region parameters in the ray generation have an effect on the perceived visual quality?
- **RQ3:** Does the fixation type have an effect on the perceived visual quality?

6.3.1 *Procedure and Apparatus*

The setup used for the user study differed from the benchmark configuration. It comprised an Oculus Rift DK2 (SDK 0.8) on a Windows 10 system including an Intel Xeon E5-2609 (2.4 GHz), and 64 GiB of RAM. The DK2's native refresh rate of 75 Hz was used as the baseline. As both the foveated rendering and the OpenGL-based reprojection process had to be parallelized in order to achieve this refresh rate, it was necessary to deploy two Quadro K6000 cards. These were additionally required due to the unavailability of a Linux driver for the utilized eye tracker. This in turn ment only Windows could be used, which does not allow for multi-GPU rendering on NVIDIA's consumer cards. The Oculus was equipped with an SMI binocular eye tracker running at 60 Hz (asynchronous). This eye tracking refresh rate can be considered a lower bound for foveated rendering (Appendix A.2).

The experiment was conducted as a within-subject study, employing a $4 \times 4 \times 3$ full factorial design. Each participant completed 96 trials in a randomized order. These trials were the result of a full factorial combination of four scenes {*Sponza*, *Tunnel_geom*, *Tunnel_maps*, *Rungholt*} presented in Figure 76, four FRCs {small, medium, large, full} and three fixation types {fixed, moving, free}, described below, as well as two repetitions. Full ray tracing was included as the FRC *full*, representing the control group. Each trial consisted of an 8-second-flight with a one-factor combination. With some minor optimizations, frame rates of at least 75 Hz were achieved for all scenes, including the final image warping for display in the Oculus. However, all sequences used for full ray tracing had to be pre-recorded (excluding any optimization like reprojection), then loaded at runtime, and optionally augmented with the specific trial fixations in order to be displayed at 75 Hz. In the following, the chosen factors are detailed and it is explained why they were selected to answer RQ1-3.

RQ1: DIFFERENTIATION BETWEEN FOVEATED AND NON-FOVEATED RENDERING. In order to provide answers to RQ1, the test scenes were varied to study the effect of graphical contexts on the noticed visual artifacts. This way, more reliable statements can be made when determining if there is a perceivable difference between foveated and full rendering. While *Sponza* represents the most real world-like scene with only a few pronounced discontinuities (and thus hard edges) usually visible, the scene also contains some smoother curves resembling real objects. *Rungholt*, a scene generated from a Minecraft map, has many visible depth discontinuities, which are both challenging for perception and reprojection methods. The artificial test scene *Tunnel_geom* contains a tunnel consisting of noisy, displaced geometry. Depending on the point of view, this scene can contain both hard edges and smooth, continuous surfaces. *Tunnel_maps* is a tunnel textured with a checkerboard map and a noise texture. Both scenes were designed to contain challenging discontinuities, either due to a great variance in depths or in contrasts.

RQ2: EFFECT OF FOVEAL REGION SCALES. The following eccentricity thresholds and minimum sampling probabilities were selected to test the influence of the chosen FRCs: *small* ($r_0 = 5°, r_1 = 10°, p_{min} = 0.01$), *medium* ($10°, 20°, 0.05$), *large* ($15°, 30°, 0.1$) and *full* ($\infty, \infty, 1$). FRCs were determined by using the angular size of the fovea for r_0 with a steep fall-off for the smallest setting and increasing the foveal region and minimum sampling probability while reducing steepness for the other settings. The smallest FRC was expected to yield visible artifacts for most participants, as the foveal region used for rendering barely matches the extents of the *fovea centralis*. The *medium* and *large* FRC extended the foveal region to include the *parafovea* and *perifovea*, respectively (Section 2.1.1).

RQ3: EFFECT OF FIXATION TYPES. Finally, while eye tracking determines a user's focal point in the scene (defining the foveal region), fixation may affect visual attention, potentially leading to visual tunneling (Section 2.2.5). In order to determine a potential influence, fixation types were varied to trigger different levels of visual attention. The *fixed focus* mode contained a static fixation cross at the image center to be focused for the entire trial. For the *moving target* mode, a set of paths across the image plane was generated. Here a green sphere as fixation target that moved along the paths was displayed. The velocity of this movement was static and did not exceed $18°/s$. Also, in order to provide a more natural experience and allow users to focus on this target, it was adapted to the scene depth as

it was directly located on each underlying surface. While identical paths were selected for the repetitions, paths were varied in all trials. This was deemed necessary to avoid learning effects and to have a better spread of potential fixation locations. In order to avoid a negative influence of the eye tracker's inaccuracies (relatively low refresh rate, inaccurate tracking towards outer display regions), the foveal region for *fixed focus* and *moving target* was always centered around the fixation target. Here, the *moving target* fixation mode was expected to cause higher visual tunneling as the user had to concentrate on following the target. Only those trials with *free focus fixation* enabled the user to look around freely with the foveal region following the user's gaze.

After signing an informed consent and receiving instructions, participants were seated and equipped with the HMD. Prior to the main experiment, six test trials of a flight through *Sponza* were shown, including the smallest FRC, full rendering, and all fixation types. In order to avoid learning effects, the flight through *Sponza* in this introductory part differed from the one used in the actual study. Still, the introductory part allowed participants to familiarize themselves with the range of configurations and visible artifacts. After each main trial, the participants were presented with the following statements

- **Q1:** The sequence shown was free of visual artifacts
- **Q2:** I was confident in giving this answer

to be rated on a 7-point Likert scale from *strongly disagree* (-3) to *strongly agree* (3). Eye tracking data was recorded during all the trials to enable a comparison to be drawn of the measured PoRs and the fixation paths as well as to provide more insights into tracking accuracy.

6.3.2 *Results*

15 subjects participated in the user study (10 male/5 female, all with academic background), aged between 26 and 51 ($M = 33$, $SD = 7.24$). All reported to have normal or corrected-to-normal vision ($< \pm 1\,\mathrm{D}$) and no other known serious visual impairments. Eight participants ($53, \overline{3}\%$) played computer games on a regular basis.

The next paragraph presents the data that concerns the visible artifacts of the proposed method. This will allow insights into the *visual quality* (responses to Q1 and Q2) of the rendering method to be provided. Afterwards, the following paragraphs present the quality of the recorded eye tracking data to show its validity and the link between the mental workload and the subjectively perceived image quality.

RESPONSES TO Q1 AND Q2 The responses of the Likert-ratings on the presence of visual artifacts (Q1) can be found in Figure 78 and Figure 79. A multifactor Analysis Of Variance (ANOVA) [FH03] was performed on the data (1440 trials) by Thorsten Roth, presented in our previously published paper [Wei+16]. ANOVA contrasts were configured to always compare to both the full-sized FRC and the factor mean values for the *scene* and *fixation type*. Significant interactions and the observed main effects were analyzed with post-hoc t-tests using Holm's method for p-value adjustment. Confidence values (Q2) were mostly high

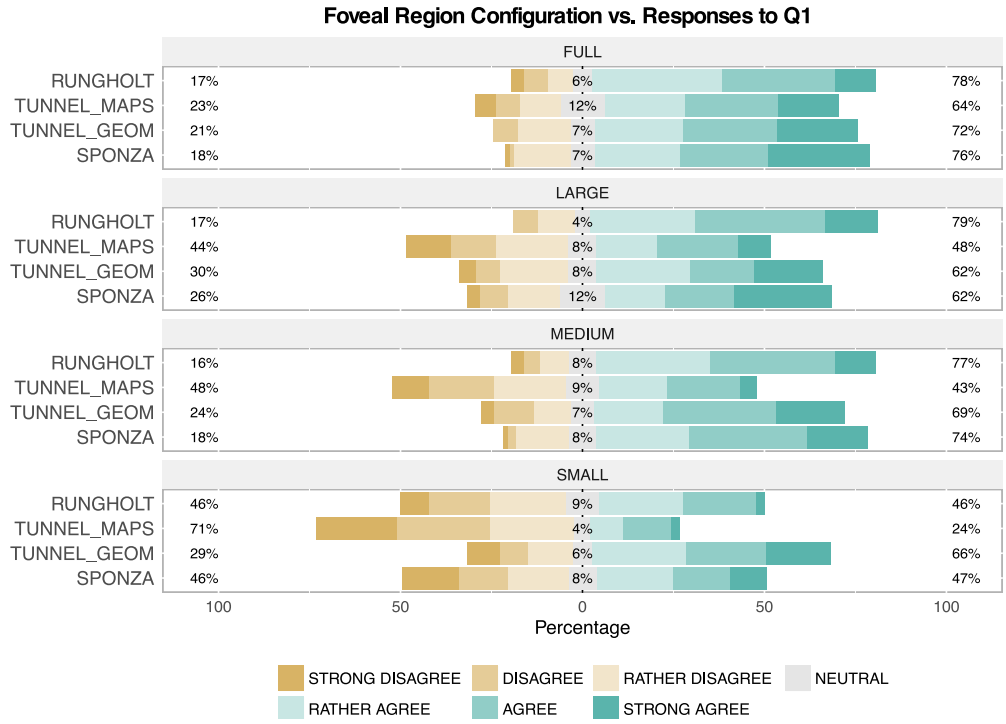


Figure 78: Likert-scale ratings for Q1 (*Was the sequence shown free of visual artifacts?*) for all scenes grouped according to foveal region configurations. The percentages on the left and right represent the fraction of all participants who had a tendency towards *disagree* and *agree*, respectively. The smallest foveal region configuration revealed a significant number of artifacts, while the larger foveal regions were close to the full renderer in this regard.

($M = 1.62$, $SD = 1.14$) with negligible differences. Consequently, it was not deemed necessary to further consider the confidence scores in this analysis.

In Figure 80 the average Q1 scores can be seen, grouped according to the different *fixation types* and FRCs. On average the visible artifacts were rated to be less perceivable for all scenes when a *moving target* was presented. To shed some light on the influence of the actual visual quality, Figure 81 illustrates the data for the individual scenes. These plots present all three fixation modes with respect to all FRCs up to full rendering, including red lines to show the means for each of the fixation modes. A discussion of these results will be presented in the next section.

RECORDED EYE TRACKING DATA One essential property of the eye tracker is its precision. To get an idea of the tracking accuracy, the distance between the recorded PoR and the actual location of the *fixed focus* and *moving target* that people were instructed to fix their vision on, can be determined. It is noteable here that all distances are average values of the left and the right eye in order to achieve a more compact analysis. The preceeding experiments to determine usefull parameters ranges for this user study, made quite noticeable eccentricity-dependent tracking inaccuracies evident. It became obvious, that the precision degraded towards outer image areas. Hence, a discussion on tracking accuracies is a necessity in order to put results of this user study into perspective.

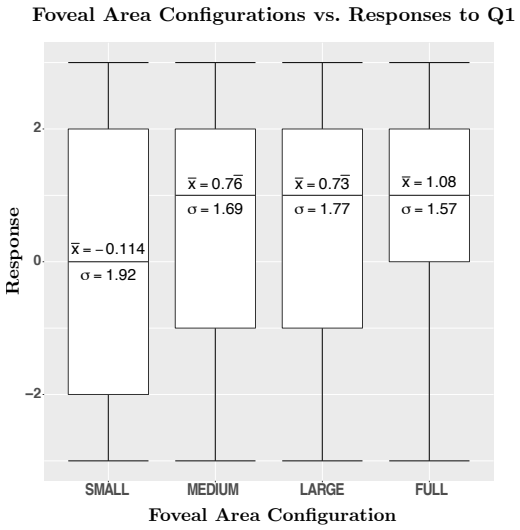


Figure 79: Likert-scale ratings of perceived visual artifacts for different foveal region configurations. Ratings for *medium* and *large* are not significantly different from *full* rendering.

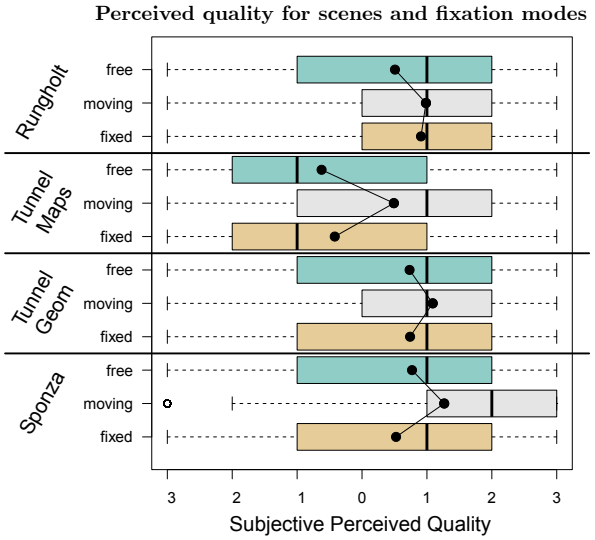


Figure 80: Quality for all combinations of scenes and fixation modes. Quality ratings were the highest for all scenes in the *moving target* mode, although the fixation accuracy was worse than for the *fixed target*.

In order to prove the initial observations, the deviations of the recorded PoR from the fixation target's current position were investigated for all trials showing a *moving target*. The analysis is based on the assumption that, if there are no eccentricity-dependent tracking inaccuracies, the deviations should be to a great extent independent of the target's position in the image. If this is the case, an increase of deviations towards outer image areas is most likely a result of tracking inaccuracies. To this end, as presented in the paper by Roth et al. [Rot+17], an analysis by my co-authors was performed as follows: To show a potential relationship between eccentricity and accuracy, data is sorted into eccentricity-dependent bins. After this, the means are computed for each bin. To achieve a high resolution, bins were determined as having a width of $w = 0.1°$. Each bin stores a tuple of $B_j = (\bar{F}_j, \bar{G}_j), 0 \leq j < n$, with

$$F_j = \{F_{p,t}(i) \,|\, j \cdot w \leq F_{p,t}(i) < (j+1) \cdot w\}, \tag{6}$$

$$G_j = \{G_{p,t}(i) \,|\, F_{p,t}(i) \in F_j\}. \tag{7}$$

Here, $F_{p,t}(i)$ is the distance between the fixation target's current position and the image center, while $G_{p,t}(i)$ represents the distance between the gaze and the fixation target in trial t at the frame i for the participant p. The chosen bin width results in a total of $n = \lceil \max(F_{p,t}(i))/w \rceil$ bins. Here, $\bar{G}_j$, i.e. the mean value for the respective bin j, provides an approximate tracking quality measure at a specific eccentricity. Finally, to analyze if the tracking precision relates to the actual eccentricity, second order polynomial regression with $\hat{G}_j = \beta_0 + \beta_1 F_j + \beta_2 F_j^2$ is performed. The quadratic prediction for gaze deviation is plotted in Figure 82. This plot clearly shows the dependency of tracking precision and eccentricity. As expected, the results showed statistically significance, yielding a correlation coefficient of $r = 0.989$ with $\beta = (1.05, 0.024, 0.008)$. The unitless correlation coefficient r ranges from -1 to 1. The closer the value is to one or the other, the better is the correlation of the fit.

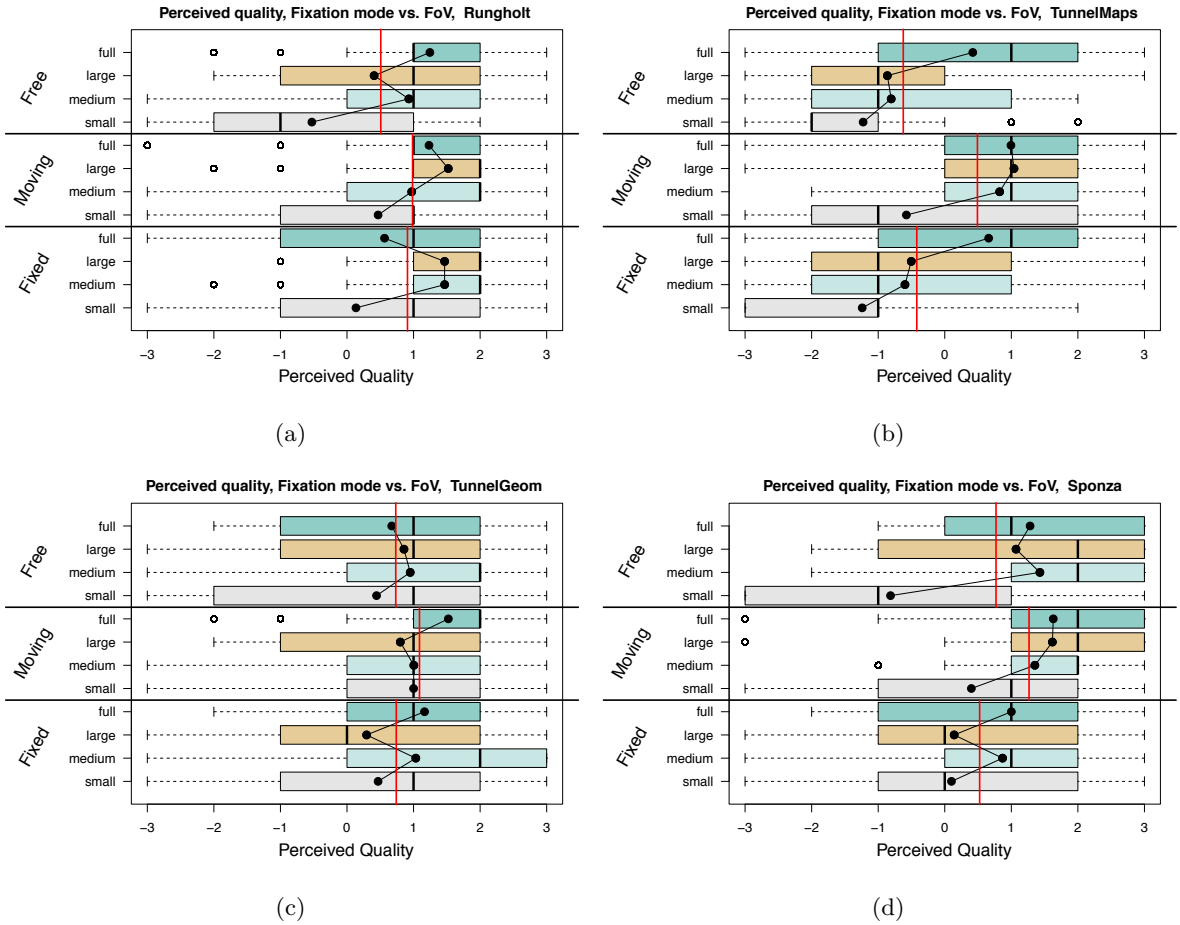


Figure 81: Quality ratings for fixation modes, Foveal Region Configurations and all scenes. The black dots inside the boxes represent the respective mean quality ratings.

The coefficient of determination is $R^2 = 0.978$ is another statistical measure of how well the regression predicts the measured samples. A value of one is reported for perfect fits. However, the high quality of the fit of the mean deviations between the tracking targets and the PoRs is also clearly visible in Figure 82.

Besides the decreasing tracking inaccuracy, it is interesting to investigate where the PoR in the image was located in relationship to the *moving target* and the fixation cross in the *fixed focus* mode. A heatmap showing the deviations for these two factors and for the test scenes is presented in Figure 83. In addition, an alternative way to look at the deviations is by computing Cumulative Distribution Functions (CDFs). The results of these computations are presented in Figure 84 for the *fixed focus* and the *moving* fixation target for all four scenes. In these plots the horizontal axes represents the respective angular distances between the participant's gaze and the fixation target. The vertical axes represent the probability of having an equal or smaller angular deviation.

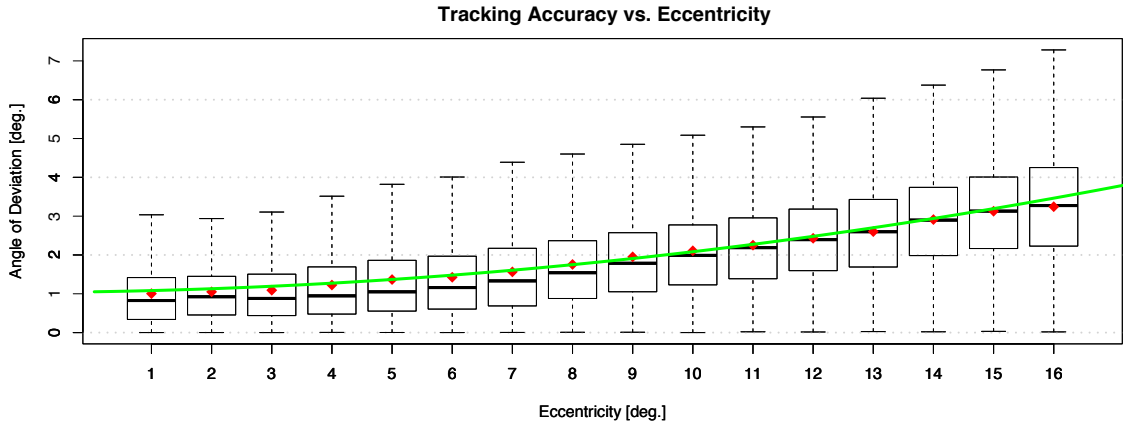


Figure 82: Tracking precision vs. fixation target's distance to the image center. The green line represents the result of linear regression with a quadratic equation. The red points show the binned means.

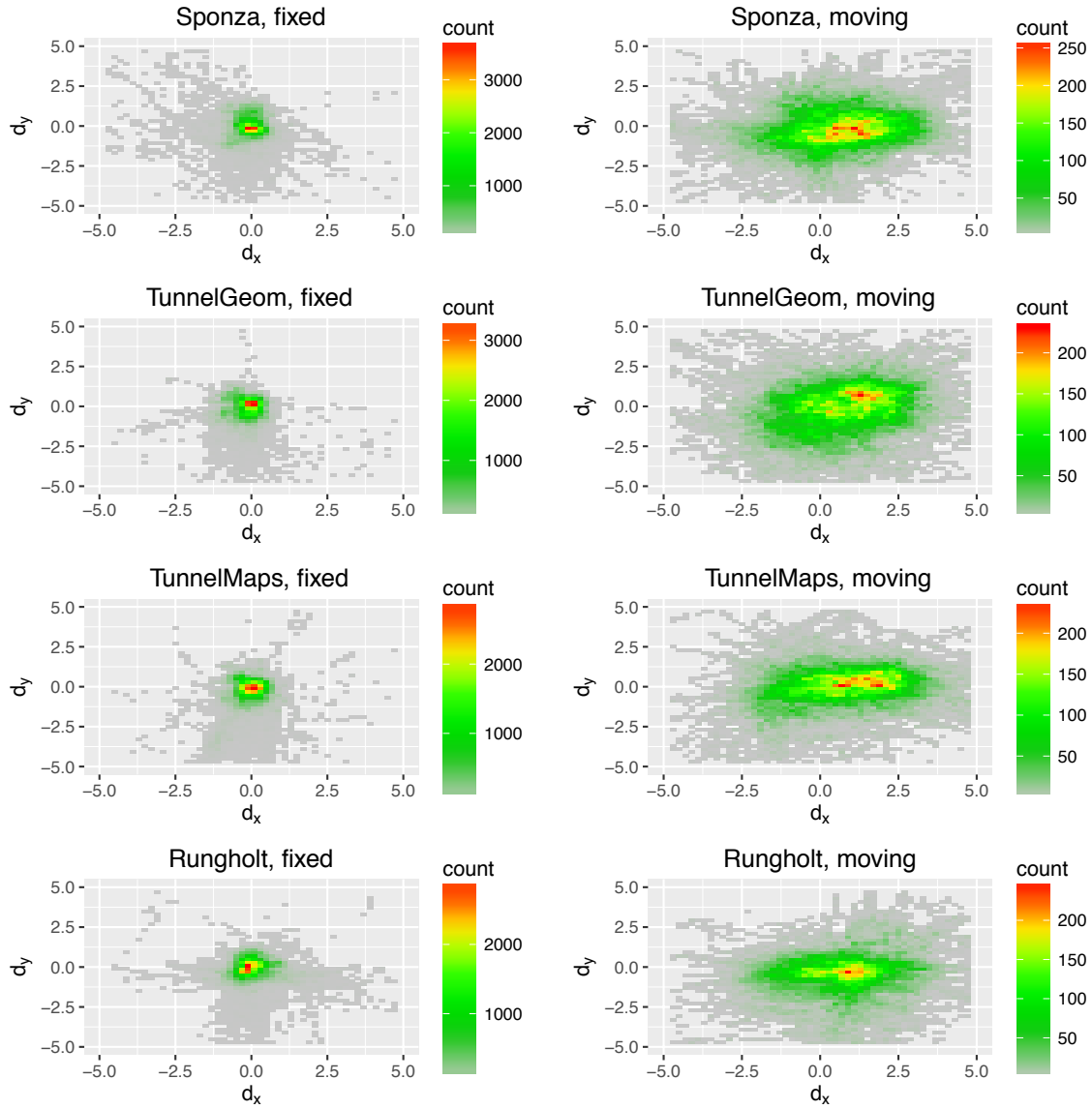


Figure 83: Gaze deviation for all individual scenes, *fixed* and *moving targets*. *Image from Roth et al. [Rot+17]*

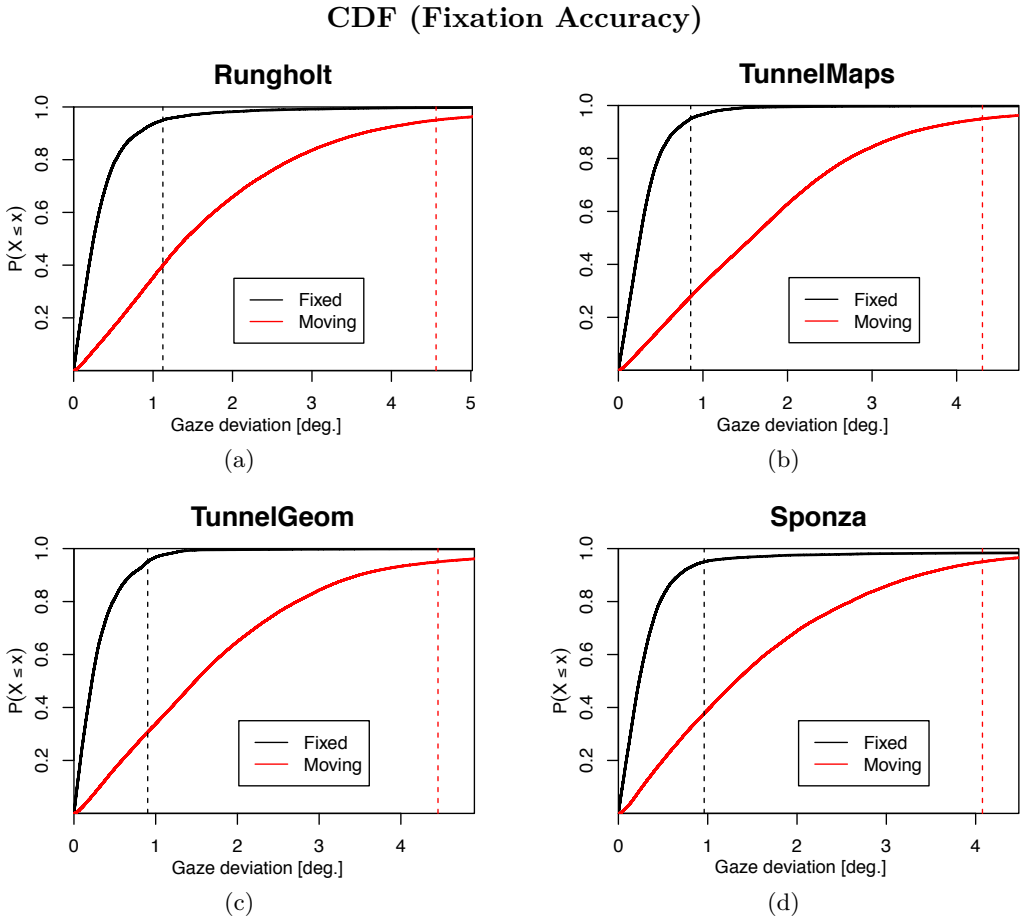


Figure 84: Cumulative distribution functions (CDFs) of the measured fixation accuracy for *fixed* and *moving targets*. The 95% quantiles of gaze deviations for each scene are illustrated with dotted lines. There are significant differences between the fixation accuracy for the *fixed* and the *moving fixation targets*. X is the actual gaze deviation. *Image from Roth et al. [Rot+17]*

6.3.3 *Discussion*

This section provides answers to RQ1-RQ3, shows an evaluation of the tracking data as well as potential interactions to subjective quality measurements. As illustrated in Figure 82, the precision of the utilized eye tracking device decreases with increasing eccentricities. Adults can physically rotate the eye up to 50° horizontally, 42° upwards and 48° downwards away from the line of sight in the eye's resting position [Adl+11]. In practice, however, humans usually do not rotate the eye to the physiologically possible maximum. After a certain angular deviation, a human would most probably begin to turn his head. As detailed in Section 2.2.3, this Comfortable Viewing Angle (CVA) is considered to be $\approx 15°$ around the normal line of sight. Thus, it is important to note that fixation target eccentricities larger than the CVA in the precision measurements were not accounted for. In this user study, head tracking was not activated in order to present identical visual stimuli to all participants. This would not have been possible if users had been able to look around freely. However, for fixation target positions further away from the image center than the CVA, users would probably not just rely on eye movement to fixate a target, but instead, incorporate head movement.

The solution for this in the user study has been to limit the area concerned to only include fixation targets up to the CVA. This issue can also be approached in a different way. The calibration step of the eye tracker could be investigated more thoroughly. Interestingly, the built-in 9-point calibration procedure and the respective fixation patterns of the SMI-based eye tracker covered only a small inner subset of the FoV.

As presented in Section 7.3.1 other eye trackers show significantly different accuracies when evaluated using the approach presented here. However, even though Figure 82 suggests, that the eye tracker is fairly accurate (with a mean error of 1° to 3° depending on the eccentricity), the result of this type of tracking precision analysis should not be interpreted as a direct measure for tracking precision. Latency-based deviations, saccadic or Smooth Pursuit Eye Motion (SPEM) movements, as well as other possible disturbances, have not been filtered from the data. The actual behavior of the measured gaze deviation, however, yields a reasonable estimate of the eccentricity-dependent precision fall-off.

Closer study of Figure 84 reveals that there is a significant difference between the fixation accuracy for the *fixed target* (below 1.1°) and the *moving target* (approx. 4° to 4.5° for all scenes). By further analyzing the paths using Figure 83 as performed by Roth et al. [Rot+17], revealed that the fixation target moved left more often than right. Initially, this apparent shift to the right for the gaze deviation was explained by the utilized fixation paths not being equally distributed regarding the fixation target's movement. Another explanation is that the fixation accuracy between the *moving target* and the fixation cross differs due to the SPEM. The movement of the target was not predictable for the participants. This unpredictability naturally leads to a reduced SPEM precision. Moreover, accuracy is reduced because the background is at the same distance from the eyes as the target to be followed by the gaze. Also, head tracking was inactive, as prerecorded camera paths were presented. Thus, other signals, for example by the vestibular system, could not be used by the Human Visual System (HVS) to distinguish between target and background [Adl+11, p. 229].

Nonetheless, the presented data shows the validity of the acquired eye tracking data. Even at large eccentricities, the average gaze deviation is well below the selected FRCs. Also, exceeding the CVA, the paths for trials showing a *moving target* were limited to always stay within the innermost $\approx 50°$ of the FoV. Taking into account the aforementioned considerations, a further analysis of the initial research questions RQ1-3 based on the user study as well as a discussion of the results can be found below.

RQ1: DIFFERENTIATION BETWEEN FOVEATED AND NON-FOVEATED RENDERING. By looking at the results presented in Figure 79, generally subjects cannot reliably differentiate between full and foveated rendering. This is the case for foveal regions not smaller than approx. 10°. However, looking at Figure 78, where responses are grouped by scenes for varying FRCs, *Tunnel_Maps* shows a decrease in quality. Scenes with too much high-frequency contrasts appear to be problematic. Nonetheless, what should be apparent now and revealed by statistical data is that differentiation depends significantly on all the test variables: the size of the FRC, the fixation mode as well as the displayed scene. While the factor FRC shows a significant main effect ($F \approx 30.54$, $p \approx 0$), there is a strong interaction between FRC and the scene ($F \approx 3.09$, $p < 0.005$). Interestingly, the performed t-tests showed that significant differences between the *medium*, *large*, and *full* FRC were only present in the scene *Tunnel_Maps*. All other scenes only showed significant differences when the *small* FRC was

involved. First, this was mainly attributed to the regular, rather extremely high-contrast checkerboard pattern in *Tunnel_Maps*. However, as the eye tracking system is susceptible to inaccuracies in outer image regions, the logged data was filtered for analyzing the tracking information in order to include only the region utilized by SMI's calibration method. This region extends to maximum eccentricities of approx. 10.3° left/right and 11.68° up/down. These numbers were taken from the SMI SDK's 9-point calibration method and converted to angles. Angular differences were analyzed between the fixation point and the tracked gaze. Participants stayed closer to the fixation point for the *fixed mode* ($M = 0.31°$, $SD = 0.4°$) than for the *moving target* ($M = 1.9°$, $SD = 1.52°$). Bearing in mind that *Tunnel_Maps* had the greatest number of visible artifacts for the participants, it is important to mention that the median angular differences for *Sponza*, *Tunnel_Geom*, and *Rungholt* were between 1.25° and 1.58°, while for *Tunnel_Maps* a median difference of 2.24° was present. As this larger distance to the foveal region's center indicates that the gaze was closer to sparsely sampled regions, this offers another explanation for the relatively low Likert ratings for this scene.

RQ2: EFFECT OF FOVEAL REGION SCALES. The previous paragraph already showed that if the FRC is not too small, subjects will hardly notice visual artifacts using foveated rendering. Studying Figure 79 once more, where the responses to Q1 for varying FRCs are plotted including the mean values and standard deviations, the small FRC scored significantly lower. Medium and large FRCs were almost identical as regards perceived visual artifacts. The difference to full rendering was limited to a larger standard deviation. As Figure 78 illustrates, this can again be mainly attributed to the artifacts visible in *Tunnel_Maps*. In addition, there were no significant differences between angular deviations for varying FRCs for that scene. The median values over all scenes for the four FRCs were all within $[1.62°, 1.7°]$ for the *moving target* and $[0.22°, 0.25°]$ for the *fixed mode*.

RQ3: EFFECT OF FIXATION TYPES. Fixation types, associated with different levels of visual attention, had a significant main effect ($F = 3.46$, $p = 0.03$) on the perceived visual quality. While *free* ($M = 0.43, SD = 1.89$) and *fixed* ($M = 0.43, SD = 1.81$) modes showed nearly identical responses, the *moving target* was rated significantly better ($M = 0.99, SD = 1.63$). Thus, fewer visual artifacts were noticed with the presumed higher visual attention of the *moving target*, as subjects were more probably less aware of details outside the focus area. This is remarkable as it could further reduce the sampling rate outside the foveal region. Furthermore, the foveal region matched the gaze when the target was perfectly followed. Figure 80 shows that the average Q1 scores are highest for the *moving target* for all scenes. In addition, as shown in Figure 81, the visual quality for the individual scenes in dependency to the FRC was consistently superior with the *moving target* condition. However, it also becomes apparent that the increase in rendering quality between the medium and the large FRC did not result in a consistent improvement of subjectively perceived quality. Differences from a medium FRC up to full rendering are mostly negligible for the moving fixation target. Interestingly, in some cases, even a larger FRC results in lower subjective perceived quality.

While the dependency of the fixation type to the visual quality can be estimated by looking at deviations of the real and measured PoR, analyzing eye tracking data for the *free viewing* task requires a different examination. Details on this matter can be found in the paper by

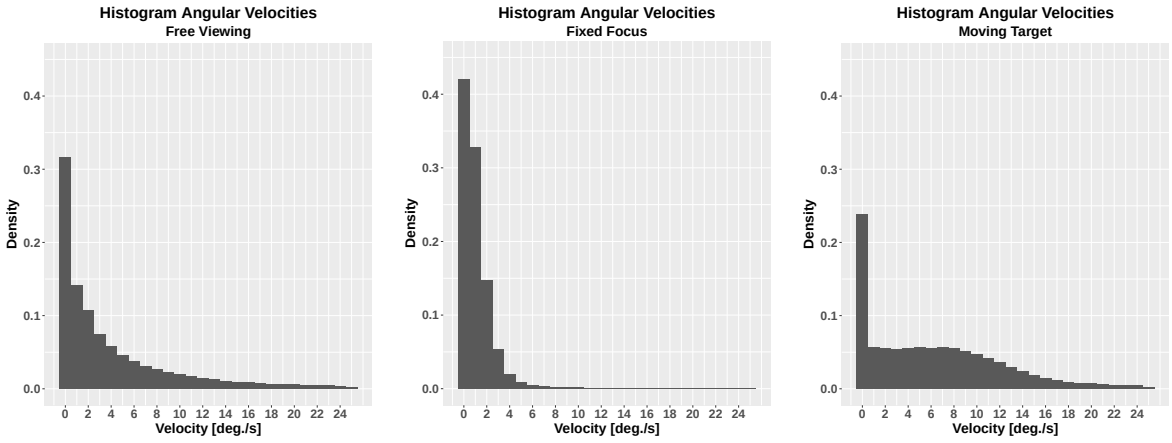
Roth et al. [Rot+17]. As the differences between the scenes proved too small to draw any further conclusions, they are not discussed here.

However, there might be another cause for the *moving target* mode being rated better that has not been considered so far, namely *retinal velocity*. While there is strong evidence that visual tunneling and mental workload play a significant role when judging the visual quality, the retinal velocity also greatly influences visual acuity (Section 3.1.2). In an initial step, the mean angular velocities were computed. These computations resulted in *free viewing* ($M = 22.20°/s$, $SD = 62.24$), *fixed focus* ($M = 1.26°/s$, $SD = 2.85$) and *moving target* ($M = 9.88°/s$, $SD = 24.69$). As expected the lowest mean angular velocities was present for the fixed focus mode. There is a striking difference here between *free viewing* and the *moving target*. The mean velocities for *free viewing* are more than twice as high as for the *moving target*. Hence, this is potentially an additional strong indicator for visual tunneling effects, as the visual quality when following a *moving target* was rated to be highest, despite the lower mean angular velocity. It can also be argued however that there is a difference in eye tracking behavior.

While following a *moving target*, the user's gaze has a much more constant velocity. Participants are exhibiting SPEM (Section 2.2.3). On the other hand, while being able to freely look around, users have much higher angular peak velocities but are able to spend more time inspecting details at lower angular velocities. The increase in peak angular velocities is observable by looking at the standard deviations. In order to show the difference in behaviors, a plot of the occurrence of different velocities for the first $50°/s$ is presented in Figure 85. As expected, the *free viewing* mode (Figure 85a) does contain a hyperbolically decreasing number of values with low peak angular velocities. Outliers at high angular velocities did, in fact, have an influence and resulted in a high mean. Also, by looking at the *fixed focus* results (Figure 85b) its low mean values become clear. However, using the *moving target* mode (Figure 85c) produces a striking difference in the results. Overall, a greater number of velocities for $> 4°/s$ is present. These velocities lead to a greatly reduced visual acuity (Figure 85c). Also, a much smaller number of low angular velocities $\leq 4°/s$ is present. The most obvious cause for this is that users were simply physiologically unable to detect artifacts while following the *moving target*. Hence, despite the visual tunneling, the reduced visual acuity at high retinal velocities and the change in eye tracking behavior might be another reason that the visual quality was rated highest for the *moving target* mode.

The user study revealed the perceived visual quality for even the moderately-sized FRC *medium* $(10°, 20°, 0.05)$ is almost identical to that of full rendering. Further improved outcomes are to be expected if a better and more accurate eye tracker is used. It is worth noting that the main causes for the low accuracy, besides the tracking precision itself, are tracking latency and possible unpredictabilities of the target's movement. However, the evaluation also showed that increasing the FRC did not always result in improved quality ratings. One possible explanation is that, eventually, all rendering methods can suffer from aliasing artifacts – in this case either with a finer and more regular or with a coarser pixel distribution. As reprojection methods hide visual artifacts, they may be able to conceal certain artifacts "more effectively" using the foveated mode presented.

While intuition may suggest a worse outcome for the *fixed targets* and the *moving targets* mode due to the high gaze-deviation as illustrated in Figure 83 and Figure 84, this is only



(a) (b) (c)

Figure 85: Histogram (Density) of peak angular velocities for the three focus modes *free viewing* (a), *fixed focus* (b) and *moving target* (c). While *free viewing* and *fixed focus* have a great number of values at low angular velocities, they are more evenly distributed when users are instructed to follow a *moving target*.

true for the *fixed target* mode. Figure 80 and Figure 81 reveal that the quality ratings for *moving target* fixation are better in all scenes tested. It can be argued that this difference in perceived quality between fixation modes and their counter-intuitive nature when taking tracking precision and temporal effects into account can be seen as evidence for the possible presence of visual tunneling effects. Another cause could be that the decrease of visual acuity is dependent on the retinal velocity. Either way visual artifacts that appear in the rendering system are effectively filtered by human perception.

There is a clear tendency towards negative ratings for the small FRC. In order to overcome this issue the foveal region can be enlarged with respect to the eye movement. This approach, however, poses a significant challenge. Increasing the rendering quality results in a performance hit, making it even more difficult to achieve the necessary refresh rate. As has been seen, the mean values for gaze velocities were highest for the *free viewing* mode. These high angular velocities pose a critical challenge when updating the image based on the PoR, despite the fact that some of them are filtered by *saccadic suppression* (Section 2.2.3).

6.4 FUTURE WORK

Although rasterization methods are faster on current GPU generations, ray tracing does provide a higher degree of flexibility. Accordingly, ray-based methods could become the first choice for performance critical real-time VR rendering in head-mounted devices [Hun15; Fri+16]. It would be interesting to evaluate how this system performs with moving objects, highly glossy materials, and dynamic light sources, also in the context of stochastic GI methods. In general, strongly view-dependent effects cannot be well captured with reprojection techniques. However, ray tracing has the advantage of being able to re-sample individual pixels efficiently. This could be used to include view-dependent effects in the resampling process. In the worst case however, this could in turn also lead to a fully sampled image,

bringing performance below the necessary refresh rates. It is planned to dynamically adapt the foveal fall-off and minimum sampling probability in the peripheral visual field to meet performance requirements consistently even for highly-complex scenes. Despite the fact that a choice has to be made, the author is convinced that users will more readily accept a few artifacts in the visual periphery than lower frame rates, especially in HMDs. This developed approach is an important step in the process of making realistic real-time ray tracing suitable for head-mounted devices.

Another exciting field for further research is *frameless rendering* (Section 4.3.1). The coarse reprojection is able to run in a separate thread, outputting images while at the same time matching and resolving reprojection errors to the display's refresh rate. Other threads asynchronously generate and merge new samples based on the user's gaze. If resolutions of displays continually increase or wireless transmission becomes available, we face a situation where a transmission of the computed pixels to the HMD is limited by the bandwidth of the interconnect. In this case, methods are preferable that enable images to be partially and asynchronously updated.

The accompanying user study and its evaluation also revealed the effects of an acuity loss based on the retinal velocity, visual tunneling, and mental workload. Thus, there are more circumstances which make it possible to reduce visual quality besides gaze. This is certainly the case in games, where events can be triggered that produce a change in the visuals, or task-driven environments. Task or navigation complexity may lead to high mental workloads. Moreover, certain events may allow hints to be derived about which part of the scene it is that attracts attention. Thus, visual quality can be reduced even further when attentional models are used in combination with gaze and saccade landing position predictions (Section 3.3.2). Another challenge yet to be solved is the issue of HMDs getting out of place in the process of performing a task or in user studies. Also, re-wearing the headset remains an issue for calibration. Accurate gaze-measurements might not be possible without recalibration once the user has put the HMD on and off the head. This problem might also exist when HMD slips. Even slight movements of the HMD on the user's head may lead to inaccurate eye tracking results and asking the user to repeat the calibration step each time the HMD become tilted is not a viable option. To this end, new continuously updating, online calibration methods, ideally embedded into the task, would make HMDs with eye trackers more practical for everyday applications.

6.5 CONCLUSION

This chapter has discussed a foveated rendering method that uses adaptive ray tracing and reprojection from previous frames to increase temporal stability and reduce artifacts. Sparsely sampled image data is reprojected to new views using a depth mesh generated from a low-resolution G-Buffer. The number of errors arising from the reprojection in regions critical for perception are reduced by an update strategy that allows these to be (re-)sampled by incorporating the samples' quality. The method enables the visualization of static scenes with millions of triangles within the Oculus Rift DK2 at a refresh rate of 75 Hz. Using the approach presented here, the benchmarks have shown significantly improved performance, while the user study has revealed that the perceived visual quality for even moderately sized FRCs is almost equal in quality to full rendering. Analyzing the tracking precision regarding its angular

dependencies has revealed a significant loss of tracking quality for higher eccentricities. When inspecting these inaccuracies of the tracking device further, it becomes clear that applications need to adjust the specific parameterizations for the given devices and users. An analysis of the user's ability to focus on static and moving fixation targets has revealed effects on the perceived visual quality. While the results appeared to be contradictory at first, given the intuitive assumption that worse fixations should result in worse quality ratings, the mean quality ratings were best for the *moving target* mode and independent of the scene – even though there was less of a match between the measured PoR and the actually focused PoR than for the static fixation mode. It can be observed how the mental workload and retinal velocities greatly degrade and limit our ability to judge visual artifacts.

Due to vast improvements in eye tracking solutions integrated into modern HMDs, research on gaze-contingent rendering is gaining increasing popularity. Nonetheless, the key question for the future remains: *"How can locally changing rendering and shading quality make the most effective use of perceptual limits to produce photo-realistic scenes with the required flexibility?"* [Wei+17] While rasterization can be used to adapt quality to an acuity model, e.g. by using deferred rendering or multi-resolution shading, the author of this thesis argues, that ray tracing is proving a higher degree of flexibility and does allow individual samples to be more efficiently updated (Section 4.2.3). Nonetheless, ray tracing sparse samples reduces ray coherency and in turn negatively influence vector processing and memory accesses. Hence, efficient foveated rendering requires a decision in favor of either visual quality or performance. While the presented reprojection and resampling strategy does provide great visual quality, the overhead costs are considerable. On the other hand, simpler methods might show artifacts. Thus, in the following chapter, a technique is presented that filters potential rendering artifacts by exploiting the eye's Depth-of-Field (DoF). This enables a simpler foveated rendering pipeline to be designed with which we can omit resampling salient regions for out-of-focus areas by filtering high-frequency noise using the inherent low-pass nature of the DoF.

7 GAZE-CONTINGENT DEPTH-OF-FIELD

Exploiting the Limitation of the Optics

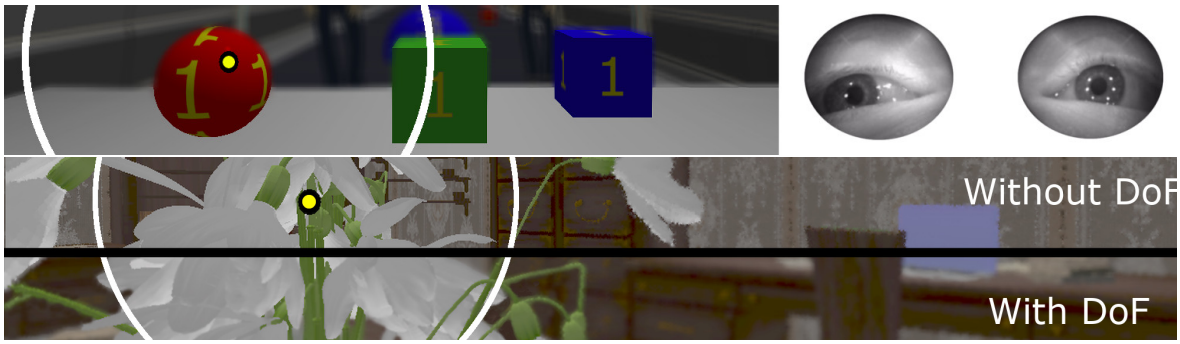


Figure 86: A highly integrated Depth-of-Field (DoF) filter can conceal artifacts arising from simple, yet efficient foveated rendering pipelines.

In addition to the system introduced in Chapter 6, in recent years several gaze-contingent rendering methods have been proposed to reduce the computational workload. The majority of them attempt to exploit the limitations of the Human Visual System (HVS) by adapting rendering quality to the user's retinal capabilities. Unfortunately, these foveated rendering methods mean that priorities have to be set. While they effectively reduce shading costs by disregarding a significant amount of pixels during rendering, this also leads to decreased coherence and to the necessity of additional filtering steps in order to fulfill perceptual requirements. The approach presented in the previous chapter attempts to alleviate such artifacts by using a Temporal Anti-Aliasing (TAA) scheme that is coupled with a resampling strategy. Similar approaches have been performed for rasterization pipelines. Methods commonly resample the scene or they apply post-process filtering in those areas that are critical for (peripheral) perception, e.g. to enhance contrasts in the visual periphery (Section 4.2.3). Unfortunately, depending on the specific situation, these approaches may even lead to an increase in the total rendering time.

Along with the retina's decreasing visual acuity, the HVS is also limited by its optical properties. Among the most prominent effects of this is Depth-of-Field (DoF), which occurs when focusing objects. While naturally, in an environment with objects at different distances, *accommodation* adjusts the focus distance to the fixated object. It is usually not possible to perceive all observed objects as one sharp image. While objects at the focused distance from the eye are perceived clearly, other objects appear increasingly blurred depending on the distance between their depth and the focused distance. At the same time, it is less important to create a high number of samples in image areas that are out-of-focus as these areas of course are not perceived sharply.

Figure 87: An especially challenging object for gaze-depth estimation using spatial measures or vergence measures alone. This view on the flower vase was used in our experiments to calibrate and predict gaze-depth estimates.

This chapter describes a system to exploit this knowledge about the DoF when rendering in a foveated fashion. To this end, DoF is applied in a post-processing step to conceal visual artifacts by removing high-frequency signals from the visual periphery using the inherent blur of the DoF (Figure 86). This allows more computational effort to be invested in regions that are more important either because they are fixated or because they are in focus.

In order to do so, several subproblems need to be addressed. Firstly, the focused gaze-depth is not as readily available as the Point-of-Regard (PoR). However, the accommodative state of the eye must be modeled to render a correct image using DoF. Secondly, a foveated rendering system, including a low-latency, high-performance DoF filter must be developed. As demonstrated in the previous chapter, the attempts to reuse samples temporally, as well as saliency-based resampling work impressively well. Unfortunately, the maintenance overhead for the system is high. For example, the system uses rasterization for the reprojection. Hence, besides, the central ray tracing routines, code for a separate rendering approach must be maintained. Also, the management overhead for all buffers is high, and for dynamic scenes, detecting reprojection errors and image regions with high saliency will become even more complex. While simpler methods are likely to elicit visible artifacts, they can be implemented more efficiently. A "simpler" fast system is needed, yet providing a sufficient visual quality at least when filtered with DoF.

Last but not least, a filter needs to be developed that allows the DoF to be quickly simulated to conceal those artifacts. While ray tracing makes it possible to sample a lens model to acquire physically accurate DoF, this increases sampling density to a point where it would have been more efficient to render in full detail directly. Hence, this chapter introduces a fast approximate DoF filter in image space. Also, when rendering in a foveated fashion and sampling the image plane sparsely a "final" image must be reconstructed. The information for those pixels must be recovered that have not been sampled. Using the method presented, DoF computation can be closely linked to this image reconstruction process in order to improve the perceived visual quality of foveated rendering.

Although it is possible for trained users to freely focus in space, in a typical application scenario the focused depth will likely belong to a point on the surface of a fixated object. This makes it feasible to use a regular eye tracker to derive a gaze-depth. If there were a perfect spatial calibration of the eye tracker, then tracing a ray through the PoR to the scene's geometric hit point and then computing its depth, would be sufficient to derive a perfect depth value. However, the spatial calibration is susceptible to inaccuracies. Hence, methods that attempted to rely on this spatial measurement have proven not to be feasible [MBT11]. Serious errors can occur in the depth measurements in the case of complex, thin objects or also when looking at a flat object at an angle and being only a fraction away from the actual PoR instead of the one acquired by the eye tracker. The accuracy of the eye tracker might not be high enough. If a user focuses on a thin object in the foreground, but the eye tracker reports a PoR right next to the thin object on the background in the distance, the reported depth and as a result the DoF computation is incorrect. A challenging object to estimate accurate depths, that was used throughout our user experiments is presented in Figure 87. Alternative methods for depth estimation include using the eyes' *vergence* (Section 2.2.3). However, there is less vergence with an increasing distance to the fixated object and the central optical axes of the eyes generally become parallel, depth estimation using vergence is reported to only work well for the first meter [EPR04; Wan+14]. Both the vergence and the depth measurement at the PoR suffer from a lack of spatial tracking accuracy. As evident from the discussion in the previous chapter, even high-quality devices inside Head-Mounted Displays (HMDs) yield an accuracy of about 1° of visual angle and this only in a limited region of the visual field (Section 6.3.2).

In order to increase the precision of gaze-depth estimation, a machine learning approach is presented that combines several gaze-depth measurements, including vergence and various spatially-obtained measurements. All of these are used in a calibration step as mixed input to train a regression model that allows more accurate predictions. In order to investigate the accuracy of this model, the required gaze data was collected by performing a user experiment. Finally, this machine learning model is used to control a filter to conceal artifacts when rendering in a simple, yet efficient, foveated fashion. As in the previous chapter, the foveated renderer is based on ray tracing. While exploiting Temporal Coherence (TC), a simple image space reconstruction technique is used to compute "dense" images from the sparse sample sets. At the last step, the gaze-depth is used to control a fast, layered and guided bilateral image space filter to add the appropriate DoF. The quality of this filter is determined by evaluating the results of another user study that demonstrates that visual quality can be greatly improved.

In summary, in this chapter, the following contributions are presented:

- A machine learning approach that is trained with multiple gaze-tracking and depth measurements, providing an improved gaze-depth estimator.
- A calibration procedure based on this feature set used to collect the required training data in synthetic and realistic scenes.
- An evaluation of the estimator in order to determine the accuracy, the required training set sizes, and an estimation of the performance across users and across scenes.
- A DoF model, incorporating knowledge about tracking inaccuracies when obtaining 3D gaze points

- A gaze-contingent rendering system with a tightly integrated DoF filter concealing potential visual artifacts
- A user study, evaluation, and benchmarks showing the potential of the proposed method to filter artifacts when rendering in a foveated fashion.

CONTRIBUTIONS BY THE AUTHOR This chapter is based on work published in the papers:

> Martin Weier, Thorsten Roth, André Hinkenjann, and Philipp Slusallek. *"Foveated Depth-of-Field Filtering in Head-Mounted Displays."* In: *ACM Transactions on Applied Perception (TAP).* Vancouver, Canada, Aug. 2018. Best Paper Award, invited article.
>
> Martin Weier, Thorsten Roth, André Hinkenjann, and Philipp Slusallek. *"Predicting the Gaze Depth in Head-mounted Displays using Multiple Feature Regression."* In: *Proceedings of the ACM Symposium on Eye Tracking Research and Applications (ETRA).* Warsaw, Poland, June 2018.

I was the primary investigator for both papers, developed the machine learning model to improve the gaze depth prediction and the rendering framework for gaze-contingent DoF. I also designed and executed the benchmarks, the user study, and the user experiment to evaluate both approaches. The data evaluation of the user experiment presented in Section 7.3 was performed in collaboration with my colleague Thorsten Roth. Here, I wrote the data evaluation and plotting routines for RQ1, RQ2, RQ5, and RQ6. At the same time, I assisted in developing the routines for RQ3 and RQ4.

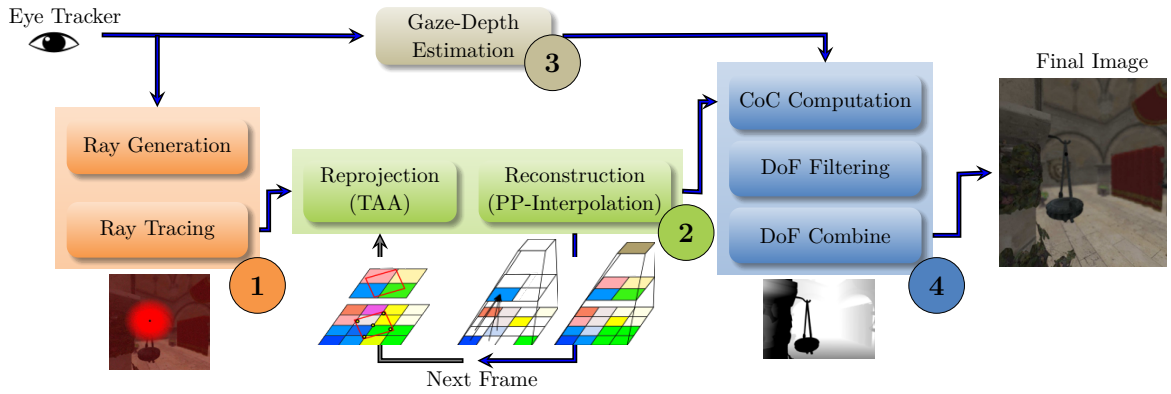


Figure 88: Rendering pipeline of the presented approach. First, ray tracing is used to sample the image plane sparsely based on a visual acuity model. Next, reprojection is used to increase the temporal stability of the sparse samples. Afterward, a reconstruction kernel reconstructs a dense image from the sparse samples. Finally, depth-of-field is computed to conceal artifacts in the final image.

7.1 METHOD

In order to conceal perceptually disturbing artifacts, a foveated rendering system is combined with an approach to compute gaze-contingent DoF in image space to filter the final image. The entire rendering pipeline is depicted in Figure 88. First, a ray tracing step samples the image sparsely according to a visual acuity model. Afterward, the temporal stability of peripheral image regions is improved by using backward reprojection-based TAA. Next, the full image is reconstructed using Pull-Push Interpolation (PP-Interpolation) [MKC07]. This efficient approach provides a high degree of flexibility, allowing to use arbitrary acuity models and (re-)sampling strategies [Ste+16; Wei+16]. Finally, to further improve the perceived image quality, gaze-contingent DoF is computed in a post-processing step. A more detailed description of each pipeline stage is presented in the following sections. All components are implemented using NVIDIA CUDA.

7.1.1 *Ray Generation and Ray Tracing*

The ray generation, relies on the same visual acuity model as introduced in Section 6.1.1. Here, a linear fall-off in a transitional region is modeled between the area of central and peripheral vision (Figure 89). Again, this model is configured using a triplet (r_0, r_1, p_{min}), referred to as the Foveal Region Configuration (FRC). However, in contrast to the approach presented in Section 6.1.1, this time the sampling points are precomputed at program launch and stored as two binary lookup tables S_{in} and S_{out} that represent pixels on the image plane. For each frame a CUDA ray generation kernel is launched for every pixel on the image plane. A ray is only generated for pixels with the associated bits, either in S_{in} or S_{out}, set to one.

The sampling pattern for the peripheral vision is represented with S_{out}. The pattern contained in S_{out} corresponds to a uniform distribution with a sampling probability of p_{min}. For central vision and the transitional region, S_{in} is used to mark the foveal and sampling points in the transitional region between central and peripheral vision. The pattern S_{in} is translated

based on the PoR obtained by the eye tracker. The translation is performed by changing the addressing of this pattern in the ray generation kernel.

The reason for using a static precomputed patterns is twofold: Firstly, static patterns can be computed offline. Thus, they allows for using high-quality low-discrepancy sampling (Section 4.2.1). Secondly, the static patterns minimize temporal inconsistencies and flickering. When using stochastic sampling, the pattern is constantly changing from one frame to the next. As the camera moves and not all pixels are updated, the image flickers. This is especially noticeable along discontinuities, such as depth or contrast edges, in the image. Exploiting the TC helps to reduce flickering. However, it cannot remove such artifacts completely, whilst maintaining a sharp image. Hence, using static precomputed patterns becomes necessary as salient parts such as edges or contrast discontinuities are also not specifically resampled in contrast to the approach presented in Chapter 6.

The binary lookup table S_{in} consists of two regions, limited by angular thresholds r_0 and r_1, $r_0 < r_1$. Angular distances $d < r_0$ are sampled with a probability of 1, while samples within the transitional region ($r_0 \leq d < r_1$) are generated with an importance sampling approach in polar coordinates:

$$r = r_0 + (r_1 - r_0)\frac{\sqrt{({p_{min}}^2 - 1) \cdot u + 1} - 1}{p_{min} - 1} \tag{8}$$

$$\phi = 2\pi v \tag{9}$$

Here, $u, v \in [0, 1]$ are random variables computed using Halton sequences. The fractional part of the function for generating samples for r (Equation (8)) is derived by applying the inversion method to $f(x) = 1 - (x \cdot (1 - p_{min}))$, describing the linear fall-off in the transition region, with $\int_0^1 cf(x)dx = 1$. These values are then transformed to the range $[r_0, r_1]$. At runtime, a ray generation kernel looks up which pixels to sample by querying both S_{out} and S_{in}. The latter is shifted based on the pixel's distance to the current PoR. If one of the queried bits is set, a ray is generated. Eventually, CUDA threads are launched to compute pixel values for all generated rays. In order to do so, the ray caster presented in Chapter 6 has been extended to use the irregular grids developed by Pérard-Gayot et al. [PKS17] in order to accelerate ray-geometry intersection.

7.1.2 *Reprojection and Reconstruction*

As sparsely sampled image regions lead to temporal instabilities, exploiting TC has become a standard for foveated rendering systems (Section 4.2.3). In contrast to the *forward reprojection* using OpenGL as performed in the system introduced in the previous chapter, this time *backward reprojection* is used (Section 4.3.1). This eliminates the necessity of a separate rasterization step and the construction of a warping geometry. The code complexity is significantly reduced.

The *backward reprojection* is performed as follows: The pixel footprint described by each ray and its differentials [Ige99] is adapted based on the eccentricity-dependent sampling probabilities. Now, the pixel footprint is transformed into world space and reprojected to the previous frame. As the view is changing, the pixel footprint might be translated, scaled and rotated between frames. Therefore, the color information of the old frame must be evaluated by sampling the reprojected pixel's footprint extent multiple times. As described below, the

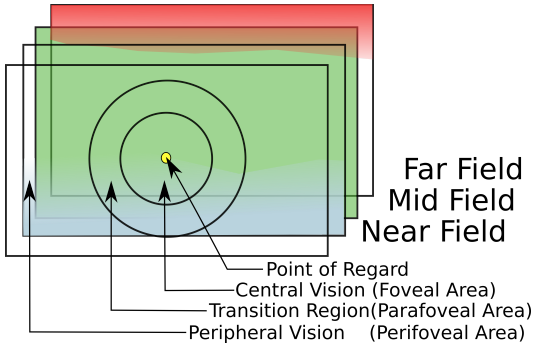


Figure 89: Layered blur to compute the depth-of-field. For peripheral vision, a blur using a mipmap representation is computed. For central vision, a separable Gauss is used to blur the values for each layer. The transition region is blended between the blur approaches.

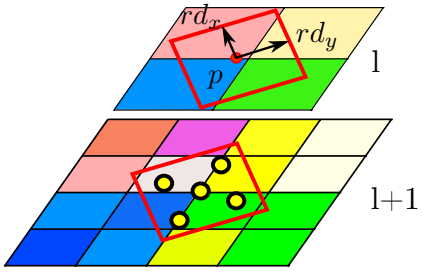


Figure 90: Sampling in the image pyramid of the last frame with respect to the backward projected pixel footprint (rd_x, rd_y) at the ray's hitpoint p. For sampling a Quincunx pattern is used (yellow dots).

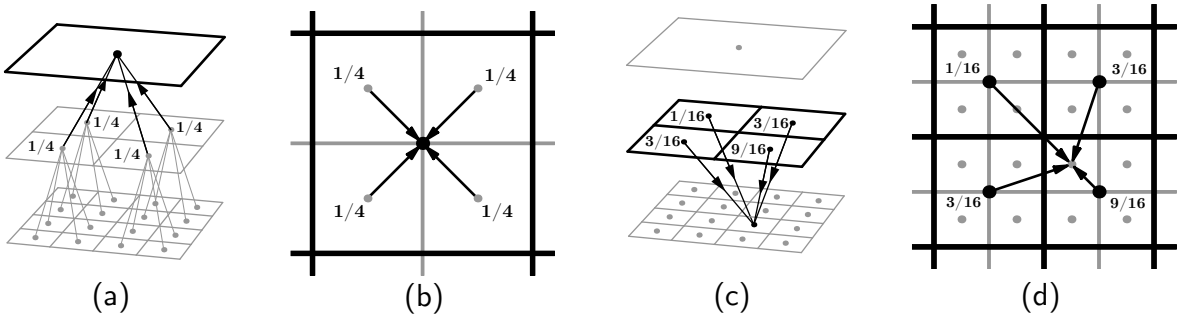


Figure 91: Pull-push reconstruction method to render 3D point sets. In the pull phase, valid samples are pulled and combined upwards *(a)*, level by level, to fill a mipmap pyramid. For each pixel, the respective valid pixels from the finer level are averaged *(b)*. Afterward, in the push phase *(c)* the mipmap pyramid is processed from the coarsest to the finest level. Missing pixels are filled in by sampling the pixel value in the next higher mipmap level with the coarser resolution *(d)*. *Image from Marroquim et al. [MKC07]*

old frame's information is not only a single image, but an image pyramid, similar to a mipmap with depths stored in the alpha channel. In order to improve the precision of the reprojected colors and depths, five samples are evaluated in a Quincunx pattern for a higher mipmap level as illustrated in Figure 90.

The depth values obtained in the alpha channel facilitate for eliminating cache values in case of perspective-related occlusions. The reprojection kernel can check if the depths of the samples between to frames difference is above a critical ϵ threshold [Yan+09]. An occlusion might have occurred if the reprojected depth values between subsequent frames are too different. In this case, samples from the cache should not be considered. Now a running estimate combines the colors from the new and the old frame (Appendix A.3).

As the system is still dealing with a sparse image, i.e. an image with "unfilled" gaps between pixels, eventually, these temporally smoothed samples are used to reconstruct a new image using PP-Interpolation [MKC07]. The idea of PP-Interpolation has been introduced for point-based rendering. It is divided into two distinct phases. The *pull* and *push phase* are

illustrated in Figure 91. In the *pull phase*, mipmap pyramid levels are computed in bottom-up order. The attributes (colors, normals, etc.) of a pixel of a coarser level are determined by averaging the corresponding four pixels of the finer pyramid level. However, only pixels that have (valid) data are included in the mean. If all (up to four) pixels are invalid, the new pixel is also marked invalid and is left to be computed during the push phase. Pixels can also be invalidated, if they are considered to be occluded depending on their depth value – a necessity for point-based rendering. In the *push phase*, the algorithm works in top-down order, i.e., from coarser to finer levels and only the attributes of invalid pixels are (re-)computed. Here, the attributes of four pixels of a coarser mipmap pyramid level are used to interpolate the attributes of a pixel of a finer level. Missing pixels are filled in by sampling the pixel value at the image in the next higher mipmap level with the coarser resolution. Finally, the mipmap pyramid becomes the input for the reprojection phase of the next frame and is also used as input for the gaze-contingent DoF filter presented in the next section. To this end, the focused gaze-depth has to be known.

7.1.3 *Gaze-depth Estimation*

Most recently, direct measurements of the eyes' accommodative state for HMDs using autorefractors have been performed [Pad+17; Mer+17]. While these devices are bulky and slow, rendering the DoF effect does not drive the physiological accommodation at all [Pad+17]. Hence, it is simply not possible to use such a device inside a consumer level HMD with a single screen and a fixed-focus lens, i.e. when it is not using multifocal lenses or lightfield displays. However, for gaze-contingent DoF it is critical to obtain a gaze-depth in order to model the eye's accommodative state. As the focused depth will likely belong to a point on the surface of a fixated object, it is feasible to use a regular eye tracker. Using binocular eye tracking and displaying synthetic images, several measurements can be used to estimate the depth of the fixated object. However, spatial and vergence measurements might lack accuracy due to the precision of the eye tracker and physiological constraints. Mantiuk et al. [MBM13] improved tracking accuracy and stability when rendering gaze-contingent DoF by applying object and scene knowledge. Essentially, a Hidden-Markov model is used to derive probabilities on how likely an object in a 3D scene is fixated. Probabilities are computed by combining 3D scene information, gaze positions, and velocities. Generally, methods are needed that combine more measurements in order to increase the gaze-depth prediction for off-the-shelf eye tracking hardware. The approach by Mantiuk et al. presents remarkable results. It can be used complementary the system here introduced. Besides, this system is not limited to tracking distinct objects in the 3D scene and is likewise not limited by a discrete set of fixation locations.

Our main idea is to combine several measurements obtained by the eye tracker. Initially, multiple samples from the depth buffer of the rendered image in a region centered at the PoR are taken. This region can be scaled according to a potentially eccentricity-dependent tracking accuracy. For these samples, various measurements are taken and combined into a *feature set*. These measurements include information about the eye's vergence, spatial depths at and around the PoR and depth variances as illustrated in Figure 92. In a calibration phase, these measurements are used to train a regression model based on a Support Vector Machine

(SVM). At runtime, these measurements are used as input to the SVM to obtain the focused depth.

For training and runtime evaluation the eye tracker is queried to compute a PoR and gaze vectors. The gaze vectors describe the central line of sight emerging from each eye. This information is used to derive the two vergence-based depth estimates described below.

Method for depth estimation by Wang et al.: The technique by Wang et al. [Wan+12] uses the PoR of the left $L_{PoR} = (l_x, l_y)$ and right eye $R_{PoR} = (r_x, r_y)$ in image space. Here, Wang et al. average the PoRs' heights $y = (l_y + r_y) \cdot 0.5$. The horizontal vergence in screen space can now be described as distance $\Delta x = r_x - l_x$. This assumes that the PoRs are referenced on a single screen. However, HMDs are often centered around two screens, and thus two different PoRs in two reference spaces are given. We assume that both screens are a single display and shift the right eye's PoR to the right, resulting in $\Delta x = ((1 + r_x) - l_x) \cdot 0.5$. With an Interocular Distance (IOD) either assumed to be $d_{IOD} = 0.063\,\text{m}$ [Duc+14] or measured using the eye tracker and an experimentally determined distance to screen d_{screen}, the gaze-depth can be computed as

$$z_{wang} = (\Delta x \cdot d_{screen})/(\Delta x - d_{IOD})$$

Ray-based depth estimation: As the skewed gaze vectors in 3D space do not necessarily intersect in a single point, the distance to the points of closest approach can be computed for both gaze vectors. The points of closest approach on gaze vectors $\vec{r_d}$ and $\vec{l_d}$ of the right and left eye can be described as $P = R_o + t \cdot \vec{r_d}$ and $Q = L_o + s \cdot \vec{l_d}$, with $R_o = (d_{IOD} \cdot 0.5, 0, 0)$ and $L_o = (-d_{IOD} \cdot 0.5, 0, 0)$. For the points of closest approach, it holds that the vector between them is perpendicular to the gaze vectors. Thus we have $(P - Q).\vec{r_d} = 0$ and $(P - Q).\vec{l_d} = 0$. This condition makes it possible to set up a system of equations to compute t and s by solving:

$$\begin{bmatrix} \vec{r_d}.\vec{r_d} \cdot t + \vec{r_d}.\vec{l_d} \cdot s = -R_o.\vec{r_d} + L_o.\vec{r_d} \\ \vec{l_d}.\vec{r_d} \cdot t + \vec{l_d}.\vec{l_d} \cdot s = -R_o.\vec{l_d} + L_o.\vec{l_d} \end{bmatrix}$$

The mean distance to P and Q is computed and assumed to be the gaze-depth as

$$z_{ray} = (P_z + Q_z) \cdot 0.5$$

As additional features, the information of the device's spatial tracking accuracy and the scene's depth buffer is taken into consideration to estimate the gaze-depth. In a first experiment, data about the tracking accuracy of the HMD was collected. Users were requested to fixate and follow a tracking target, guiding the user's gaze through the scene, much like it was performed to evaluate the system presented in the previous chapter. This data showed that this time spatial calibration of the eye tracker used provided an accuracy of roughly $\pm 1°$ of visual angle. This inaccuracy was later confirmed by the primary study that was conducted in order to evaluate the approach here presented (Section 7.3.2). Now, besides computing a single depth estimate at the PoR from the depth buffer, more data in the region of spatial tracking uncertainty is collected. For this purpose, n depth samples are drawn around the PoR (Figure 92), covering the $1°$ radius of tracking inaccuracy. Throughout the experiments, $n = 20$ samples are drawn. These samples are used to compute a mean depth and a normalized variance. Moreover, the PoR's distance to the screen center (eccentricity) was recorded.

Figure 92: Depth samples (red) drawn around the PoR (yellow) to estimate the gaze-depth. The size of the PoR and the sampling pattern are exaggerated for visualization purposes. The upper right image shows a magnification of the sampled region. The lower right image shows the same region with the tracking target used to guide the user's gaze.

All of the previously describe gaze-depth measurements and statistics can be combined into a set of gaze-depth features (feature set), summarized in Table 8.

The model presented assumes that if the variance is close to zero and the mean of all depth samples results in a depth value close to the central depth measured at the PoR, the depth at the PoR is likely the gaze-depth. However, if there is a high variance in the depth samples, but their mean depth is close to the virtual camera, the underlying model should give the vergence measures a higher weight when deriving the gaze-depth estimate. The challenge thus is when and how to weight a vergence-based and a spatial measurements. However, building such a model can be achieved using machine learning.

Feature	**Description**
`Center`	Depth at the PoR (depth buffer)
`Mean`	Mean of the samples around the PoR (depth buffer)
`Var.`	Variance of the samples around the PoR (depth buffer)
`Ray`	Ray-based depth estimate (vergence)
`Wang`	Method by [Wan+12] (vergence)
`Ecc.`	Eccentricity of the PoR regarding the screen center

Table 8: Feature set used for training the regression model to improve gaze-depth estimates.

Although various approaches exist that allow for regression of such feature sets to train a predictive model, here Support Vector Regression (SVR) with radial basis functions [Gun98] is used. While the training phase of SVR models is computationally demanding, predicting new values is highly efficient. This allows for low latencies, which is essential when rendering

to an HMD. In order to decrease training times, computationally faster approximations for large kernel machines are available, for example by Rahimi et al. [RR07]. Compared to other machine learning approaches, one advantage of SVR with radial basis functions is that the quality of the model is only determined by a small parameter set, depending on the type of the underlying SVM implementation. This makes it an ideal tool for quality estimation. In order to get the best results, it is recommend to tune the parameters based on preliminary cross-validation. For runtime evaluation inside the renderer libSVM [CL11] is used.

7.1.4 *Depth-of-Field Filter*

Given the depth estimate, the reconstructed image, and the mipmap pyramid from the PP-Interpolation, the DoF effect can be computed to produce the final image. Its distinct steps of computation are detailed below.

7.1.4.1 *CoC Computation*

Although the upcoming evaluation does show that combining multiple measures into a single regression model improves gaze-depth estimation substantially, it is still suffering from inaccuracies. One way to tackle these inaccuracies is to employ temporal filters such as higher-order Butterworth filters in order to temporally smooth the estimates, as for example performed by Duchowski et al. [Duc+11]. The system presented here makes use of such a filter to obtain smoother depth estimates but the user can choose a high cut-off frequency for a more responsive adaptation. In addition, a conservative model is proposed that accounts for potential tracking inaccuracies by extending the focus range based on the accuracy of the depth estimate.

The conservative model assumes a thin lens that can be expressed using the general lens equation $1/f = 1/d_o + 1/d_i$. In this work, d_i, the distance to the image plane is assumed to be fixed with 22.4 mm, which is itself an estimate based on the average of measurements of the human eye [Gro05, chp. 36.4]. The object's distance d_o to the focused distance is obtained using the machine learning model. Now two focal lengths f_n and f_f are computed to obtain a near and far focal plane. This is achieved by using the mean tracking inaccuracies t (Figure 98 green line) to adapt d_o as $d_n = d_o - t$ and $d_f = d_o + t$ in order to solve the general lens equation to determine the focal lengths. Finally, these values can be used to compute a signed Circle-of-Confusion (CoC) as

$$CoC = \begin{cases} -(V_d(f_n, g) - V_f(f_n, d_n)) \cdot (k \cdot V_d(f_n, g)) \cdot E & \text{if } g < d_n \\ (V_d(f_f, g) - V_f(f_f, d_f)) \cdot (k \cdot V_d(f_f, g)) \cdot E & \text{if } g > d_f \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

with $V_d(F, G) = (F \cdot G)/(G - F)$, $G > P$ and $V(F, D) = (F \cdot D)/(D - F)$, $D > F$. A derivation of these formulas is presented by Mulder and van Liere [ML00]. The distance to the unfocused object is denoted g. E is a measure of the retinal resolution, such as the reciprocal of the size of the receptive fields of the photo-sensitive cells, and k is the pupil's diameter. While having experimented with various estimates of E using, for example, the size of the ganglion cell's [Bar99, pp. 66-74] or the Minimum Angle of Resolution (MAR)

(Section 3.1.2), it was decided to determine it experimentally to control the strength of the DoF effect. k is mainly dependent on the retinal illumination. It can either be estimated, e.g. by using Le Grand's approximation formula (Section 3.1.1), measured directly using the eye tracker, or selected by the user in order to control the intensity of the DoF effect.

The given formula to compute the CoC can be used in post-processing DoF approaches such as the method introduced by Mulder et al. [ML00; Buk+13]. However, if one wants to model and sample an actual lens model using ray-based [KMH95] or wavefront models [Kak+07], a lenses optical imaging qualities are usually not described with two focal lengths, f_n and f_f. However, in those cases, it is possible to alter the parameter of the retinal resolution E, i.e. the sensor size, to have a wider focus range in order to take the region of tracking inaccuracy into account.

In order to compute the CoC, a CUDA kernel is launched that calculates a signed value for each newly computed sample. The sign marks if a pixel is located in the near or the far field. Finally, each thread stores the color information and the signed CoC's size encoded in the samples alpha channel.

7.1.4.2 *DoF Filtering*

As gaze-contingent rendering for HMDs requires low latencies to cope with fast eye movements and to match the displays' refresh rates, performance is a critical aspect when computing gaze-contingent DoF. It must be remembered that approaches that sample a lens model with the aim to compute physically correct DoF require casting many rays, which is contradictory to the idea of foveated rendering that aims to reduce the number of shaded samples. Although various more efficient techniques have been developed to approximate DoF [Dem04; MRD12], they are still barely usable inside a performance-critical rendering pipeline as even a simple adaptive Gaussian blur at the native HMD resolution can take several milliseconds.

In general, the presented approach for DoF computation follows the idea of Buchowski et al. [Buk+13]. However, in order to calculate the DoF effect more efficient its computations are closely coupled with the foveated sampling and reconstruction scheme. Based on the pixel-wise computed CoC, the rendered image is divided into three different layers: One layer for pixels in the far field (with a CoC$> \epsilon$), another layer for pixels in the focused mid field (with $\|\text{CoC}\| < \epsilon$), and a third layer for pixels in the near field (with a CoC$< -\epsilon$). These three different layers are illustrated in Figure 89. Pixels are blurred, distributed on the according layer, and combined to a final image.

In fact, for a plausible DoF effect, the blur has to take depth discontinuities continuously into account and without a limited set of layers. Blurry distant objects should not bleed over closer objects in the focused field. However, the assumption in this work and the approach by Buckowski et al. is that as long as the order of the layers is preserved, users are likely to tolerate potential blurring inconsistencies within each layer.

Despite this, even blurring a few buffers is computationally expensive. This is especially the case when objects are focused at close range. Objects in the background, which are then out-of-focus, must be heavily blurred. Filter kernels for blurring these become very large, and as a result, the blurring operation becomes very slow. Therefore, Bukowski et al. work with buffers of reduced resolution in order to obtain very blurred regions with a small filter kernels and to stay within frame time budgets. However, the reduced resolution is not suitable when

using an HMD. The reason here is the low angular resolution of the HMD's display compared to its large Field of View (FoV). Block and subsampling artifacts become visible. Therefore, it is necessary to develop a method that can efficiently blur large areas at high resolutions.

In order to efficiently blur images, the presented approach samples the appropriate levels from the mipmap pyramid, considering the CoC values for each pixel. Here, the algorithm samples the mipmap several times in order to further improve the quality of the blur. In addition, the type of blur and its quality are adapted based on the position in the visual field. Filtering using the mipmap pyramid is particularly suitable for areas in the peripheral visual field that already have low visual acuity. Higher quality blurs and better blending weights to combine the final image layers are computed for central vision.

The blur kernel processes the image as follows: In the first pass of the filter, the pixels are read horizontally. If a pixel is in a peripheral vision field, the mipmap is sampled. The radius of the CoC is used to select the appropriate mipmap layer. In order to improve the quality of the blurring operation, the mipmap is sampled multiple times [W3C16] using the same Quincunx scheme as already used in the *reprojection phase* (Figure 90). Now, based on the sign of the CoC and ϵ, the pixel value is stored either in the near or far field image buffer. In order to blend between the layers, an additional coverage value is computed based on the samples from the mipmap pyramid and the current pixel. To this end, the sign of the CoC stored in each sample's alpha value is examined. This makes it possible to only consider those samples in the blurring operation that are valid for the current near or far field. More details on the computation of the coverage values are provided by Buchowski et al. [Buk+13].

In the area of central vision and in the transition area to peripheral vision, a high-quality blur is required. High visual quality matters here. At this point, if a pixel is in a position of the visual field that is between central and peripheral vision, both pixels in the near and in the far field are blurred with a Gaussian blur.

Still, the transition between central and peripheral vision may be visible in the final image due to the different approaches to blur and combine the image. In order to overcome this issue, both types of blur are blended in the vertical pass of the separable filter as follows: pixels that are visible by peripheral vision are already blurred and remain untouched. Pixels that are located in the central and transition region of the visual field are blurred in the vertical direction using a Gaussian blur. However, for pixels that reside in the transitional region, the blur resulting from sampling the mipmap is also computed. This allows blending between the mipmap and the high-quality Gaussian blur in the near and far field buffer in a single pass. The blending weights are chosen to be linearly controlled by the falloff parameters r_0 and r_1, describing the central and transitional region of the visual field (Section 7.1.1). After two separate blurred images for the near and far field with coverage values have been computed, all available images are combined to a final image.

7.1.4.3 *DoF Combine*

In order to obtain the final image, the three available image layers are blended: the blurry near field buffer, the blurry far field buffer, as well as the unmodified original image. Firstly pixels are interpolated between the original unmodified image and the far field buffer based on the CoC. Near field values are blended over the resulting image using alpha blending with

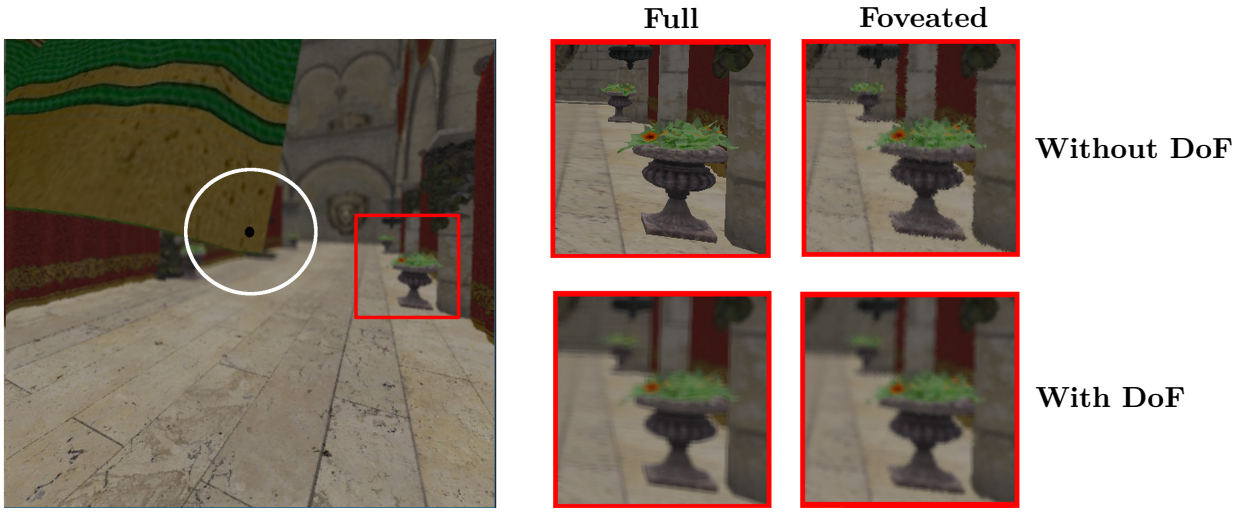


Figure 93: Different rendering configurations showcasing the potential of the depth-of-field filter.

the associated coverage information stored in the alpha channel. This way, computing the transitions between the regions can be handled efficiently using simple linear interpolations.

Another technique to counteract the quality degradation for peripheral vision is eccentricity-dependent contrast enhancement. Patney et al. [Pat+16b] show that sampling rates in a foveated rendering system can be reduced when contrast enhancement for peripheral vision is applied. Several other researchers have also shown that enhancing contrasts helps to conceal lost details [Gru+06; KRK11]. Patney et al. achieved the contrast enhancement by weighting each pixel's color with a blurred version of its surrounding using a kernel width adapted to the eccentricity. However, computing such a blur takes considerable time, and Patney et al. found that box blurs do not provide the necessary fidelity [Pat+16c]. However, the mipmap helps to obtain blurred images at a high quality. Again, this is achieved by consecutively sampling the mipmap. Essentially the box blur represented by the mipmap is used to approximate a Gaussian blur [W3C16]. Having an efficient way to blur images adaptively, the color value for each pixel p'_{ij} in the final image is computed by evaluating

$$p'_{ij} = \bar{p}_{ij} + f_e \cdot (1 + \sigma_{ij}) \cdot (p_{ij} - \bar{p}_{ij})$$

using the unmodified pixel color p_{ij} and the blurred version $\bar{p}_{ij}$. The value σ_{ij} measures the filter width. It is zero for central vision and increases with the eccentricity in the visual field. The parameter f_e is user-defined. Patney et al. found $f_e = 0.2$ to yield satisfactory results. Finally, the combined image is presented to the user.

A comparison of different rendering configurations using DoF for the presented foveated renderer in contrast to full rendering is shown in Figure 93. Please note how enabling DoF filters potential rendering artifacts.

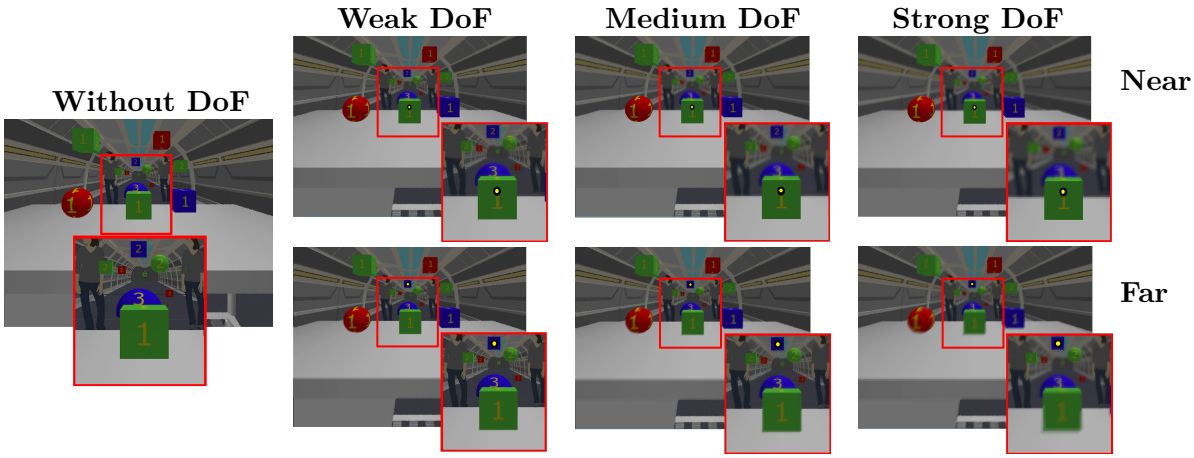


Figure 94: The final renderings as presented in the gaze-contingent renderer. The images show the test scene that was used in the user experiment. The PoR is marked as a black dot with a yellow center, either fixating a target close to the user (near) or in the distance (far). Moreover, this image shows the different DoF modes as used in the user study. The scene itself consists of various targets at different depths. Targets labeled one range from 0.5 m to 1 m. Targets labeled with a two range from 1 m to 6 m. The big blue ball labeled three is located at a distance of 6 m.

7.2 BENCHMARKS

In this section, the benchmarks of the renderer and filter are presented. The benchmarks and the accompanying user study were performed on Windows using an Intel Core i7-3820 machine clocked at 3.6 GHz equipped with 16 GiB RAM and two NVIDIA GeForce GTX 1080 Ti graphics cards with 11 GiB VRAM each. For the benchmarks, the scene *space shooting range* [Ros14] was modified. It consists of a long tunnel extended by the different targets as illustrated in Figure 94. The version of the scene used here consists of 227568 triangles. Please note that the runtimes of the kernel that performs the reprojection, the reconstruction, and the DoF effect are independent of the scene's geometric complexity; only samples in image space are processed. The runtime of each pipeline stage is shown in Table 9. The scene was rendered for a single eye, matching the HMD's native resolution of 1280×1440. Run times were averaged over 1000 frames. For these renderings, the same foveated configuration as for the user study was chosen. Foveated rendering is configured with a FRC of $(r_0 = 10°, r_1 = 20°, p_{min} = 0.2)$. The radius that is specifying the central vision (r_0) and the transitional

				DoF				
Mode	**Ray Tracing**	**Reproj. (TAA)**	**Reconstr. (PP-Interp.)**	**CoC**	**Filter**	**Combine**	**Total**	**# Samples**
Foveated (Ours)	3.09	1.35	1.107	0.81	0.88	0.41	7.64	**558945**
Full Ray Tracing	5.52	-	1.09	0.81	0.88	0.38	8.68	**1843200**

Table 9: Benchmarks of the presented pipeline in ms for a single eye rendered at a resolution 1280×1440 averaged over 1000 frames. The approach presented reduces the number of shaded samples by 69% compared to full ray tracing.

region (r_1) have been chosen based on the experience from the foveated rendering system introduced in the previous chapter. This configuration has proven to be sufficient to generate images that could mostly not be distinguished from full rendering. Also, these settings are chosen that central and transitional region covers the *parafovea* and *perifovea*, respectively (Section 2.1.1). Compared to the system that was introduced in the previous chapter, higher values for p_{min} (0.2 instead of 0.05) were chosen. This is because no salient image regions are resampled here.

DoF was computed with the *medium* setting as illustrated in Figure 94. For foveated rendering, the inaccuracy of eye tracking can be compensated by using a larger foveal area or by predicting saccadic movements [Ste+16; Ara+17]. However, determining the eye's focused depth in order to control the DoF is probably more inaccurate. Fortunately, the use of DoF is supported by the rather slow speed of accommodation [TM89]. Hence, the update rate of the focused depth is usually not critical when using eye tracking devices. A simple solution for counteracting inaccuracies is to render more of the scene in focus. However, in this case, filtering quality might be influenced negatively. Fewer parts of the image are affected by the blurring of the DoF. As a result, more artifacts are potentially visible. Nonetheless, the method proposed to compute the CoC enables to compensate for inaccurate depth estimates. As a result of the measurements presented in the next section, the mean depth estimation inaccuracy t can assumed to be 0.2 m over the entire critical depth range of 6 m. Although the total performance increase of about 1 ms does not appear to be noteworthy, the scene was rendered using only primary rays and simple shading. No secondary contributions like shadows, ambient occlusion or Global Illumination (GI) were computed. With an increasing computational complexity of each shaded sample, the difference between foveated and full rendering is expected to be much higher. Thus, the most important measure is the difference in the number of rendered samples between foveated and full rendering. This reduction is quite substantial, with the number of samples being reduced by 69%.

If a look is taken at the percentage of reduced samples using the method introduced in the previous chapter – it was up to 79% – it becomes clear that the system here compares well. However, note that in contrast to the system proposed here, this number of 79% is scene- and view-dependent as salient image regions may need to be resampled using the previous method. Also, DoF has the potential to increase realism and depth perception as well as to reduced motion sickness [Hil+07; Hel+10; Lan+16]. Comparing to the rest of the competing state-of-the-art, Guenter et al. [Gue+12] rasterize the image in three layers with different resolutions and render only 7% of the pixels. The image is strongly undersampled. Stengel et al. [Ste+16] report that shaded pixels are decreased by 65% for the same resolution of 1280×1440. Patney et al. [Pat+16b], relying on Coarse Pixel Shading [Vai+14], do not reduce the visibility rate (pixel writes) but the number of shading computations on the shaded quads to about 50% compared to the work by Guenter et al. [Gue+12]. In contrast to Coarse Pixel Shading [Vai+14], the PP-Interpolation used provides high flexibility at a reasonable cost. The ray casting approach presented here has a runtime of 7.64 ms. In order to stay within the V-Sync limits of the HMD, two render threads were launched on two Graphics Processing Units (GPUs). Besides rendering, most time is spent on reprojecting information from previous frames.

Concerning the visual quality in the foveal region the approach presented mostly resembles the work by Bukowski et al. [Buk+13]. Identical quality is achieved there. However, quality is reduced for parts in the peripheral visual field as the high-quality Gaussian blur is replaced

by the box blur from samples of the mipmap pyramid. Nevertheless, as the latter is sampled multiple times to compute a final color and coverage information, differences are hardly noticeable – especially, due to the general acuity loss at increasing eccentricities. Interestingly, in this setup, the difference in runtime between the various DoF settings is rather minimal. Once reprojection and reconstruction have been performed, the total time to compute the DoF (CoC, Filter, Combine) with *weak*, *medium* and *strong* settings amounts to 2.08 ms. The almost constant runtime of the DoF filter can be accounted to the heavy usage of the mipmap pyramid in the peripheral visual field, making the amount of blurriness largely independent of the runtime. However, slightly worse runtimes are to be expected if the focused region in the foveal and parafoveal region contains a higher amount of objects not in focus.

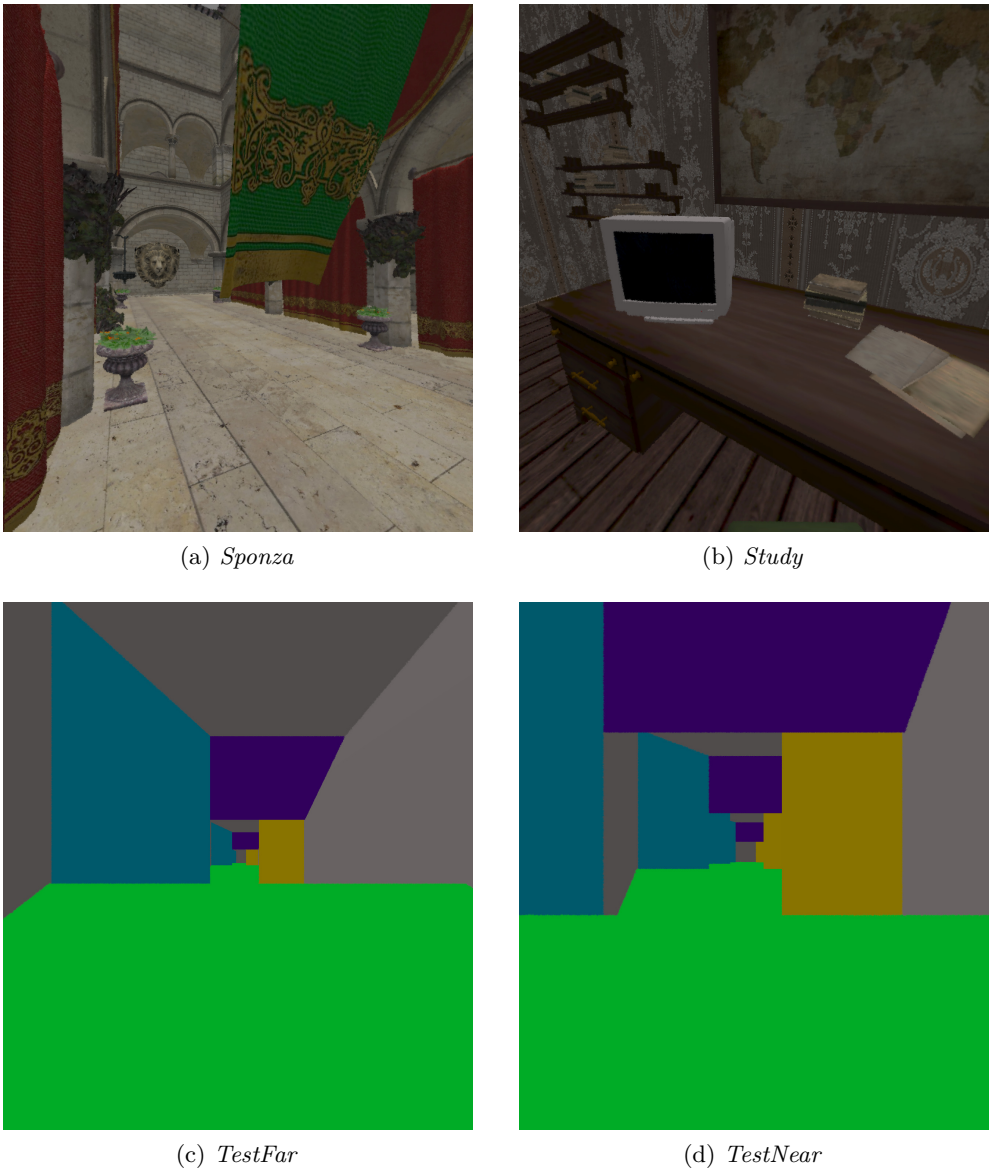
7.3 EXPERIMENTAL EVALUATION - TRACKING DATA

In this section, the results of an experiment with the aim to evaluate the quality of the gaze-depth estimation are presented. The following research questions provide insight into the software's potential:

- **RQ1:** Can the depth estimate be improved by combining measurements and how accurate is the resulting model?
- **RQ2:** How much does accuracy depend on the target's depth when using vergence alone vs. the introduced model?
- **RQ3:** How much does accuracy depend on the number of training samples?
- **RQ4:** Can the model be used across users and scenes?

7.3.1 *Procedure and Apparatus*

An experiment to collect the required gaze data was performed to evaluate the gaze-depth estimation framework. In contrast to the setup used to evaluate the foveated rendering framework in the previous chapter, a Fove 0 Headset [Fov17] was used. This HMD is natively equipped with a binocular eye tracker running at up to 120 Hz with a precision of 1° and a latency of 14 ms. The experiment was conducted as a within-subject user experiment, employing a $4 \times 6 \times 3$ full factorial design. Each participant had to perform 72 trials. The trials consisted of four scenes, two typical video game scenarios (*Sponza* and *Study*) as well as two test scenes (*TestFar* and *TestNear*), as shown in Figure 95. While for the test scenes a single static camera position was chosen for all trials, the typical video game-like scenes *Sponza* and *Study* were presented with six different camera positions (Appendix A.5). These positions were chosen to have a high variability of depth changes, potentially spanning the entire depth range in question, with a maximum of 12 m. The test scenes were modeled to range from 0 m to 12 m (*TestFar*) and from 0 m to 3 m (*TestNear*), rendered with simple shading to limit the number of possible distractions. Each of these scenes was composed of skewed boxes at fixed depth intervals (Figure 95). All camera positions were randomly shuffled and presented to the subjects three times. For each of these camera positions, a tracking target was generated that followed a randomly generated path through the scene. It was presented in front of a gray background for two seconds before the actual movement and data acquisition began in



(a) *Sponza* (b) *Study*

(c) *TestFar* (d) *TestNear*

Figure 95: Test scenes and exemplary views used in the user study to acquire gaze data and to evaluate the proposed method.

order to enable participants to find the target's initial position. The paths were generated to be within a radius of the innermost 20° of the visual field around the central optical axis to cover the Comfortable Viewing Angle (CVA) of 15° (Section 2.2.3). Moreover, this was intended to ensure the target's visibility throughout the entire trial for each participant.

As illustrated in Figure 92 the tracking target was a semi-transparent red ball with an opaque black center. This target was adapted to the scene's depth at the current location on the tracking path. As depth changes of the target along the randomly generated path can potentially be high and for physiological reasons the participants are not able to perform the vergence movement sufficiently fast, the accepted speed of depth changes was artificially reduced. The maximum depth change allowed at which a user can still focus on an object is said to be approx. 0.7 m/sec (Section 2.2.3). In order to ensure that the user can focus on the object at all times, the movement was further slowed down. Otherwise, double vision might

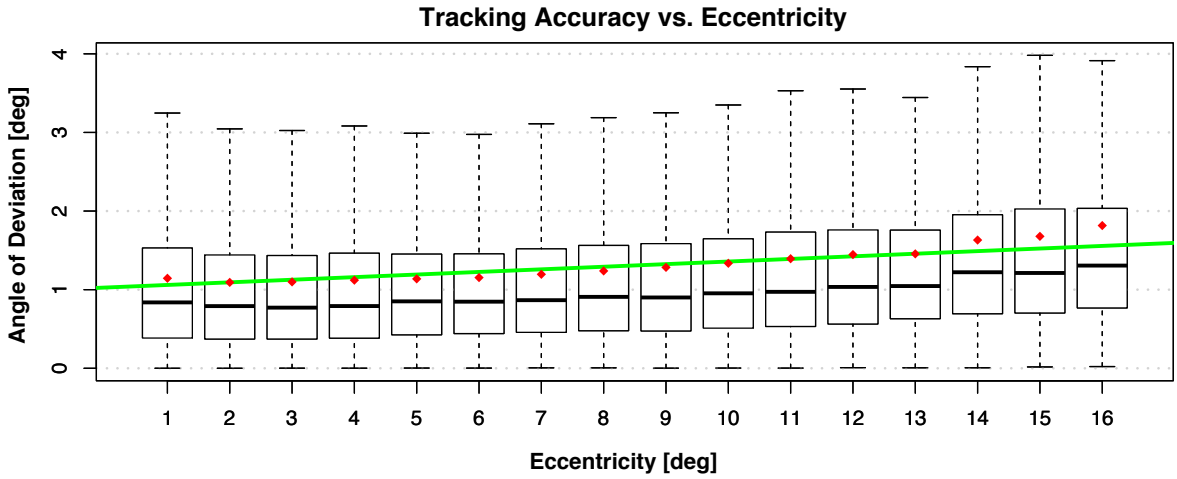


Figure 96: The tracking accuracy shows how well participants could follow the calibration target through the scene wrt. to the target's eccentricity. It is computed based on the binned differences of the tracking target positions and the measured points-of-regard, obtained by the eye tracker. Red dots show the deviation's mean. The green line shows a linear fit through the binned means.

occur as the eyes may not be capable of keeping focused. While the target was translating into depth and free space, no training samples were generated until the object was placed on a surface again. Please note that the depth samples used for training the model were not taken from the tracking target surface itself but the mesh of the underlying scene. The tracking target behaves like a rendered 3D sphere and changes its size once it is further away from the camera in the correct perspective manner. However, it was rendered in a second render path and overlayed on top of the original image in order to prevent it from modifying the original depth buffer. Note, that the semi-transparent tracking target was only meant to give a notion of what to look at in order to steer the user's gaze. It did not influence any of the measures used in the feature set.

A total of 14 subjects (8 male/6 female, all with academic background and experience in VR) aged between 22 and 50 ($M = 33$, $SD = 7.09$), participated in the experiment. All participants reported having a normal or corrected-to-normal vision ($< \pm 1\,\mathrm{D}$) without known serious visual impairments. After signing informed consent and receiving instructions, participants were seated and equipped with the HMD. Before presenting each scene, the participants were asked to perform a spatial calibration provided by the Fove SDK. Following that, the 18 paths per scene where shown to the subjects, where they had to track and fixate the moving target throughout the experiment while feature sets were collected.

7.3.2 *Results and Discussion*

In this section, the evaluation of the data collected in the user experiment is presented. Initially, the eye tracker's spatial tracking accuracy is determined. Figure 96 shows the angle of deviation of the tracking target's position to the measured PoR for various eccentricities. Based on this measure, the tracking accuracy is determined to be roughly 1° of the visual field, getting slightly worse with increasing eccentricities.

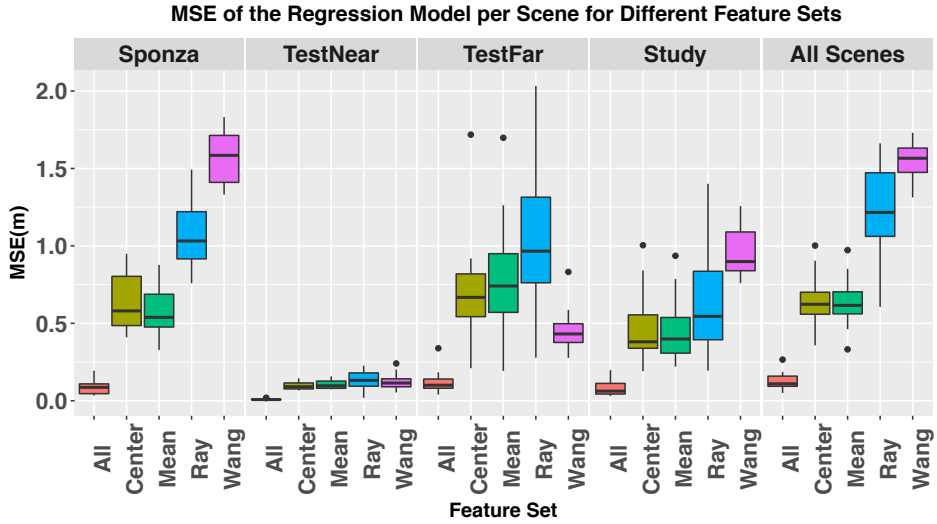


Figure 97: Mean squared error (MSE) in meters as the deviation from the ground truth for the calibration per user with different features for every single scene and for all scenes using 50% of the collected data for training the regression model and 50% for testing.

RQ1: ACCURACY OF THE RESULTING MODEL. In order to demonstrate that the model presented improves the depth estimation, the Mean Squared Error (MSE) of the regression model in meters for each user individually, per scene, and for all scenes, were computed. This was either performed using the proposed feature set (*All*) or using only individual features (*Mean*, *Center*, *Ray*, *Wang*). In order to get a fair comparison, the *Eccentricity* feature of the tracking target was used for all combinations. The MSE was estimated by training the regression model using a Leave-Group-Out-Validation (LGOV). For each user and each scene, 50% of the acquired gaze data was used to train the SVM and the remaining 50% to test it. This was repeated ten times with randomized subsets. Figure 97 shows the averaged MSE in meters as the deviation from ground truth including the data from the repetitions of all participants.

The figure reveals that the combined feature set (*All*) provides a substantial improvement over using individual features only. For all scenes, the MSE could be reduced to less than 50% as compared to using a single feature only, with a MSE of the combined feature sets of 0.1 m for *Sponza* and 0.01 m for *TestNear*. In general, the methods appears to perform better for scenes that have a limited depth range like *TestNear*. The most important reason for this is the fact that the impact of inaccurate predictions on the error is less for limited depths. Also, the vergence provides more accurate results if the target is close to the user, which is most probably the case in scenes with a small depth extent. However, the purely vergence-based estimates, *Ray* and *Wang*, perform worse for scenes that have a great extent of depth as estimates become less accurate if the tracking target is positioned at greater distances.

Figure 98 shows the accuracy of the depth as intervals of 0.25 m, starting at a distance of 0.25 m. This plot is continued over the entire range of 6 m. This distance is assumed to be critical in order to estimate the accommodative state of the eye (Section 2.2.3). For each user, 50% of the samples acquired per scene for that particular user were used to train the regression model. The remaining 50% were used to test the model computing the deviation from the ground truth. Again this was performed for ten repetitions with randomized training and test sets. Although the median errors of the model's accuracy are almost constant over the

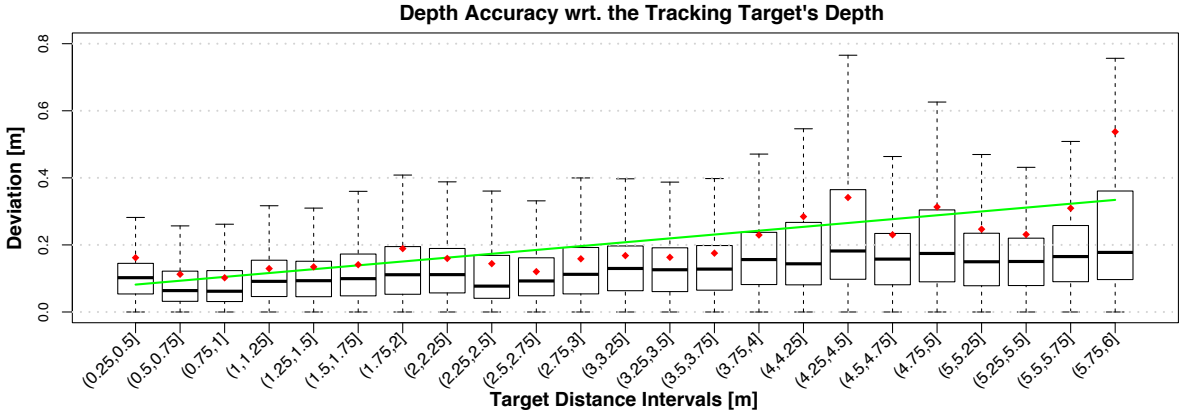


Figure 98: The depth accuracy showing how well the regression model predicts the calibration target's depth in meters wrt. the target distance. The figure is based on cross-validating the collected data. Red dots show the deviation's mean. The green line illustrates the linear fit of the means.

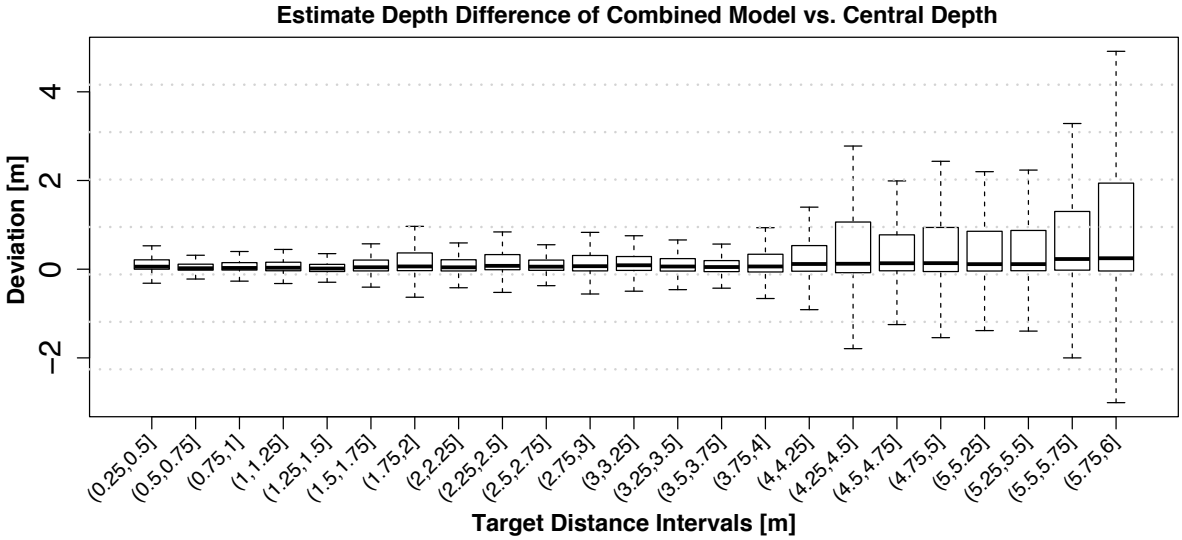


Figure 99: Depth differences comparing the combined model to a regression model that only uses the depth at the point of regard obtained with a standard ray casting approach. If the difference in depth is negative, the predictor trained using only the central depth is closer to the actual target by that amount. If it is positive, the combined feature set is more accurate.

entire depth range of 6 m with a deviation of approx. 0.1 m, the mean deviations are becoming worse with increasing distance. Due to a higher spread of samples, the mean deviation from the target (Figure 98, red dots) ranges from 0.08 m at a distance of 1 m up to a deviation of 0.5 m at a distance of 6 m.

Using the same LGOV scheme, a unit-less measure of the importance of the individual features for the prediction can be computed [Kuh17]. For the entire depth range of 6 m, the mean (95.03) and central depth (84.39) are the most important measurements, with the importance of *Ray* (43.86) and *Wang* (40.1) being almost identical for the model. Surprisingly, the eccentricity (11.43) is even less important than the variance (15.67). Keeping in mind that uncalibrated vergence measurements are trained, the eccentricity of the target from the

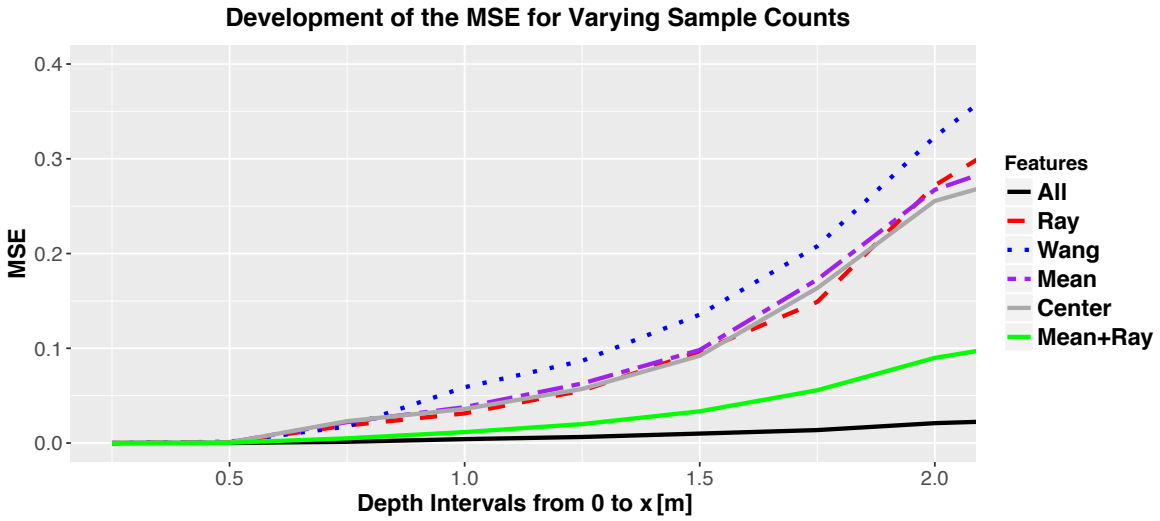


Figure 100: The averaged MSE per user learned from varying the intervals of samples to consider. The proposed feature set (All) outperforms purely vergence-based measures (*Ray* and *Wang*).

screen's center was expected to greatly influence the quality of the vergence-based measurements. According to the data presented, this is not the case.

The standard approach to estimating the gaze-depth in a head-mounted device is sampling the depth at the PoR using ray casting or sampling the depth buffer. Figure 99 illustrates the difference in depth estimates that result from using the central depth alone as opposed to using the proposed combined feature set. In order to compute these values, the model was trained with 50% of the data and predictions were made with ten repetitions. The distance of the predicted values to the ground truth, i.e., the depth of the displayed tracking target, was computed. If the difference in depth is negative, using only the central depth at the PoR is closer to the actual target by that amount. If it is positive, the combined feature set is more accurate. Figure 99 illustrates that using the depth at the PoR's center only, without considering the other features, leads to a decrease in the tracking accuracy over the entire depth range, especially at the far end of the range between 4 m and 6 m. Nonetheless, in summary the evaluation shows that using the proposed feature sets for depth estimation improves accuracy substantially.

RQ2: TARGET DEPTH AND THE ACCURACY OF VERGENCE ESTIMATES. The proposed combined feature set outperforms the purely vergence-based measurements. However, it is questionable to which depth the vergence-based estimation is equally accurate or perhaps even better. Figure 100 shows the regression model trained for increasing depth intervals. First, all samples individually per user ranging from 0 m to 0.25 m were used to train and predict values, following the introduced LGOV scheme. This is performed individually for the participants and scenes, with ten repetitions. The averaged MSE is computed including all repetitions, users, and scenes. Next, the interval from 0 m to 0.5 m is processed. This evaluation is repeated until the last interval (0 m to 6 m) is reached. The plot of the computed average MSEs is illustrated in Figure 100. It becomes clear that after the first 0.5 m the vergence-based measure is becoming increasingly inaccurate. This correlates with the related work on vergence-based depth estimations. Wang et al. [Wan+12] investigate the tracking accuracy using targets ranging from 0 m to 0.5 m. However, it has to be noted that they use

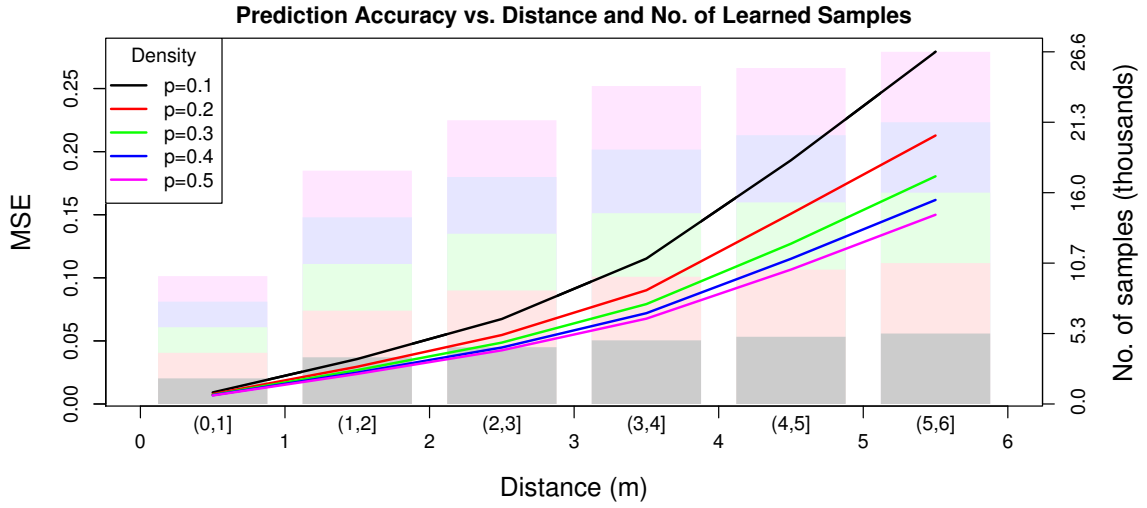


Figure 101: Prediction Accuracy (MSE) plotted vs. distance for different sampling densities p. These determine the number of samples taken into consideration for the learning process.

a more elaborate method to calibrate the vergence estimates using Periodic Self-Organizing Map (PSOM) [WNR00] including the spatial position of the current PoR. Interestingly, Figure 100 also shows that the image-based measures computing the central depth and mean depth perform comparably to the vergence measures over the examined depth range. However, for these direct depth measurements, this comparable behavior results from the increased probability of the eye tracker missing the correct position of the focused object at greater depths: an effect caused by the reduced projected size of distant objects. Figure 100 also shows that combining a vergence-based (*Ray*) and an image-based measure (*Mean*) diminishes the effect. Combining features decreases the introduced error substantially (*Mean+Ray*).

RQ3: EFFECT OF THE TRAINING SET SIZE. Calibration is a cumbersome process and might negatively influence the user experience if the process takes too long. Therefore, it is essential to describe how the accuracy of the prediction is influenced by the number of samples used to train it. Figure 101 shows how the prediction accuracy improves when using an increasing amount of samples for the training process. The plotted lines show the development of the MSE for data in the range of up to n meters on the x-axis. The number of samples taken into account was chosen to be a percentage p (*sampling density*) of the available samples, namely 10% to 50% in steps of 10% as larger sample sets lead to very long training times. The measurements were performed by learning the respective percentage p of the available tracking data from all scenes at once, but for each user individually. Depth estimates were then computed for all available tracking data of all participants and errors were averaged, as shown by the lines. The resulting number of samples for each sampling density is shown by the bars in the background for each depth interval. The plot shows clear improvements in the prediction accuracy when a higher number of samples is used for training. However, these improvements happen in a somewhat logarithmic fashion. Assuming that the calibration process is not a one-off task, but has to be performed on a regular basis, it is necessary to limit the time it takes to perform this task. For each scene, collecting the data using the calibration paths took each participant approximately five minutes. Judging from the analysis, it is assumed that reducing this time by 50% would indeed be possible,

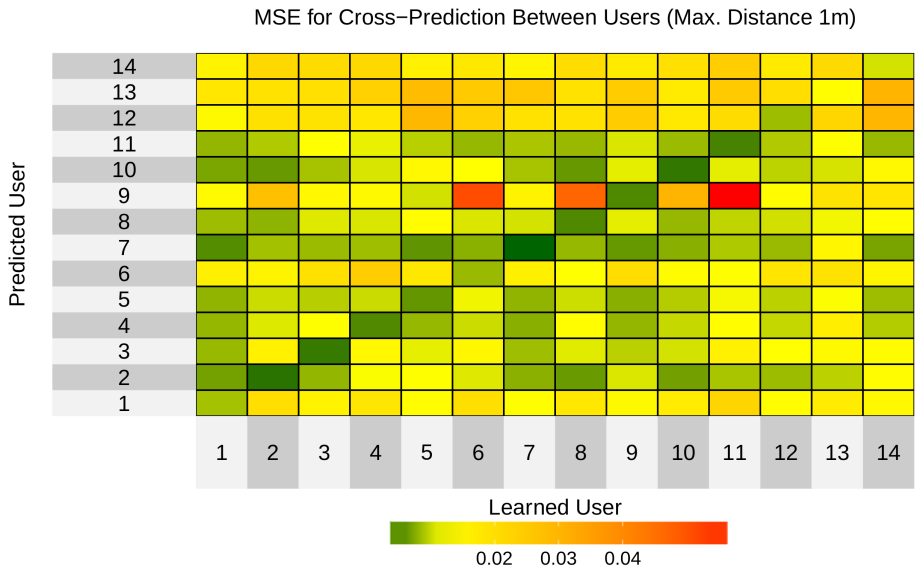


Figure 102: Mean square errors for cross-prediction between users for the first meter only. Data was learned for each user individually and then predicted for all subjects.

especially if the main requirement is a high accuracy for low distances. It also has to be kept in mind that there is in fact an unknown dependency between chosen fixation paths, viewpoints, and the accuracy of the predictor. This should be further analyzed in future work. Gaining knowledge on the influence of the shape of chosen fixation paths regarding prediction accuracy could enable further reductions of required sample set sizes.

RQ4: EFFECT OF USING THE CALIBRATION ACROSS USERS AND SCENES. As the calibration process can still take some time it would be desirable to have a *universal* calibration that does not require each user to run the calibration process every time they use the system. In order to test if one user's trained data can be used to predict another user's target, physiological properties such as differences in the IOD need to be taken into consideration. This made it necessary to add a stage where the target depth was predicted solely based on vergence-related features (Wang, Ray, and Eccentricity). Essentially, the data is first split into two distinct sets. All vergence-related features are then used to train and predict the target depth. This prediction is used as a new feature for the second training step together with the *mean depth*, *center depth*, and *variance*. The goal of this approach is to reduce the dependency of features on the physiological properties of a user. The resulting model is then used to predict the gaze-depth of each of the 14 participants. It should be noted that all computations in this evaluation are only performed for recorded target distances of up to 1 m. The results are shown as a heatmap in Figure 102, with learned subjects on the horizontal axis, predicted subjects on the vertical axis and the MSEs of the cross-prediction being color-coded. Unfortunately, this reveals that a reliable cross-prediction between subjects is indeed difficult. While not having identical calibration paths for all participants may be one of the causes, choosing identical paths would not be a realistic scenario: Eye movements in the prediction phase will not depend at all on a fixation path in real-world applications.

Even though using the calibration across subjects does not yield the same accuracy as training participants individually would, analyzing the accuracy across scenes is still worthwhile.

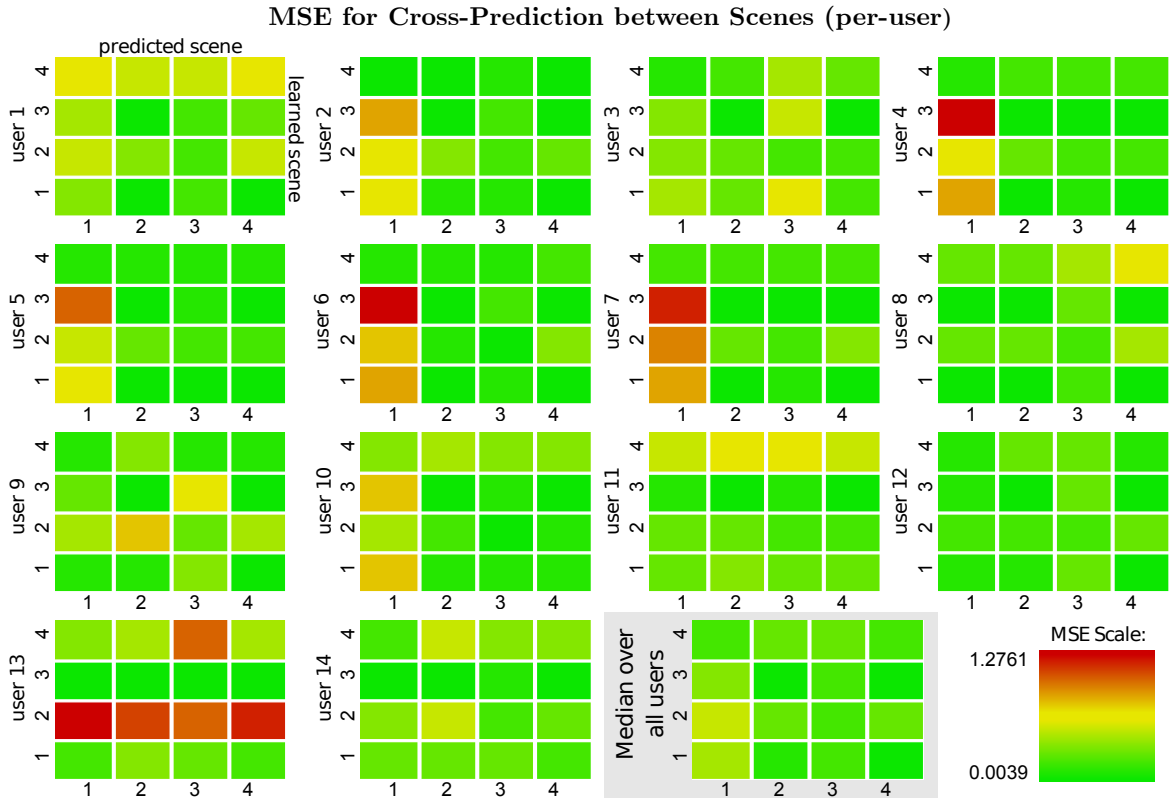


Figure 103: The mean square error (MSE) in meters when training each of the 14 individual users across scenes (1=Sponza, 2=TestNear, 3=TestFar, 4=Study). Rows represent learned scenes, columns represent predicted scenes. A total of 50% of the data per trained scene was used for building the model. The MSE was computed for all training samples of the respective predicted scene.

Figure 97 illustrates the MSE computed using the 50% LGOV of data to train the regression model per user, but for all scenes. By looking at the results for all features, it becomes apparent that training across scenes does not result in substantially worse MSEs, leading to the assumption that the calibration does work well in such a scenario. Hence, the approach was evaluated using a Leave-One-Scene-Out (LOSO) scheme. This way the model was trained with the same number of samples that would have been used to train a single user, i.e., 50% of the number of an average user, but this time the training samples were selected from all scenes except one. Following that, the model was used to predict the scene that was omitted. In order to make a comparison possible, the MSE when predicting and learning from the same scene was computed yielding a ground truth. Interestingly, the LOSO scheme provided lower and thus better MSEs for *Sponza* (0.164 vs. 0.244) and *TestFar* (0.160 vs. 0.234) but higher and thus worse for *Study* (0.155 vs. 0.102) and much worse for *TestNear* (0.202 vs. 0.008). This difference can be attributed to the vastly different characteristics of the test scene *TestNear*, which is very limited in depth. For this scene, most predictions are dependent on the vergence information. However, using the LOSO scheme, samples were drawn from scenes where the depth could potentially be larger and fewer training samples were generated for the extreme near field. For further investigation the method's cross prediction capabilities were evaluated for each participant. The results are presented in Figure 103. Here, the regression model was trained for each scene and participant individually, and predictions were made for all scenes

for the respective subject. Rows represent the learned scenes, while columns represent the predicted scenes. The heatmap on the lower right shows the median error values from all 14 participants. It can be noted that apart from Subject 13, even Scene 2 works well when used as training data for predicting other scenes despite its lower depth extent. For vice versa, the same is true, as Scenes 1, 3, and 4 lead to a good prediction of Scene 2 in most cases. It is also important to note that there are some outliers, (e.g., for Subject 13) leading to a significant increase in the total MSE, even though the system works well for most users. In summary, Figure 103 shows that cross-prediction across scenes works well which is in direct contrast to cross-prediction across participants. This means that the performed calibration data is not only valid for a specific scene, which is supported by the low errors illustrated on the diagonal in Figure 103, where the data was also learned independently of the scenes.

The evaluation shows that the presented approach that combines multiple measurements to estimated gaze-depth is superior over using single measurements only. A tracking accuracy with a median error of 0.1 m and a mean of 0.1 m to 0.5 m in the critical depth range of 0.2 m to 6 m could be achieved. In contrast, vergence-based measurements were only accurate up to a focused distance of 0.5 m. However, the quality of the machine learning approach depends on the number of training samples, and there is an unknown dependency between chosen fixation paths, viewpoints, and the accuracy of the predictor. Also, while calibration accuracies do not depend on the scene, the calibrated models should not be used across users.

7.4 USER STUDY - DEPTH-OF-FIELD

Having developed and evaluated the system to obtain gaze depths, this section presents the results of a user study in order to evaluate the perceptual quality and implications of the presented gaze-contingent DoF filtering framework. It was driven by the following research questions:

- **RQ5:** Does gaze-contingent DoF conceal visual artifacts?
- **RQ6:** Does gaze-contingent DoF increase depth perception?

7.4.1 *Procedure and Apparatus*

In order to evaluate the foveated rendering system with gaze-contingent DoF, the same Fove 0 Headset as in the previous experiments was used. This time, the study consisted of three parts. First, users were asked to put on the HMD and do the spatial eye tracking calibration provided by the Fove SDK. Following that, the calibration of the gaze-depth estimation was performed as described in the experiments in the previous section. In order to collect the training data in reasonable training time, only six randomly generated calibration paths per participant were used. This took about three minutes. Although higher accuracies for gaze-depth estimation are to be expected for bigger training set sizes, this value was selected to balance accuracy, time, and the estimated subjects's patience. The main part was conducted as a within-subject study, employing a $4 \times 3 \times 2$ full factorial design with four DoF *settings*, three *focus modes*, and two *rendering modes*. Trials were generated with two repetitions and were randomly shuffled resulting in a total of 48 trials per participant. A single camera position

was chosen for all trials using the scene shown in Figure 94. The scene was presented with four different DoF settings (*DoFMode* = *NONE*, *WEAK*, *MEDIUM*, or *STRONG*) (Figure 94) and was designed to contain multiple labeled targets (spheres and boxes). Targets within a range of 0.5 m to 1 m were labeled as "1". Targets within a range from 1 m to 6 m were labeled as "2". For each trial the participants were either asked to fixate targets labeled "1", "2" or to freely look around in the scene, thus corresponding to the factor levels *FocusMode* (NEAR, MID, or FREE). As the influence of the DoF is scene and focus point dependent, this ensured a wide variety of objects at different depths were focused on by the user. Moreover, the scene was tested with full ray tracing (*FoveatedMode* = *FULL*) vs. the presented foveated mode (*FOVEATED*). For the latter the same settings as used in Section 7.2 were selected for the ray generation, resulting in 558,945 updated samples per frame, regardless of the DoFMode. Each configuration was presented for 6 seconds. After each trial the participants were presented the following statements:

- **Q1:** There were no visual artifacts in the periphery.
- **Q2:** The visual artifacts were not distracting.
- **Q3:** I could focus on scene elements based on my gaze reliably.
- **Q4:** Rate the intensity of depth perception.

Q1 to Q3 were rated on a 7-point Likert scale from -3 (strongly disagree) to 3 (strongly agree), while Q4 was rated on a numerical scale from -3 (no depth perception) to 3 (strong depth perception).

7.4.2 *Results and Discussion*

The study was performed with 12 participants (7 male/5 female, all with academic background) aged between 25 and 50 (M = 35, SD = 7.4), who reported to have normal or corrected-to-normal vision ($<\pm 1$ D) without known serious visual impairments. Plots of the ratings for Q1–Q3 are presented in Figure 104. The mean ratings for depth perception are shown in Table 10. As presented in a previously published paper [Wei+18a], an Analysis Of Variance (ANOVA) on the data was carried out. Here, a nonparametric, rank-based ANOVA approach from R's ARTool package [KW17] was used, as Levene and Shapiro-Wilk tests show that the data's homoscedasticity and normality cannot be relied upon. Performing a 3-way ANOVA with factors *FoveatedMode*, *FocusMode* and *DoFMode* and accounting for interactions produced the results presented in Table 11. Post-hoc tests have been carried out using F-tests with Holm's method for p-value adjustment.

The ratings presented in Q1 and Q2, as well as the depth perception rating in Q4 were filtered using the focus reliability rating illustrated in Figure 104c. If the trial is rated below zero and participants could not focus reliably (*Rather Disagree*), it was removed. This way the influence of inaccurate tracking and depth estimation results was limited. In contrast to the previous experiment to determine the quality of the gaze-depth estimation, this user study is only intended to provide insights into the quality of the DoF filter. Looking at Figure 104c as well as the results from the ANOVA, it can be seen that there is an interaction between the focus reliability and the DoF mode. As the intensity of the DoF effect is increased, people are less likely to tolerate below-perfect gaze-depth estimates. Even if the estimate is close to the

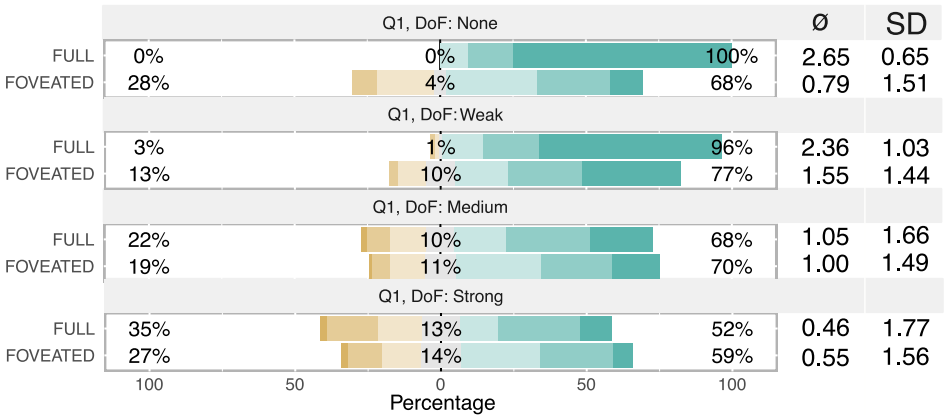
focused depth, stronger DoF modes will most probably blur the image in regions which the user expects to be in focus. Moreover, we observed that participants were unforgiving when the DoF effect did not change with the correct speed. However, the speed of accommodation and its range are dependent on a variety of influences (Section 2.2.3). The effects are very individual. Nevertheless, it is worth looking at the filtered ratings for the perceptibility and distractiveness of artifacts (Q1 and Q2).

RQ5: DIFFERENTIATION BETWEEN FOVEATED AND NON-FOVEATED RENDERING. Studying the results of the Likert ratings presented in Figure 104a, a shift in ratings between the various levels of *DoFMode* can be observed. Although the highest visual quality is reported for full rendering without using DoF, the ratings between foveated and full rendering become increasingly similar with an increasing strength of the DoF effect. This becomes apparent when comparing ratings for *DoFModes* factors *WEAK*, *MEDIUM* and *STRONG* for both full and foveated rendering. Interestingly, the mean Q1 ratings for *DoFMode STRONG* were even higher for foveated than for full rendering. The existence of such differences is also apparent from the interaction between factors in the ANOVA shown in Table 11. While the results presented for Q2 in Figure 104b show that an increase of the DoF with changing *DoFMode* factor levels result in worse ratings for full rendering, foveated rendering clearly benefits from a *WEAK* DoF effect in terms of an improved artifact reduction. This positive effect becomes especially apparent when comparing the means. Nonetheless, the results shown in the figure provide strong evidence that both ratings are similar. Since visual artifacts were least noticeable and disturbing for *DoFModes WEAK* and *MEDIUM* with foveated rendering, it can be assumed that the DoF mode successfully conceals artifacts. Unfortunately, enabling DoF appears to negatively influence the perceived quality using full ray tracing. In addition, perceived image quality worsens with increasing DoF intensities for both full and foveated rendering. The loss in apparent image quality coincides with previous works [Duc+14; VAF16].

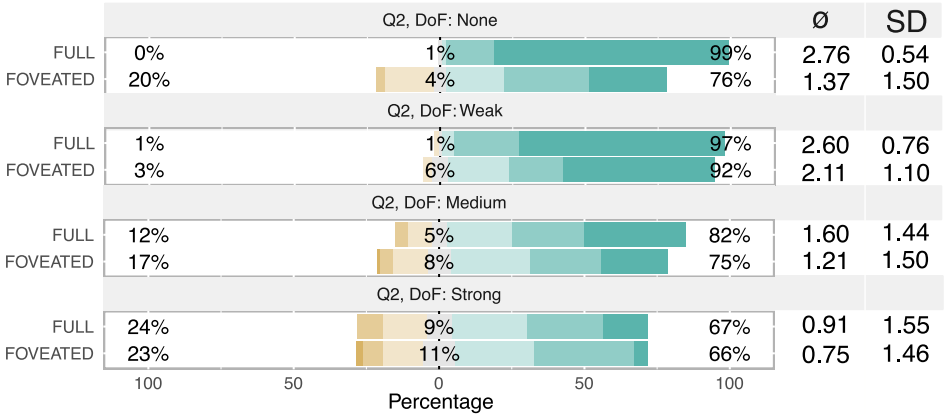
RQ6: EFFECT ON DEPTH PERCEPTION. With the ANOVA showing statistical significance for the differences in depth perception between the various levels of *DoFMode* ($p < 0.05$, $F(3, 484) = 5.6$), Table 10 shows the corresponding means and standard deviations. While the lowest DoF setting shows the highest depth perception rating, differences between the various factor combinations are small. In a questionnaire after the study, the participants were asked for their level of agreement with the statement that DoF could increase depth perception. Here mixed results could be observed (M=1, SD=1.13). While five of the participants rated their level of agreement as neutral, eight users rather agreed with the statement. While it is certain that depth perception depends on the tracking accuracy, the response to the DoF effect appears to be highly individual. Depth perception is rated best for *DoFMode* = *WEAK*, with decreasing ratings in the order *MEDIUM*, *NONE*, *STRONG*.

Although the results show that the overall visual quality is reduced with respect to the chosen *DoFMode*, a *WEAK* setting has a positive influence on the amount and the subjective distraction caused by artifacts in the peripheral visual field. At the same time, for this setting, the overall visual quality remains relatively high, while a slight increase in the overall sensation of depth could be observed.

a: Q1: There were no visual artifacts in the periphery



b: Q2: The visual artifacts were not distracting



c: Q3: I could focus scene elements based on my gaze reliably

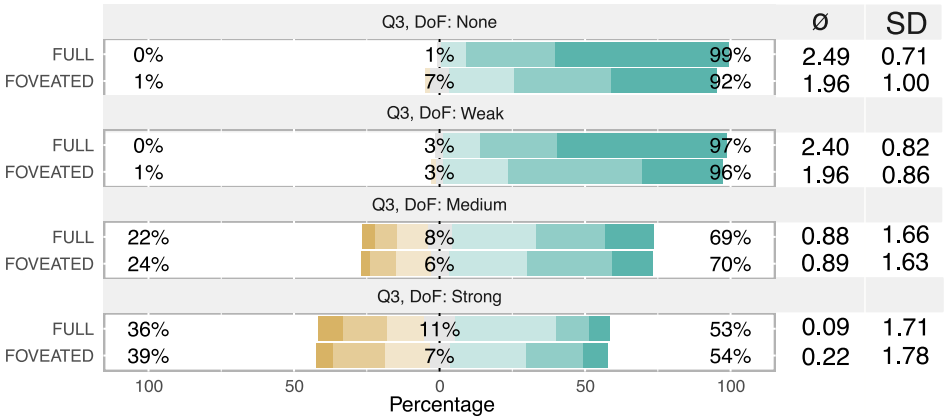


Figure 104: Likert scale ratings of Q1-Q3 to provide insights into the visual quality of the DoF filter.

		DoF			
		NONE	**WEAK**	**MEDIUM**	**STRONG**
FOVEATED	Mean	1.75	2.06	1.82	1.70
	SD	0.73	0.63	0.83	0.85
FULL	Mean	1.81	2.03	1.94	1.74
	SD	0.88	0.73	0.80	0.91

Table 10: Mean and SD for Q4 "Rate the intensity of depth perception" for all level combinations of *FoveatedMode* and *DoFMode*.

Q1	**FoveatedMode:DoFMode**	$F(3, 484) = 11.78$	
	FOVEATED - FULL	NONE - WEAK	$F(1, 484) = 13.7$
		NONE - MEDIUM	$F(1, 484) = 23.92$
		NONE - STRONG	$F(1, 484) = 25.16$
Q2	**FoveatedMode:DoFMode**	$F(3, 484) = 7.72$	
	FOVEATED - FULL	NONE - WEAK	$F(1, 484) = 19.16$
		NONE - MEDIUM	$F(1, 484) = 8.82$
		NONE - STRONG	$F(1, 484) = 12.82$
Q3	**FoveatedMode:DoFMode**	$F(3, 484) = 2.73$	
	FOVEATED - FULL	NONE - WEAK	$F(1, 484) = 3.87$
		NONE - MEDIUM	$F(1, 484) = 11.34$
		NONE - STRONG	$F(1, 484) = 15.88$
Q4	**DoFMode**	$F(3, 484) = 5.6$	

Table 11: Significant results ($p < 0.05$) for the performed ANOVA. Main effects were omitted if significant interactions were present. For Q1–Q4 significant results are shown as main effects or interactions together with their difference of differences (DoD) for significant factor levels.

7.5 FUTURE WORK

While the evaluation has shown the potential of gaze-contingent DoF, there are several directions that future work could take. As user studies are time-consuming, it is worthwhile computationally measuring the influence of DoF on various artifacts with different scenes for example by using a wavelet analysis as presented by Patney et al. [Pat+16b]. The author is confident that using knowledge about the DoF is crucial when scheduling potentially salient regions for (re-)sampling. Another possible path is to further reduce the number of cast rays by considering DoF progressively. Although the system can already re-sample the scene multiple times employing knowledge of the CoC, in this case the necessary frame rates in order to meet the V-Sync limit of the HMD could not be achieved. However, meeting these requirements is essential to reduce fatigue and to cope with fast eye movements [Alb+17]. This way, it may well be possible that the perceived visual quality using a DoF filter can ultimately exceed full rendering modes while achieving equal or even lower render times at lower sampling rates.

Besides the positive effects on the computational complexity, better and more precise eye tracking solutions or methods that apply scene as well as object knowledge such as the method

by Mantiuk et al. [MBM13], that allows more accurate gaze-depth estimates to be derived, will extend the applicability of the approach presented. More elaborate physiological models on the speed of accommodation might further improve perceived quality.

Surprisingly, the variance of the samples appears to have a higher influence on the accuracy of the proposed model than the eccentricity of the PoR. Especially for challenging situations, such as looking at the edge of a significant depth discontinuity, the variance of the samples is vital. Nonetheless, it has been assumed the eccentricity is more influential because of its direct connection to vergence. Future work will need to investigate how calibration paths, scenes, and cameras should be chosen that are ideal for such a process. Collecting the data took approximately three to five minutes to obtain per scene and user, and training the model with only 50% of the samples resulted in low MSEs. This shows that such a scene-based calibration is feasible in consumer-level devices as actual virtual content can be used for calibration. This improves the user experience of the process. Nonetheless, the number of calibration samples had to be reduced and so also the duration of the procedure for the user study to evaluate the DoF filter, compared to the number presented in the evaluation of the tracking accuracy (Section 7.3). However, the highest achievable accuracy will possibly only be reached with specifically designed scenes and paths. Training the SVM with 50% of the trained samples took a few seconds, whereas training huge datasets, for example all data collected per user, takes between several minutes and several hours. If learning such larger amounts of data is required, choosing other machine learning techniques such as linear SVMs or neural networks is to be recommended.

Regarding the intensity of depth perception, it may be worthwhile to take a closer look at the influence of artifacts and over-blurring in the visual periphery. Also, investigating methods that render to multifocal and lightfield displays, so enabling a more natural presentation of the rendered content to the user [Pad+17; Mer+17], might be another field that can greatly benefit from sampling strategies that take the optical limitations of the HVS into account.

7.6 CONCLUSION

In this chapter a gaze-contingent rendering and filtering approach exploits knowledge of the retinal and optical abilities of the HVS to accelerate rendering. Samples can be reduced by 69%, and gaze-contingent DoF has been proven to be a viable solution for concealing rendering artifacts. A user study has shown that quality ratings between foveated and full rendering were almost identical using gaze-contingent DoF, although the visual quality is slightly reduced.

Filtering using DoF is made possible by employing an estimator for the gaze-depth using an eye tracker inside an HMD. In a calibration step, eye tracking data was recorded and then analysed to compute different depth features. These were then consolidated into a feature set in order to train an SVM for depth prediction. It has been demonstrated that the proposed method provides a high accuracy regarding the mean deviation from the reference depth of 0.1 m for the first 2 m up to 0.3 m for targets at a distance of 6 m. In contrast, solely vergence-based estimates are only precise for the first 0.5 m. Once trained this model works well when used across scenes but becomes less precise when applied across users. As determining the PoR should lead to similar results for all users, it is presumed that the decrease in performance

is caused by an imprecise vergence calibration, especially since the IOD and the distance of the eyes to the display was assumed to be fixed. More accurate results across users are to be expected if more work is invested into calibrating the vergence and if the individual IOD is taken into account. This also affects re-wearing the HMD, as this might significantly reduce the tracking and estimation accuracy. Even the slightest movement and slipping of the HMD on the user's head might result in inaccurate tracking. However, as the computations rely on very precise data to accurately derive the focused depth, it is anyhow beneficial to perform a calibration per user each time before rendering a new scene as is also advisable for the spatial calibration process.

As DoF is essentially a guided low-pass filter applied to the image, it is useful to hide high-frequency artifacts that are challenging for peripheral vision. However, the influence of the DoF effect is scene-dependent, as artifacts are only reduced for areas of the scene that are out-of-focus. Hence, ultimately, for optimal results, several strategies for foveated rendering need to be combined. Filtering the image using DoF is just another tool for perception-driven rendering. Knowing the size of the CoC allows the image in regions that are in focus to be resampled. Using PP-Interpolation already allows samples to be integrated based on perceptual requirements, such as image saliency. This way, it can be argued that DoF will become another essential factor when deciding which regions or pixels need increased computational effort. Flexible and faster ray casting and ray tracing solutions like NVIDIA RTX [Sti18] and HVVR [HMN18], as well as new hardware generations, will provide the necessary computing power. This also makes it possible to study fully dynamic scenes, which is as yet too slow for HMDs in the current implementation.

Part IV

EVERYTHING MUST COME TO AN END

There are things known and there are things unknown, and in between are the doors of perception.

Aldous Huxley (⋆1894 - †1963)

The eye sees only what the mind is prepared to comprehend.

Tempest-Tost
William Robertson Davies (⋆1913 - †1995)

It's all in the mind.

George Harrison (⋆1943 - †2001)

8 FINAL WORDS

A close understanding of the Human Visual System (HVS) and the various processes of perception has tremendous potential to improve the quality, the speed of generation, and the comprehensibility of computer-generated images. For efficient rendering techniques, the central question is *how* to exploit the limitations or use the potentials of perception to enhance the quality of a method whilst maintaining its performance and vice versa. This thesis showed some of the limitations and potentials as well as how to exploit them in modern rendering systems. In this final chapter we summarize the major contributions, the results of this work, and discuss future work.

This thesis has presented a wide variety of physiological and perceptual limitations of the HVS and the models to describe these. Building such models is an active field of research that dates back centuries (Chapter 2). Thanks to the growing understanding of the physiological components, the availability of better optical appliances and instruments, high-quality medical imaging as well as an ever-increasing mathematical and physical toolset, perception research is constantly progressing. In this field, user studies and experiments are fundamental tools to increase the understanding of limitations and potentials of the HVS. A large number of the most common models and systems that have proven themselves in the graphics community were presented here (Chapter 3). By discussing the most fundamental concepts of efficient rendering, the state-of-the-art in perception-driven accelerated rendering has been presented (Chapter 4). This thesis shows how existing methods such as perception-driven Level-of-Detail (LoD), sampling, or methods that exploit temporal coherence can be combined into new rendering systems. Based on these considerations, this thesis has described our efforts to push the boundaries of perception-driven rendering further. Here, we have specifically targeted view- and gaze-contingent approaches that use active measurements of the visual system.

Large-scale projections and Head-Mounted Displays (HMDs) allow the visual perceptual channel to be addressed more directly. However, images must be rendered in high resolution and possibly in real-time in order to take full advantage of such systems. Nevertheless, meaningful 3D models that use the advantages of such display systems are visually complex. When rendering such models, aliasing artifacts greatly limit the visual quality. Therefore, methods are needed that dynamically adjust the rendering quality to increase the visual quality but at the same time maintain low render times. To this end, we have introduced a Graphics Processing Unit (GPU)-based voxelization and octree construction scheme to build a Hybrid Sparse Voxel Octree (HSVO), combining voxel and polygonal information (Chapter 5). This method has the potential to aid image quality, especially for large outdoor scenes. Here, a voxel representation for distant objects might provide a sufficiently high visual quality but, at the same time, can significantly reduce aliasing artifacts. Images rendered with four sample-per-pixel (spp) using the HSVO achieve quality ratings that compare to $8 - 16$ spp images rendered with triangles only. The detailed evaluation of the visual quality provides valuable

insights when evaluating flickering artifacts and noise resulting from undersampling artifacts. Likewise, more robust tracking solutions have become an enabling technology that allows active input to be incorporated into image synthesis approaches. Therefore, we presented an approach that adapts rendering to the user's field of vision in front of a large, high-resolution display wall. Here, insights gained from the evaluation and the user study paved the way for foveated rendering systems presented in the following chapters.

We presented two methods that adapt sampling to the retinal capabilities, i.e. *foveated rendering* (Chapter 6), including a method to filter potential artifacts using the eye's inherent Depth-of-Field (DoF) when using HMDs (Chapter 7). This research shows that a deep understanding of perception is becoming increasingly important when designing Virtual Reality (VR) and Augmented Reality (AR) systems, starting with the necessary display and eye tracking hardware through to the rendering techniques that synthesize images. Besides, the presented approaches would not have been possible without efficient ray tracing cores. Research from this thesis shows the relevance of resampling for perception-driven rendering pipelines, and ray tracing cores gives the freedom to do so efficiently. In addition and in order to maintain a high visual quality, all of the foveated rendering approaches that were developed for this thesis heavily exploit Temporal Coherence (TC) between subsequent frames. Our methods demonstrate how ray tracing can be used to sample the image to resolve reprojection artifacts selectively.

This thesis has presented the results of several user studies and all show the limits and potential of the respective methods. Foveated ray tracing with forward reprojection allows for the reduction of the number of sampled pixels by 79% (Chapter 6) retaining a visual quality that is on-par with full ray tracing. This makes speedup factors of two to three possible compared to rendering all pixels of the image and not considering the retinal limitations. Now, interactive ray tracing becomes possible, even for demanding rendering processes for VR and AR that require frame rates exceeding 60 Hz.

Likewise, we showed that the process of filtering images with gaze-contingent DoF filters has the potential to reduce the complexity of rendering pipelines but also to increase the presence in the virtual world (Chapter 7). Knowledge about the design, the execution, and evaluation of user studies and experiments are valuable skills when developing more efficient perception-driven image synthesis approaches. The analysis of user studies with statistical tools such as with and Analysis Of Variance (ANOVA) is needed to draw conclusions from the data and gain more insights into the connection of different factors.

Also, this thesis demonstrated new ways to use and analyze eye tracking data. For example, this thesis presented a method to derive more accurate gaze-depth locations using machine learning techniques on multiple gaze-depth measurements (Chapter 7). This makes it possible to estimate gaze depth with an average error of about 0.2 m for points in the critical depth range of 0 m to 6 m. In addition, interesting conclusions from user studies and research experiments can be drawn when analyzing eye tracking data. This way, we could show how solving tasks significantly influences the way we perceive our surroundings. Such algorithms, that target inattentional blindness and visual tunneling, bear another great potential: They allow for the next level of rendering optimization. Here, additional knowledge about other factors such as retinal velocity is important when putting results into perspective.

High display resolutions enable the perception of finer grained details, and a wider coverage of the visual field. Large, high-resolution, projection-based displays as well as high-resolution

tiled display walls have already become well-established installations in research institutions around the world. In addition, the continuing interest in consumer level VR and AR technologies is driving the market for an ever-increasing range of head-mounted displays. Even today, rendering high-quality images for those devices at the necessary refresh rates is challenging. For hardware, strong trends towards novel displays can be recognized that allow rendering multiple views, lightfields, or holographic imagery. Also, retinal projection systems are advancing beyond the prototype stage. Multi-panel installations and adaptable lenses will create a more natural experience when viewing images in VR and AR headsets. These displays can help in increasing presence and in reducing fatigues as well as limitations like the vergence-accommodation conflict. Also, the field of computational displays that exploits the limitations of the HVS in order to improve devices beyond their specifications might result in devices with a better visual quality at a lower cost and within the technological constraints of the time. Here, tricking our senses to improve the perceived contrast and apparent resolution enhancements already enable image details to be shown that are beyond the physical limitations of the display device. Still, in the pursuit of the dream of an ever-increasing image quality, these approaches will further increase the requirements of rendering approaches. Hence, based on these developments and the results of this thesis, the author sees several promising research directions, which can lead to the development of even more efficient and more robust rendering algorithms.

In the times when ray tracing methods were too slow to be practical, interactive graphics had to rely on rasterization-based rendering technology. However, such techniques have limits when sampling the image plane selectively. Still, hardware improvements concerning perceptual algorithms, such as efficient pixel-precise shading and multi-resolution rendering will boost the efficiency of current and upcoming algorithms. However, we already see more flexible rendering pipelines with ray tracing hardware becoming available. In the author's opinion, dynamic real-time adaptive ray tracing permits more efficient and flexible sample schemes to be implemented and thus better fit perceptual requirements. Images can be re-sampled more readily in those parts that matter more to the user, either because they require the highest visual acuity, are in-focus, or because they are more relevant to the HVS. Currently, we are focusing on primary visibility by sparsely sampling the image plane. However, as the computational requirements for shading are getting higher and higher, foveation should take higher-order lighting effects into consideration. Determining primary visibility is relatively efficient compared to those efforts necessary to compute Global Illumination (GI). This will likely benefit more from an evaluation of an eccentricity-dependent contrast sensitivity rather than visual acuity. Also, it is worthwhile to incorporate visual masking when selecting regions that potentially need higher visual fidelity. Here, establishing the right toolsets to develop for such a system is of great importance. Tools to firstly identify perceptually critical regions, then select and re-sample images as well as filtering the final images (all this potentially iteratively and within a given time-frame) might need novel approaches in comparison to traditional rendering pipelines.

Ultimately, the transmission of the computed pixels and views at very high resolutions is also limited by the bandwidth of the interconnect. Here, new approaches to frameless rendering and display controllers may be one possible solution to drive such systems. For VR and AR applications it is the author's contemplation that a rendering method should ideally be asynchronous from the display and independent from a fix refresh. A display controller might be used to warp images to the correct perspective, while an asynchronous render

thread updates individual pixels based on perceptual requirements. This way, computational resources to increase visual quality can be distributed more flexible.

Besides that, the integration of low-latency eye tracking has led to a wide field of research that aims to address usability and performance requirements by designing gaze-contingent rendering and interaction methods. In addition, the extensive use of eye tracking may simplify the creation of large-scale gaze databases. Such databases could, in turn, lead to significant improvements in scan-path predictions and learning-based saliency methods. Unfortunately, current eye tracking devices often lack precision. Even if they are head-mounted, there is always the danger that the correct position of the tracker cannot be maintained. Putting HMDs on and taking them off when using integrated eye trackers poses challenges. The problem is the reproducibility of the accuracy without recalibration for the same user and the same HMD once the user has put the HMD off and on the head again. The problem even exists when the HMD slips due to a slight movement of the head. Hence, we suggest there is a need for online calibration procedures that subconsciously steer the gaze towards tracking targets in the virtual world to improve calibration progressively. In contrast, methods that allow scan paths, saccades, and landing positions to be predicted may be another solution to loosen the requirements of the tracking hardware. Likewise, not all parameters of the eye, such as its accommodative state, can be captured as readily as gaze points (Points of Regard (PoRs)). However, there is a need for methods or devices that can accurately capture the accommodative state of the eye to enable gaze-contingent DoF but also for display systems with an adaptable focus. One approach to improve gaze-depth estimates has been presented in this thesis.

In order to provide more comprehensive models of the HVS that can be evaluated more cost-effectively, we firmly believe in the potential of modern machine learning approaches. Deep learning has already shown great potential in modeling the HVS with higher precision and for general image data. Hence, it might provide a viable tool for more robust reference and non-reference metrics, leading to new quality measures and sampling guidance methods in the coming years. Also, estimates of saliency evaluations, the stimuli that drive our attention, could become more reliable. Current methods including attentional models demonstrated the potential to accelerate rendering. However, the success of these methods is dependent on a balance between implementation effort, attention model detail, and the available model knowledge, such as task description or the scene gist. Often, the complexity of such automatic methods limits the applicability to offline rendering. Here, machine learning methods could come into their own. Once trained, they are usually faster compared to approaches evaluating every parameter individually. This way, the performance of more complex perception models based on for example an eccentricity-dependent Contrast Sensitivity Function (CSF) evaluation or visual masking (limitations that bare tremendous potentials to reduce rendering costs) might be increased. In addition, although users might be able to perceive a visual difference compared to a fully rendered image, this is not always relevant to them solving a specific task or accomplishing a defined goal. Here, user studies may help determine minimum requirements for rendering algorithms. Also, the interplay of different senses is even more complex as well as mostly unexplored and models for multi-sensory perception hardly exist. However, VR and AR applications will increasingly provide multi-sensory experiences, including vision, audio, and haptics. Hence, we foresee further rendering optimization involving “tuning” methods to account for the non-visual senses emerging.

Perception-driven rendering has become a significant topic, and new ways to simulate and exploit human vision are sure to be discovered in the coming years. With affordable tracking devices, non-obtrusive ways to capture human attention and perception, novel VR devices available on a consumer level and ever-increasing displays technologies, the evolvement of vibrant, more immersive computer-generated realities is only a question of time.

Weier et al. [Wei+17]

APPENDIX

A.1 NO-REFERENCE METRIC FOR NOISE ESTIMATIONS

Measuring the noise without a reference is a challenge. It has to be decided whether the detected high-frequency components are noise or rather part of the actual signal. Likewise, the change of the signal over time must be considered to measure temporal stability. An algorithm to measure noise in images must therefore distinguish whether an image distortion and temporal flickering are artifacts, i.e. noise, or not. Vatolin et al. [Vat+11] introduce a set of approaches, available in the *MSU Noise Estimation Filter*. Unfortunately, few implementation details are provided. Hence, this description of the metrics refers to the website of the filters [Vat+11]. In total three different noise metrics are available.

MAD Performs a HAAR wavelet decomposition for each frame. The wavelet decomposition can be considered to relate to the multi-scale processing in the visual cortex. The decomposition yields four subbands of the image, the LL, LH, HL, and HH band. While the LL band contains a low-pass filtered version of the image, the other bands represent the high-frequency components. The LH component contains mostly vertical edges; the HL components emphasizes horizontal edges [Rit02, ch. 2]. Both are considered typical structures that can be found in natural images. However, the HH band can be interpreted to contain edges in diagonal direction [Rit02, ch. 2]. Such signals are considered to be noise. Hence, the medians of HH component's absolute values are computed by the metric. The final value of the metric is the normalized median of these values. (Section 2.2.4).

Block-Based First, the frames are tessellated into a number of 8×8 blocks. Then standard deviations of intensity are computed for all the blocks. This provides a value on the strength of the intensity change in each block – the smaller the standard deviation, the smoother the block. This intensity variation may be due to noise, in which the standard deviation of the block is close to that of Gaussian noise. Based on these values the blocks are sorted from low to high. The final value of the metric is the normalized mean of the standard deviations of 30% of the blocks with the lowest values.

Spatio-Temporal Gradients Performs a wavelet decomposition for each frame. In this way temporal and spatial histograms can be computed. The initial estimation of the noise level is determined by the value at which the time or space histogram reaches its maximum value. The decision of whether to use the spatial or temporal histogram is based on its deviation from the Rayleigh distribution. Later, this estimation is corrected, using a Kolmogorov-Smirnow test [All76]. The normalized corrected estimation is the final value of the metric.

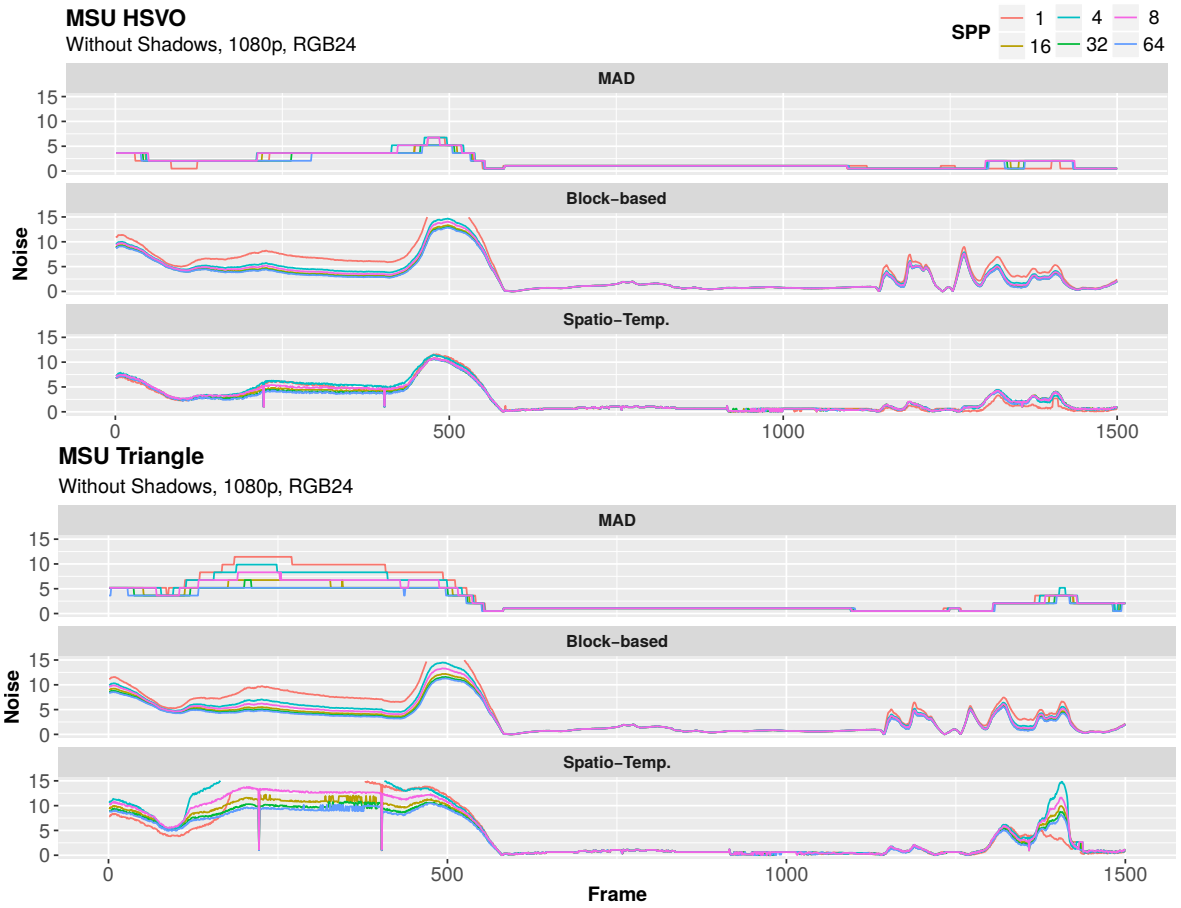


Figure 105: Results of the MSU noise estimation metrics (MAD, Block-based and Spatio-Temporal) on a high-frequency test scene of 1500 frames rendered at 1080p with different sample-per-pixel (spp). The voxel-based rendering method produces less noise compared to using triangles only if shadows are turned off.

The metrics are computed with the image sequence used to evaluate the Hybrid Sparse Voxel Octree (HSVO) (Figure 61), as presented in Section 5.3. Here, the scene rendered using the HSVO is compared to renderings produced with the full triangle data. Moreover, it has been tested with shadows turned both on and off. It becomes clear that using the HSVO significantly reduces noise if no shadows are computed (Figure 105), while this cannot be concluded when rendering the scene with shadows (Figure 106). This reduction of noise can also be confirmed by looking at the means of the noise levels as presented in Figure 107. While the mean noise level for renderings with HSVO in the case where no shadows are rendered are lower compared to using a triangle representation, the opposite is the case when rendering with shadows turned on. One possible explanation is that voxels usually provide a greater closed surface that is more likely to be directly hit by the light source. This can clearly be seen in scenes with tree and shrub vegetation. Trees in the distance appear to be brighter as the voxel levels do capture more direct light, while for the triangle representation more light is “lost” in the foliage. This increase in brightness can also be observed in Figure 62. These more intense illuminations might have influenced the intensity measurements performed by the metrics, and resulted in the slightly increased noise levels.

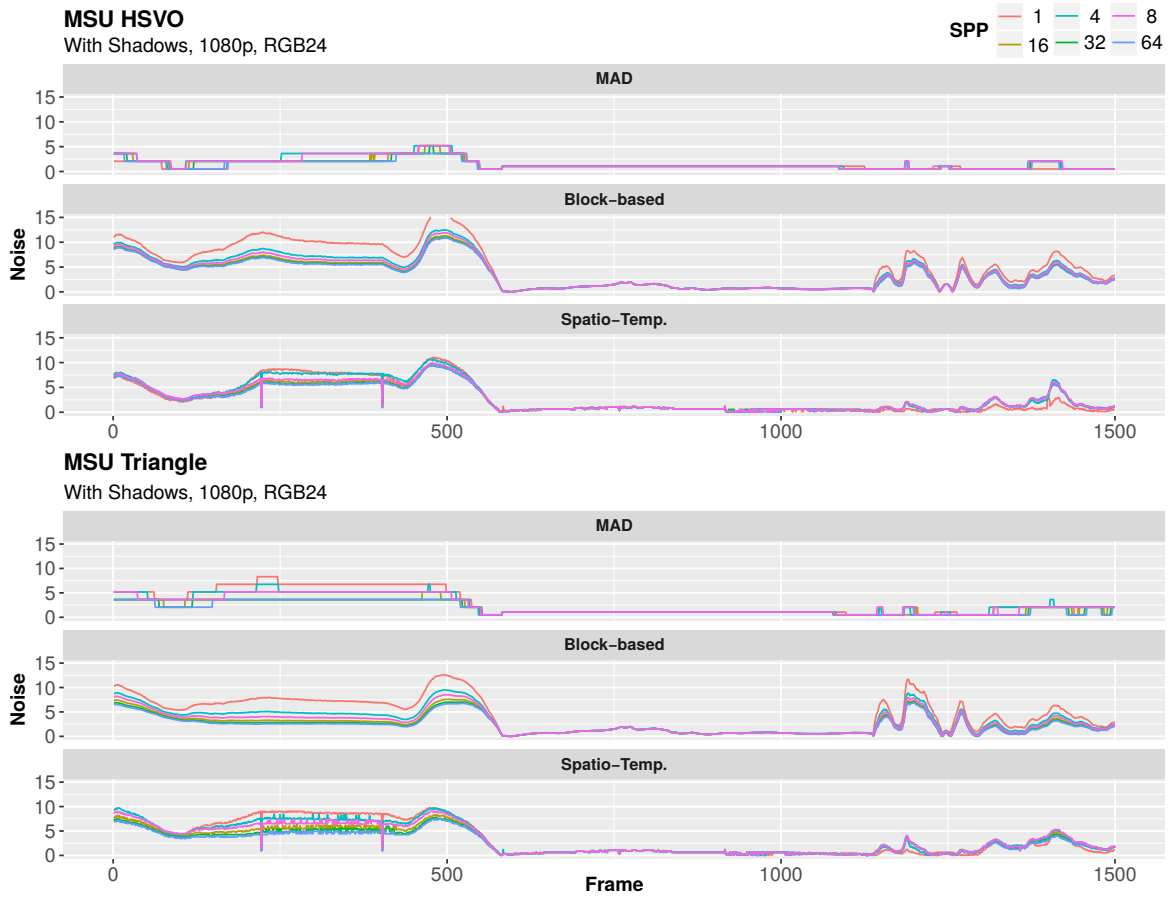
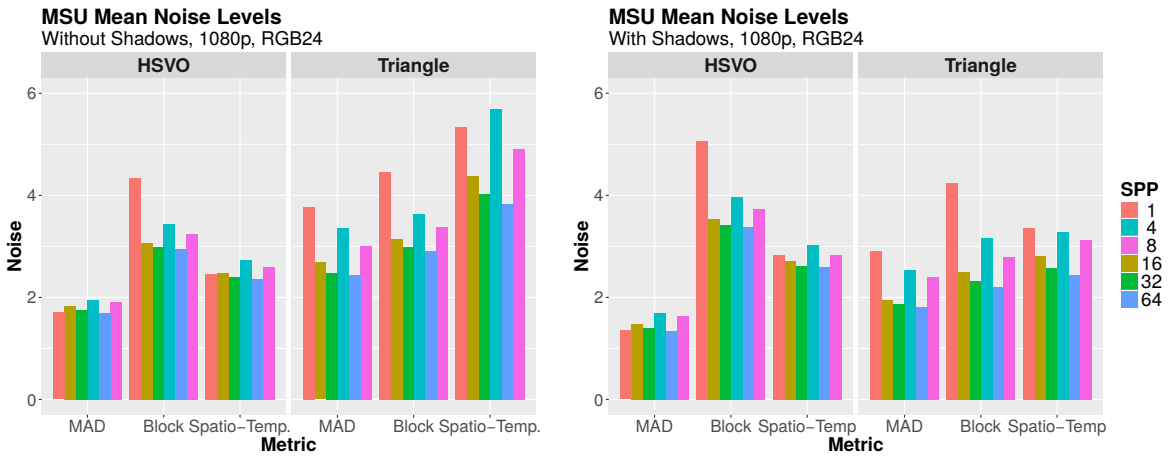


Figure 106: Results of the MSU noise estimation metrics (MAD, Block-based and Spatio-Temporal) on a high-frequent test scene of 1500 frames rendered at 1080p with different spp. Voxel-based rendering appears to produce either more or an equal amount of noise compared to using only triangles when rendering the scene with shadows.



(a) (b)

Figure 107: Mean MSU noise levels for the test scene renderer at 1080p with various spp. While the noise levels are lower when using the HSVO when shadows are turned off (a), turning shadows on provides less clear results compared to using only triangle data (b).

A.2 EYE TRACKING LATENCIES FOR GAZE-CONTINGENT RENDERING

Dedicated measurements of acceptable latencies for gaze-contingent displays have been conducted in several studies [LW07; San+07; SW14; Rin+14]. The measured *end-to-end latency* comprises the full gaze capture and rendering pipeline, starting with capturing data from the eye tracker and ending with the reception of display-emitted photons on the retina. The gaze-contingent display system presented by Santini et al. [San+07] renders at a frame rate of 200 Hz and achieves an end-to-end latency of only 10 ms with dedicated hardware. Loschky and Wolverton [LW07] tested for perceptually acceptable latencies with respect to peripheral image blur of different sizes and blurring filters. Images are blurred with the blur's strength dependent on the eccentricity. The Point-of-Regard (PoR) is determine with eye tracking hardware that supports updates at 1000 Hz. Their study reveals that image update delays as long as 60 ms did not significantly increase blur detection. Besides that, the acceptable delay for image updates depends on the task to be performed in the application and the stimulus size in the visual field. However, after a certain delay, the likelihood of the detection of slow updates increases quickly [LM00; LW07]. Nonetheless, for purely attentional processes such as detecting objects, great latencies are tolerable [Fei+07]. For gaze-contingent rendering, the work by Albert et al. [Alb+17] suggests that a total system latency of 50–70 ms could be tolerated. Shorter eye tracking latencies of 20–40 ms have absolutely no effect on the amount of "foveation". Hence, even eye trackers running at only 60 Hz to 90 Hz are generally sufficient for efficient gaze-contingent rendering.

A.3 CONSIDERATIONS ON RUNNING ESTIMATES

When exploiting temporal coherence, a common task is to compute a pixel color for a new frame by combining old and newly sampled color values. This combination is usually performed as a *running estimate*. Here, a new pixel p of a frame f at time t is computed using a weighted sum of the current sample and $s_t[p]$ and a (re-)projection π_{t-1} of p, fetching the color from the old frame f_{t-1}.

$$f_t[p] \leftarrow \alpha \cdot s_t[p] + (1-\alpha) \cdot f_{t-1}[\pi_{t-1}(p)] \tag{11}$$

The weight α is used to control the amount of variance reduction and the responsiveness of the cache to changes. A small α allows more samples to be integrated from the past but has a risk of over-blurring and a filter that is less responsive to signal changes. According to Scherzer et al. [SYM10, p. 7] variance reduction is given by

$$\lim_{t\to\infty} \frac{Var(f_t(p))}{Var(s_t(p))} = \frac{\alpha}{2-\alpha} \tag{12}$$

Choosing $\alpha = 2/5$, for example, reduces the variance to 1/4 of the original.

As choosing α provides a balance between smoothness and lag, finding an optimal value for α has been a target of much research. As the camera or objects move in the scene, the reprojected position usually samples the last frame at a different position. Moreover, potential magnifications and minifications of the reprojected pixel footprint of the last frame make

it necessary to filter the information from the last frame in order to obtain the reprojected pixel color. However, so as to be efficient and under the assumption that the change between subsequent frames is minimal, most systems use simple bilinear filtering to fetch samples from the cache. Unfortunately, this bilinear filtering influences the quality of the reprojected data as it tends to attenuate and misplace high frequencies. This influence is especially apparent when filtering takes place exactly between four neighboring pixels: As the camera or objects move, this can result in challenging pixel velocities that move sampling positions between neighboring pixels. The image gets blurred and at high velocities high-frequency components show ghosting artifacts. If this happens subsequently, these bilinear filtered results are accumulated over and over again, leading to blurring. This progressive low-pass nature can also be described using a probability mass function of a Bernoulli distribution. The mathematical derivation is presented in Scherzer et al. [SYM10, p. 9]. In order to overcome this over-blurring Yang et al. [Yan+09] carefully designed the running estimate. The weight α_t for each pixel p can be expressed by including the number of accumulated samples N_t per frame into the update rules.

$$\alpha_t[p] \leftarrow \frac{1}{N_{t-1}[p] + 1} \tag{13}$$

$$N_t[p] \leftarrow N_{t-1}[p] + 1 \tag{14}$$

Then, to limit the infinite accumulation of samples and thus the blur caused thereby, Yang et al. derived a threshold for α_t. This threshold is based on a precomputed table of pixel velocities including sampling momentums and positions [Yan+09, p. 6]. As this changes the rate of accumulation, the authors have also developed a new update rule for N_t.

$$N_t[p] \leftarrow \left(\alpha_t[p]^2 + \frac{(1 - \alpha_t[p])^2}{N_{t-1}[\pi_{t-1}(p)]} \right)^{-1} \tag{15}$$

Besides this careful design of the running estimate, Yang et al. work with higher resolution buffers of 2×2 subpixel precision, updated in an irregular fashion to limit ghosting artifacts and blurring. Often temporal methods are used to improve perception and create more temporally stable images as presented in Chapter 4. In the foveated rendering pipeline, presented in Section 6.1.4 on Page 137, the decision has been made to increase temporal stability by sampling the reprojected pixel footprints multiple times in order to get more accurate samples from the previous frames. In contrast to the approaches presented, not every pixel is updated in each frame. Hence, for this work a limit for α, based on the sample's age, is proposed.

A.4 RESOLUTION ESTIMATES OF AN OPTIMAL HMD

According to a keynote talk by Warren Hunt [Hun15] from Oculus Research/Facebook Reality Labs that was presented at *High-Performance Graphics 2015*, achieving retinal resolution in commodity Head-Mounted Displays (HMDs) would require approximately $16K \times 16K = 256$ Megapixels. Here, Hunt assumed a Field of View (FoV) to be 100° which was the state-of-the-art for consumer level HMDs at that time. For a full FoV of an eye that can move, Hunt assumed 200° horizontally by 150° vertically, which resulted, according to his computations, in a necessary resolution of $32K \times 24K = 768$ Megapixels per eye. The cone spacing and optical filtering limits the Minimum Angle of Resolution (MAR), yielding a cut-off frequency

of about 60 Cycle per Degree (cpd) (Section 2.2.1). Although Hunt states that these values have been computed using the MAR of the eye of 1 arcmin (Section 2.2.2), different resolution limits were obtained in this research. In order to accurately represent such a grating pattern, a pixel should span $0.5'$ arcmin (Figure 17, Page 25). Given $0.5'$ arcmin ($= 0.008\overline{3}^\circ$) and a 100° FoV

$$100^\circ/(0.008\overline{3}^\circ) = 12000.0$$

a resolution of $12k \times 12k = 144$ Megapixels per eye is sufficient. Considering a fully dynamic FoV of 290° × 150°

$$290^\circ/(0.008\overline{3}^\circ) = 34800.0 \qquad 150^\circ/(0.008\overline{3}^\circ) = 18000.0$$

an HMD ideally needs a maximal resolution of $35k \times 18k = 630$ Megapixels. What is missing here however, is the discussion of hyperacuity. Considering the minimum discriminable acuity of 0.00024° (Table 1, Page 25) the resolution needs to be as high as approx. $1208k \times 625k = 755$ Gigapixel per eye. With a minimum visible acuity of 0.00014° (Table 1, Page 25), resolution requirements rise to approx. $1429k \times 1071k = 1530.459$ Gigapixel for a single eye. These resolutions are necessary to achieve or exceed Vernier acuity (Section 2.2.2). There is still a long way to achieving this. An overview on the development of pixel densities and refresh rates of HMD systems over the last decades is provided in Table 12

Device Name	**Year**	**Resolution**	**Refresh Rate**
Forte VFX-1	1994	263x230x2	60 Hz
Sony Glasstron PLM-S700E	1997	832x642	85 Hz
Forte VFX 3D	1998	263x480x2	60 Hz
eMagin Z800	2005	800x600x2	60 Hz
Headplay Visor	2007	800x600	120 Hz
Zeiss cinemizer plus	2009	640x480	60 Hz
Sony HMZ-T1	2011	1280x720x2	60 Hz
Oculus DK1	2013	640x800x2	60 Hz
Oculus DK2	2014	960x1080x2	75 HZ
Razer OSVR	2015	960x1080x2	120 Hz
Samsung Gear VR	2015	1280x1440x2	60 Hz
Sony Playstation VR	2016	960x1080x2	120 Hz
Oculus VR	2016	1080x1200x2	90 Hz
HTC Vive	2016	1080x1200x2	90 Hz
StarVR	2016	2560x1440x2	90 Hz
FoveVR	2017	1280x1440x2	75 Hz
HTC Vive Pro	2017	1440x1660x2	90 Hz
Pimax 8k	2018	3840×2160x2	80 Hz

Table 12: The development of pixel densities and refresh rates of HMDs over the last two decades.

A.5 GAZE-DEPTH CALIBRATION CAMERA POSITONS

In Figure 108 - Figure 110 the camera positions that were selected to evaluate the gaze-depth estimator (Section 7.3) are presented.

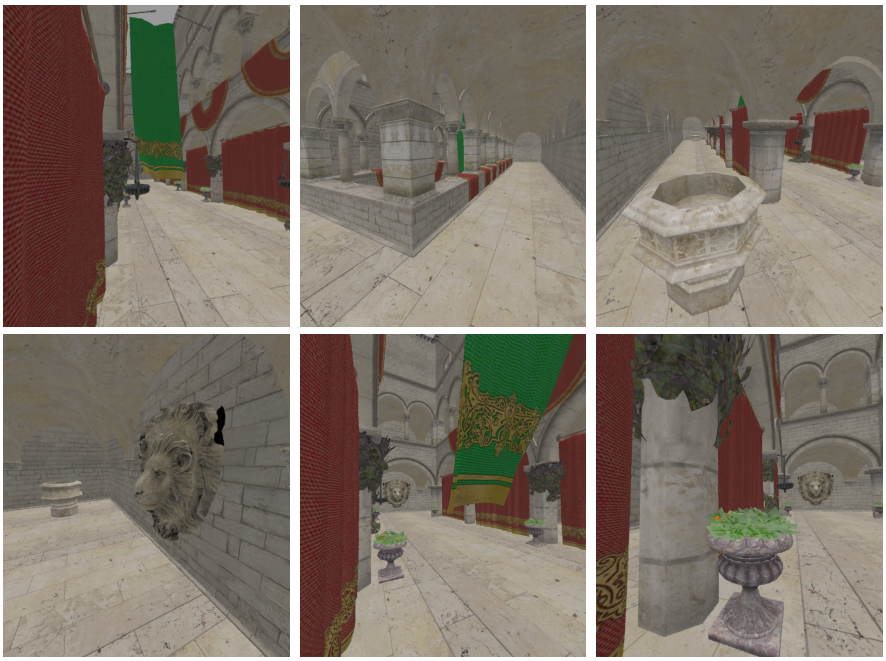


Figure 108: Camera positions for the scene Sponza used in the experimental evaluation of the gaze-depth estimation.

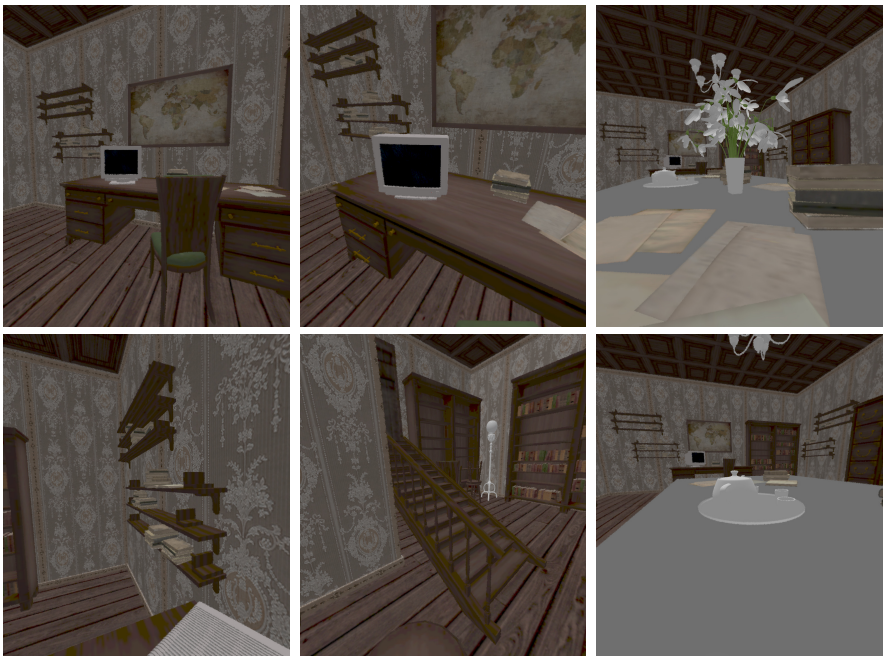


Figure 109: Camera positions for the scene StudyRoom used in the experimental evaluation of the gaze-depth estimation.

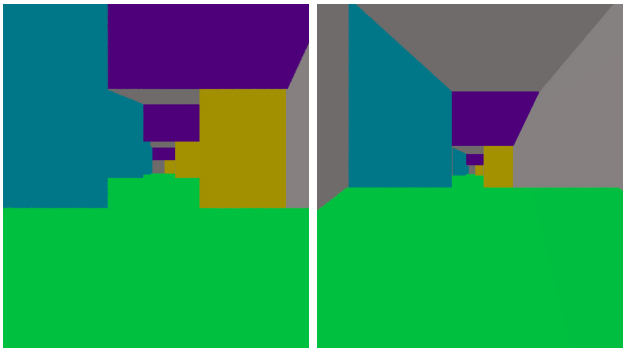


Figure 110: Camera positions for the artificial scenes TestNear (left) and TestFar (right) used in the experimental evaluation of the gaze-depth estimation.

BIBLIOGRAPHY

[Ade82] E. H. Adelson. *"Saturation and adaptation in the rod system."* eng. In: *Vision Research* 22.10 (1982), pp. 1299–1312. ISSN: 0042-6989.

[AH93] Stephen J. Adelson and Larry F. Hodges. *"Stereoscopic ray-tracing."* In: *The Visual Computer* 10.3 (1993), pp. 127–144.

[AH95] Stephen J. Adelson and Larry F. Hodges. *"Generating Exact Ray-Traced Animation Frames by Reprojection."* In: *IEEE Computer Graphics and Applications* 15.3 (1995), pp. 43–52.

[Adl+11] Francis Heed Adler, Paul L. Kaufman, Leonard A. Levin, and Albert Alm. *Adler's Physiology of the Eye*. Elsevier Health Sciences, 2011. ISBN: 978-0-323-05714-1.

[AD11] Venceslas Biri Adrien Herubel and Stephane Deverly. *"Morpological antialiasing and topological reconstruction."* In: *GRAPP*. 2011.

[AKF13] Velibor Adzic, Hari Kalva, and Borko Furht. *"Exploring Visual Temporal Masking for Video Compression."* In: *2013 IEEE International Conference on Consumer Electronics*. ICCE. 2013, pp. 590–591.

[AGL89] Mark Agate, Richard L. Grimsdale, and Paul F. Lister. *"The HERO Algorithm for Ray-Tracing Octrees."* In: *Advances in Computer Graphics Hardware*. Ed. by Richard L. Grimsdale and Wolfgang Straßer. Springer, 1989, pp. 61–73. ISBN: 3-540-53473-3.

[Ahm+16] Abdalla G. M. Ahmed, Hélène Perrier, David Coeurjolly, Victor Ostromoukhov, Jianwei Guo, Dong-Ming Yan, Hui Huang, and Oliver Deussen. *"Low-discrepancy Blue Noise Sampling."* In: *ACM Trans. Graph.* 35.6 (Nov. 2016), 247:1–247:13. ISSN: 0730-0301.

[ALK12] Timo Aila, Samuli Laine, and Tero Karras. *Understanding the Efficiency of Ray Traversal on GPUs - Kepler and Fermi Addendum*. Technical Report NVR-2012-02. NVIDIA, 2012. HPG2012 poster.

[Ake93] Kurt Akeley. *"Reality Engine Graphics."* In: *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '93. ACM, 1993, pp. 109–116. ISBN: 978-0-89791-601-1.

[AHH08] Tomas Akenine-Möller, Eric Haines, and Naty Hoffman. *Real-Time Rendering*. Third Edition. Wellesley, Massachusetts: A K Peters Ltd., 2008. ISBN: 9781568814247.

[Alb+17] Rachel Albert, Anjul Patney, David Luebke, and Joohwan Kim. *"Latency Requirements for Foveated Rendering in Virtual Reality."* In: *ACM Trans. Appl. Percept.* 14.4 (Sept. 2017), 25:1–25:13.

[All76] M.E. Allen. *Kolmogorov-Smirnov Test for Discrete Distributions*. Defense Technical Information Center, 1976. Naval Postgraduate School Monterey, CA.

[AW87] John Amanatides and Andrew Woo. *"A Fast Voxel Traversal Algorithm for Ray Tracing."* In: *EG 1987-Technical Papers*. Eurographics Association, 1987.

[APY16] Seyed Ali Amirshahi, Marius Pedersen, and Stella X. Yu. *"Image Quality Assessment by Comparing CNN Features between Images."* In: *Journal of Imaging Science and Technology* 60.6 (2016), p. 60410.

[AET96] Roger S. Anderson, David W. Evans, and Larry N. Thibos. *"Effect of window size on detection acuity and resolution acuity for sinusoidal gratings in central and peripheral vision."* In: *Journal of the Optical Society of America. A, Optics, Image Science, and Vision* 13.4 (1996), pp. 697–706.

[And10] Dmitry Andreev. *"Real-time Frame Rate Up-conversion for Video Games: Or How to Get from 30 to 60 Fps for "Free"."* In: *ACM SIGGRAPH 2010 Talks*. SIGGRAPH '10. Los Angeles, California: ACM, 2010, 16:1–16:1. ISBN: 978-1-4503-0394-1.

[ARH78] Stuart Anstis, Brian Rogers, and Jean Henry. *"Interactions between simultaneous contrast and coloured afterimages."* In: *Vision Research* 18.8 (1978), pp. 899–911.

[Ara+17] Elena Arabadzhiyska, Okan Tarhan Tursun, Karol Myszkowski, Hans-Peter Seidel, and Piotr Didyk. *"Saccade Landing Position Prediction for Gaze-Contingent Rendering."* In: *SIGGRAPH '17, ACM Transactions on Graphics (TOC)* 36.4 (2017).

[AL73] G. B. Arden and D. R. Lewis. *"The pattern visual evoked response in the assessment of visual acuity."* In: *Transactions of the Ophthalmological Societies of the United Kingdom* 93.0 (1973), pp. 39–48.

[AH05] Arul Asirvatham and Hugues Hoppe. *"GPU Gems 2: Programming Techniques for High-Performance Graphics and General-Purpose Computation (Gpu Gems)."* In: ed. by Matt Pharr and Randima Fernando. 2. Addison-Wesley Professional, 2005. Chap. Geometry Clipmaps. ISBN: 0321335597.

[ASG15] Tunç Ozan Aydın, Aljoscha Smolic, and Markus Gross. *"Automated Aesthetic Analysis of Photographic Images."* In: *IEEE Transactions on Visualization and Computer Graphics* 21.1 (2015), pp. 31–42.

[Bad88] Sig Badt Jr. *"Two Algorithms for Taking Advantage of Temporal Coherence in Ray Tracing."* In: *The Visual Computer* 4.3 (1988), pp. 123–132.

[BAS75] A. T. Bahill, D. Adler, and L. Stark. *"Most naturally occurring human saccades have magnitudes of 15 degrees or less."* In: *Invest Ophthalmol* 14.6 (June 1975), pp. 468–469.

[BL13] I. L. Bailey and J. E. Lovie-Kitchin. *"Visual acuity testing. From the laboratory to the clinic."* In: *Vision Res.* 90 (Sept. 2013), pp. 2–9.

[Bak49] H. D. Baker. *"The course of foveal light adaptation measured by the threshold intensity increment."* eng. In: *Journal of the Optical Society of America* 39.2 (1949), pp. 172–179. ISSN: 0030-3941.

[BNR09] B. Balas, L. Nakano, and R. Rosenholtz. *"A summary-statistic representation in peripheral vision explains visual crowding."* In: *Journal of Vision* 9.12 (2009), pp. 11–18.

[Bal+14] M. Balsa Rodríguez, E. Gobbetti, J. A. Iglesias Guitián, M. Makhinya, F. Marton, R. Pajarola, and S.K. Suter. *"State-of-the-Art in Compressed GPU-Based Direct Volume Rendering."* In: *Comput. Graph. Forum* 33.6 (Sept. 2014), pp. 77–100. ISSN: 0167-7055.

[BSA91] Martin S. Banks, Allison B. Sekuler, and Stephen J. Anderson. *"Peripheral spatial vision: Limits imposed by optics, photoreceptors, and receptor pooling."* In: *Journal of the Optical Society of America* 8.11 (1991), pp. 1775–1787.

[Bar79] G. R. Barnes. *"Vestibulo-ocular function during co-ordinated head and eye movements to acquire visual targets."* In: *J. Physiol. (Lond.)* 287 (Feb. 1979), pp. 127–147.

[Bar19] Colin Barré-Brisebois. *Are we done with ray tracing? State-of-the-art and Challenges in Game Ray Tracing*. Electronic. Aug. 2019. Siggraph Course Notes, https://drive.google.com/file/d/1kNzbQ7oglYQGrZz32355wU-KoizUwhqb/view, last visited 28. Oct. 2019.

[BK08] Brian A. Barsky and Todd J. Kosloff. *"Algorithms for Rendering Depth of Field Effects in Computer Graphics."* In: *Proceedings of the 12th WSEAS International Conference on Computers*. ICCOMP'08. Heraklion, Greece: World Scientific, Engineering Academy, and Society (WSEAS), 2008, pp. 999–1010. ISBN: 978-960-6766-85-5.

[Bar99] Peter Barten. *Contrast sensitivity of the human eye and its effects on image quality*. Bellingham, WA: SPIE Optical Engineering Press, 1999. ISBN: 9780819434968.

[BKM05] Josephine Battista, Michael Kalloniatis, and Andrew Metha. *"Visual function: the problem with eccentricity."* eng. In: *Clinical & Experimental Optometry* 88.5 (2005), pp. 313–321. ISSN: 0816-4622.

[Bau+15] Pablo Bauszat, Martin Eisemann, Elmar Eisemann, and Marcus Magnor. *"General and Robust Error Estimation and Reconstruction for Monte Carlo Rendering."* In: *Computer Graphics Forum* 34.2 (May 2015), pp. 597–608. ISSN: 0167-7055.

[BEM15] Pablo Bauszat, Martin Eisemann, and Marcus Magnor. *"Sample-Based Manifold Filtering for Interactive Global Illumination and Depth of Field."* In: *Computer Graphics Forum* 34.1 (Feb. 2015), pp. 265–276.

[Bau+11] Pablo Bauszat, Martin Eisemann, Marcus Magnor, and Naveed Ahmed. *"Guided Image Filtering for Interactive High-quality Global Illumination."* In: *Computer Graphics Forum (Proc. of Eurographics Symposium on Rendering EGSR)* 30.4 (June 2011), pp. 1361–1368.

[BG16] Dean Beeler and Anuj Gosalia. *Asynchronous Timewarp on Oculus Rift*. Ed. by Facebook Reality Labs Oculus Research. 2016. URL: https://developer.oculus.com/blog/asynchronous-timewarp-on-oculus-rift/. last visited 11. Nov. 2019.

[BHP16] Dean Beeler, Ed Hutchins, and Paul Pedriana. *Asynchronous Spacewarp*. Ed. by Facebook Reality Labs Oculus Research. 2016. URL: https://developer.oculus.com/blog/asynchronous-spacewarp/. last visited 11. Nov. 2019.

[Beg+13] A. Beghdadi, M.-C. Larabi, A. Bouzerdoum, and K.M. Iftekharuddin. *"A survey of perceptual image processing methods."* en. In: *Signal Processing: Image Communication* 28.8 (Sept. 2013), pp. 811–831.

[Beh+05] S. Behrendt, C. Colditz, O. Franzke, J. Kopf, and O. Deussen. *"Realistic real-time rendering of landscapes using billboard clouds."* In: *Computer Graphics Forum* (2005).

[Bem+19] Mojtaba Bemana, Joachim Keinert, Karol Myszkowski, Michel Bätz, Matthias Ziegler, Hans-Peter Seidel, and Tobias Ritschel. *"Learning to Predict Image-based Rendering Artifacts with Respect to a Hidden Reference Image."* eng. In: *Computer Graphics Forum (Proc. Pacific Graphics)* 38.7 (2019). ISSN: 1467-8659.

[Ben19] Anis Benyoub. *"Leveraging Real-time Ray Tracing To Build A Hybrid Game Engine."* In: *Advances in Real-Time Rendering in Games*. ACM Siggraph. 2019. Course, Unity Technologies.

[Ber+10] Kai Berger, Christian Lipski, Christian Linz, Anita Sellent, and Marcus Magnor. *"A ghosting artifact detector for interpolated image quality assessment."* In: *Proc. IEEE International Symposium on Consumer Electronics (ISCE)*. 2010.

[BF12] Floraine Berthouzoz and Raanan Fattal. *"Resolution Enhancement by Vibrating Displays."* In: *ACM Trans. Graph.* 31.2 (Apr. 2012), 15:1–15:14. ISSN: 0730-0301.

[Bia+16] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. *"On the use of deep learning for blind image quality assessment."* In: *CoRR, preprint* arXiv:1602.05531 (2016).

[Bil17] Filip Biljecki. *"Level of detail in 3D City Models."* PhD Thesis. TU Delft, 2017.

[BOA13] Markus Billeter, Ola Olsson, and Ulf Assarsson. *"Tiled Forward Shading."* In: *GPU Pro 4: Advanced Rendering Techniques*. Ed. by Wolfgang Engel. Boca Raton, FL, USA: A K Peters/CRC Press, Jan. 1, 2013, pp. 99–114.

[BHD10] Venceslas Biri, Adrien Herubel, and Stephane Deverly. *"Practical Morphological Antialiasing on the GPU."* In: *ACM SIGGRAPH 2010 Talks*. SIGGRAPH '10. Los Angeles, California: ACM, 2010, 45:1–45:1. ISBN: 978-1-4503-0394-1.

[Bis+94] Gary Bishop, Henry Fuchs, Leonard McMillan, and Ellen J. Scher Zagier. *"Frameless Rendering: Double Buffering Considered Harmful."* In: *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. SIGGRAPH '94. ACM, 1994, pp. 175–176.

[Bit+16] Benedikt Bitterli, Fabrice Rousselle, Bochang Moon, José A. Iglesias-Guitián, David Adler, Kenny Mitchell, Wojciech Jarosz, and Jan Novák. *"Nonlinearly Weighted First-order Regression for Denoising Monte Carlo Renderings."* In: *Computer Graphics Forum (Proceedings of EGSR)* 35.4 (June 2016), pp. 107–117.

[BSJ04] Randolph Blake, Kenith V. Sobel, and Thomas W. James. *"Neural synergy between kinetic vision and touch."* eng. In: *Psychological Science* 15.6 (2004), pp. 397–402. ISSN: 0956-7976.

[BC69] C. Blakemore and F. W. Campbell. *"On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images."* In: *J. Physiol. (Lond.)* 203.1 (July 1969), pp. 237–260.

[Bli02] Jim Blinn. *Jim Blinn's Corner: Notation, Notation, Notation (Jim Blinn's Corner)*. 1st. The Morgan Kaufmann Series in Computer Graphics. Morgan Kaufmann, 2002. ISBN: 1558608605.

[Blo83] Jules Bloomenthal. *"Edge Inference with Applications to Antialiasing."* In: *Proceedings of the 10th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '83. Detroit, Michigan, USA: ACM, 1983, pp. 157–162. ISBN: 0-89791-109-1.

[Bod+07] P. Bodrogi, W. D. Wright, J. Schanda, K. Witt, and et al. *Colorimetry - Understanding the CIE System*. Ed. by Janos Schanda. 1st ed. New York: John Wiley & Sons, 2007.

[Boe+06] Martin Boehme, Michael Dorr, Christopher Krause, Thomas Martinetz, and Erhardt Barth. *"Eye movement predictions on natural videos."* In: *Neurocomputing* 69 (16-18 Oct. 2006), pp. 1996–2004.

[BM95] Mark R. Bolin and Gary W. Meyer. *"A Frequency Based Ray Tracer."* In: *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '95. ACM, 1995, pp. 409–418. ISBN: 0-89791-701-4.

[BM98] Mark R. Bolin and Gary W. Meyer. *"A Perceptually Based Adaptive Sampling Algorithm."* In: *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '98. ACM, 1998, pp. 299–309. ISBN: 0-89791-999-8.

[Bon+10] Nicolas Bonneel, Clara Suied, Isabelle Viaud-Delmon, and George Drettakis. *"Bimodal Perception of Audio-visual Material Properties for Virtual Environments."* In: *ACM Transactions on Applied Perception (TAP)* 7.1 (2010), pp. 1–16. ISSN: 1544-3558.

[BI13] Ali Borji and Laurent Itti. *"State-of-the-Art in Visual Attention Modeling."* In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 35.1 (2013), pp. 185–207.

[Bos+18] S. Bosse, D. Maniry, K. Müller, T. Wiegand, and W. Samek. *"Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment."* In: *IEEE Transactions on Image Processing* 27.1 (Jan. 2018), pp. 206–219.

[Bos+16] Sebastian Bosse, Dominique Maniry, Thomas Wiegand, and Wojciech Samek. *"A deep neural network for image quality assessment."* In: *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE. 2016, pp. 3773–3777.

[Bot+05] Mario Botsch, Alexander Hornung, Matthias Zwicker, and Leif Kobbelt. *"High-quality Surface Splatting on Today's GPUs."* In: *Proceedings of the Second Eurographics / IEEE VGTC Conference on Point-Based Graphics*. SPBG'05. New York, USA: Eurographics Association, 2005, pp. 17–24. ISBN: 3-905673-20-7.

[Bou70] H. Bouma. *"Interaction Effects in Parafoveal Letter Recognition."* In: *Nature* 226.5241 (1970), pp. 177–178.

[Bow10] Huw Bowles. *Efficient Real-Time Stereoscopic 3D Rendering*. 2010. Master Thesis ETH Zürich.

[BD07] G.E.P. Box and N.R. Draper. *Response Surfaces, Mixtures, and Ridge Analyses*. Wiley Series in Probability and Statistics. Wiley, 2007. ISBN: 9780470072752.

[BBS09] Margarita Bratkova, Solomon Boulos, and Peter Shirley. *"oRGB: A Practical Opponent Color Space for Computer Graphics."* In: *IEEE Computer Graphics and Applications* 29 (2009), pp. 42–55.

[BO00] Bruno G. Breitmeyer and Haluk Ogmen. *"Recent models and findings in visual backward masking: A comparison, review, and update."* en. In: *Perception & Psychophysics* 62.8 (2000), pp. 1572–1595. ISSN: 0031-5117, 1532-5962.

[Bre+05] Jean-Pierre Bresciani, Marc O. Ernst, Knut Drewing, Guillaume Bouyer, Vincent Maury, and Abderrahmane Kheddar. *"Feeling what you hear: auditory signals can modulate tactile tap perception."* eng. In: *Experimental Brain Research* 162.2 (2005), pp. 172–180. ISSN: 0014-4819.

[Buk+13] Mike Bukowski, Padraic Hennessy, Brian Osman, and Morgan McGuire. *"The Skylanders SWAP Force Depth-of-Field Shader."* In: *Published in GPU Pro 4: Advanced Rendering Techniques*. Apr. 2013, pp. 175–184. URL: http://casual-effects.com/research/Bukowski2013DepthOfField/index.html. last visited 13. Nov. 2019.

[Bur18] Andrew Burne. *NVIDIA NGX Technology - AI for Visual Applications*. 2018. URL: https://www.nvidia.com/en-us/geforce/news/new-gfe-features-latest-games-get-ansel-and-highlights/. last visited 26. Oct. 2018.

[Byl+15] Zoya Bylinskii, Tilke Judd, Frédo Durand, Aude Oliva, and Antonio Torralba. *MIT Saliency Benchmark*. 2015. URL: http://saliency.mit.edu/results_mit300.html. last visited 21. Nov. 2018.

[CHF92] David C. Van Essen, Charles H. Anderson, and Daniel Felleman. *"Information processing in the primate visual system: An integrated systems perspective."* In: 255 (Feb. 1992), pp. 419–23.

[Čad04] Martin Čadík. *Human Perception and Computer Graphics*. Czech Technical University, Prague, 2004. Postgraduate Study Report DC-PSR-2004-06, Czech Technical University.

[Čad+12] Martin Čadík, Robert Herzog, Rafał Mantiuk, Karol Myszkowski, and Hans-Peter Seidel. *"New Measurements Reveal Weaknesses of Image Quality Metrics in Evaluating Graphics Artifacts."* In: *ACM SIGGRAPH Asia 2012, Transactions on Graphics (TOG)* 31.6 (2012), pp. 147–157.

[CM85] T. Caelli and G. Moraglia. *"On the detection of Gabor signals and discrimination of Gabor textures."* In: *Vision Research* 25.5 (1985), pp. 671–684.

[CG65] F. W. Campbell and D. G. Green. *"Optical and retinal factors affecting visual resolution."* In: *J. Physiol. (Lond.)* 181.3 (Dec. 1965), pp. 576–593.

[CR68] F. W. Campbell and J. G. Robson. *"Application of Fourier analysis to the visibility of gratings."* In: *J. Physiol. (Lond.)* 197.3 (Aug. 1968), pp. 551–566.

[CCW03] Kirsten Cater, Alan Chalmers, and Greg Ward. *"Detail to Attention: Exploiting Visual Tasks for Selective Rendering."* In: *In Proceedings of the 14th ACM Eurographics Symposium on Rendering*. Vol. 44. EGSR '03. 2003, pp. 270–280.

[Cer+08] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. *"Predicting human gaze using low-level saliency combined with face detection."* In: *Advances in Neural Information Processing Systems* 20 (2008), pp 1-7.

[Cha+17] Chakravarty R. Alla Chaitanya, Anton S. Kaplanyan, Christoph Schied, Marco Salvi, Aaron Lefohn, Derek Nowrouzezahrai, and Timo Aila. *"Interactive Reconstruction of Monte Carlo Image Sequences Using a Recurrent Denoising Autoencoder."* In: *ACM Trans. Graph.* 36.4 (July 2017), 98:1–98:12. ISSN: 0730-0301.

[CML11] Matthäus G. Chajdas, Morgan McGuire, and David Luebke. *"Subpixel Reconstruction Antialiasing for Deferred Shading."* In: *Symposium on Interactive 3D Graphics and Games*. I3D '11. San Francisco, California: ACM, 2011, pp. 15–22. ISBN: 978-1-4503-0565-5.

[CDS06] Alan Chalmers, Kurt Debattista, and Luis Paulo dos Santos. *"Selective Rendering: Computing Only What You See."* In: *Proceedings of the 4th International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia*. GRAPHITE '06. 2006, pp. 9–18.

[CH07] Damon M. Chandler and Sheila S. Hemami. *"VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images."* In: *IEEE Transactions on Image Processing (TIP)* 16.9 (2007), pp. 2284–2298.

[CL11] Chih-Chung Chang and Chih-Jen Lin. *"LIBSVM: A Library for Support Vector Machines."* In: *ACM Trans. Intell. Syst. Technol.* 2.3 (May 2011), 27:1–27:27. ISSN: 2157-6904.

[Che+05] H. Chen, S. Kim, S. Lee, O. Kwon, and J. Sung. *"Nonlinearity compensated smooth frame insertion for motion-blur reduction in LCD."* In: *2005 IEEE 7th Workshop on Multimedia Signal Processing*. Oct. 2005.

[CBN05] Hannah F. Chua, Julie E. Boland, and Richard E. Nisbett. *"Cultural variation in eye movements during scene perception."* In: *Proceedings of the National Academy of Sciences of the United States of America* 102.35 (2005), pp. 12629–12633.

[Cla+14] Petrik Clarberg, Robert Toth, Jon Hasselgren, Jim Nilsson, and Tomas Akenine-Möller. *"AMFS: Adaptive Multi-Frequency Shading for Future Graphics Processors."* In: *SIGGRAPH '14, Transactions on Graphics (TOG)* 33.4 (2014), pp. 141–152.

[Cla76] James H. Clark. *"Hierarchical Geometric Models for Visible Surface Algorithms."* In: *Commun. ACM* 19.10 (Oct. 1976), pp. 547–554.

[COM98] Jonathan Cohen, Marc Olano, and Dinesh Manocha. *"Appearance-preserving Simplification."* In: *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '98. New York, NY, USA: ACM, 1998, pp. 115–122. ISBN: 0-89791-999-8.

[Col01] August Colenbrander. *"Vision and Vision Loss."* In: Williams and Wilkins. *Duane's Clinical Opthalmology*. Vol. 5. 2001, p. 51.

[Cor+07] Massimiliano Corsini, Elisa Drelie Gelasca, Touradj Ebrahimi, and Mauro Barni. *"Watermarked 3D Mesh Quality Assessment."* In: *IEEE Transactions on Multimedia (TOMM)* 9.2 (2007), pp. 247–256.

[Cor+13] Massimiliano Corsini, Mohamed-Chaker Larabi, Guillaume Lavoué, Oldrich Petřík, Libor Váša, and Kai Wang. *"Perceptual metrics for static and dynamic triangle meshes."* In: *ACM Eurographics '12 - STAR, Computer Graphics Forum* 32.1 (2013), pp. 101–125.

[Cor19] Cory Corvus. *VR Eye Tracking & Foveated Rendering with VRS*. Electronic. 2019. https://developer.download.nvidia.com/video/gputechconf/gtc/2019/presentation/s91047-vive-pro-eye-tracking-and-foveated-rendering-with-vrs-presented-by-htc-vive.pdf, Presentation at GTC 2019, last visited 28. Oct. 2019.

[CR74] A. Cowey and E. T. Rolls. *"Human Cortical Magnification Factor and its Relation to Visual Acuity."* en. In: *Experimental Brain Research* 21.5 (1974), pp. 447–454. ISSN: 0014-4819, 1432-1106.

[CG12] Cyril Crassin and Simon Green. *Octree-Based Sparse Voxelization Using The GPU Hardware Rasterizer*. Ed. by Christophe Riccio Patrick Cozzi. OpenGL Insights. NVIDIA Research, July 2012, pp. 303–319. ISBN: 9781439893760.

[Cra+15] Cyril Crassin, Morgan McGuire, Kayvon Fatahalian, and Aaron Lefohn. *"Aggregate G-buffer Anti-aliasing."* In: *Proceedings of the 19th Symposium on Interactive 3D Graphics and Games*. i3D '15. San Francisco, California: ACM, 2015, pp. 109–119. ISBN: 978-1-4503-3392-4.

[Cra+09] Cyril Crassin, Fabrice Neyret, Sylvain Lefebvre, and Elmar Eisemann. *"GigaVoxels: Ray-Guided Streaming for Efficient and Detailed Voxel Rendering."* In: *ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (I3D)*. ACM. Boston, MA, Etats-Unis: ACM Press, Feb. 2009.

[Cra+11] Cyril Crassin, Fabrice Neyret, Miguel Sainz, Simon Green, and Elmar Eisemann. *"Interactive Indirect Illumination Using Voxel Cone Tracing: A Preview."* In: *Symposium on Interactive 3D Graphics and Games*. I3D '11. San Francisco, California: ACM, 2011, p. 207. ISBN: 978-1-4503-0565-5.

[Cur+90] Christine A. Curcio, Kenneth R. Sloan, Robert E. Kalina, and Anita E. Hendrickson. *"Human photoreceptor topography."* In: *The Journal of Comparative Neurology* 292.4 (Feb. 1990), pp. 497–523.

[CV95] J. E. Cutting and P. M. Vishton. *"Perceiving layout and knowing distances: the integration, relative potency and contextual use of different information about depth."* In: *Handbook of Perception and Cognition*. Ed. by W. Epstein and S. Rogers. Vol. 5: Perception of Space and Motion. 1995, pp. 69–117.

[Dac11] Carsten Dachsbacher. *"Analyzing Visibility Configurations."* In: *IEEE Transactions on Visualization and Computer Graphics, TVCG '17* 17.4 (Apr. 2011), pp. 475–486. ISSN: 1077-2626.

[Dad+16] Bas Dado, Timothy R. Kol, Pablo Bauszat, Jean-Marc Thiery, and Elmar Eisemann. *"Geometry and Attribute Compression for Voxel Scenes."* In: *Computer Graphics Forum* 35.2 (2016), pp. 397–407.

[Dal93] Scott Daly. *"Digital Images and Human Vision."* In: ed. by Andrew B. Watson. Cambridge, MA, USA: MIT Press, 1993. Chap. The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity, pp. 179–206. ISBN: 0-262-23171-9.

[Dal98] Scott J. Daly. *"Engineering observations from spatiovelocity and spatiotemporal visual models."* In: *In Proceedings of SPIE - The International Society for Optical Engineering* 3299 (1998), pp. 180–191.

[Dam+10] Holger Dammertz, Daniel Sewtz, Johannes Hanika, and Hendrik Lensch. *"Edge–Avoiding A-Trous Wavelet Transform for fast Global Illumination Filtering."* In: *Proc. High Performance Graphics 2010*. 2010, pp. 67–75.

[DW61] P. M. Daniel and D. Whitteridge. *"The representation of the visual field on the cerebral cortex in monkeys."* In: *The Journal of Physiology* 159.2 (1961), pp. 203–221. ISSN: 0022-3751.

[Day+05] Abhinav Dayal, Cliff Woolley, Benjamin Watson, and David Luebke. *"Adaptive Frameless Rendering."* In: *SIGGRAPH 2005 Courses*. Los Angeles, California: ACM, July 2005.

[Déc+03] Xavier Décoret, Frédo Durand, François X. Sillion, and Julie Dorsey. *"Billboard Clouds for Extreme Model Simplification."* In: *ACM Trans. Graph.* 22.3 (July 2003), pp. 689–696. ISSN: 0730-0301.

[DN02] M. F. Deering and D. Naegle. *"The SAGE Graphics Architecture."* In: *ACM Trans. Graph.* 21.3 (July 2002), pp. 683–692.

[Def99] Department of Defense. *Design Criteria Standard, Human Engineering, MIL-STD-1472F*. Tech. rep. United States of America, 1999.

[Dem04] Joe Demers. *"Depth of field: A survey of techniques."* In: *GPU Gems* 1.375 (2004), U390.

[Den+19] G. Denes, K. Maruszczyk, G. Ash, and R. K. Mantiuk. *"Temporal Resolution Multiplexing: Exploiting the limitations of spatio-temporal vision for more efficient VR rendering."* In: *IEEE Transactions on Visualization and Computer Graphics* 25.5 (May 2019), pp. 2072–2082.

[Deu+02] Oliver Deussen, Carsten Colditz, Marc Stamminger, and George Drettakis. *"Interactive Visualization of Complex Plant Ecosystems."* In: *Proceedings of the Conference on Visualization '02*. VIS '02. Boston, Massachusetts: IEEE Computer Society, 2002, pp. 219–226. ISBN: 0-7803-7498-3.

[Did+10a] Piotr Didyk, Elmar Eisemann, Tobias Ritschel, Karol Myszkowski, and Hans-Peter Seidel. *"Apparent Display Resolution Enhancement for Moving Images."* In: *ACM Transactions on Graphics (Proceedings SIGGRAPH 2010, Los Angeles)* 29.4 (2010).

[Did+10b] Piotr Didyk, Elmar Eisemann, Tobias Ritschel, Karol Myszkowski, and Hans-Peter Seidel. *"Perceptually-motivated Real-time Temporal Upsampling of 3D Content for High-refresh-rate Displays."* In: *Computer Graphics Forum*. Vol. 29. 2010, pp. 713–722. Eurographics '10.

[Did+10c] Piotr Didyk, Tobias Ritschel, Elmar Eisemann, Karol Myszkowski, and Hans-Peter Seidel. *"Adaptive Image-space Stereo View Synthesis."* In: *Vision, Modeling and Visualization Workshop - VMV*. 2010, pp. 299–306.

[Did+11] Piotr Didyk, Tobias Ritschel, Elmar Eisemann, Karol Myszkowski, and Hans-Peter Seidel. *"A Perceptual Model for Disparity."* In: *SIGGRAPH '11, ACM Transactions on Graphics (TOG)*. Vol. 30. ACM, 2011, pp. 96–104.

[DS07] Andreas Dietrich and Philipp Slusallek. *"Adaptive Spatial Sample Caching."* In: *Proceedings of the 2007 IEEE Symposium on Interactive Ray Tracing*. RT '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 141–147. ISBN: 978-1-4244-1629-5.

[DW85] Mark A. Z. Dippé and Erling Henry Wold. *"Antialiasing Through Stochastic Sampling."* In: *Proceedings of the 12th annual conference on Computer graphics and interactive techniques, SIGGRAPH '85* 19.3 (1985), pp. 69–78.

[Dob+05] Simon Dobbyn, John Hamill, Keith O'Conor, and Carol O'Sullivan. *"Geopostors: A Real-time Geometry / Impostor Crowd Rendering System."* In: *Proceedings of the 2005 Symposium on Interactive 3D Graphics and Games*. I3D '05. Washington, District of Columbia: ACM, 2005, pp. 95–102. ISBN: 1-59593-013-2.

[Dod04] Neil A. Dodgson. *"Variation and extrema of human interpupillary distance."* In: *Stereoscopic Displays and Virtual Reality Systems XI*. Ed. by Andrew J. Woods, John O. Merritt, Stephen A. Benton, and Mark T. Bolas. SPIE, May 2004.

[Don+14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. *"Learning a Deep Convolutional Network for Image Super-Resolution."* In: *Computer Vision – ECCV 2014*. Ed. by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars. Cham: Springer International Publishing, 2014, pp. 184–199.

[Don+04] Zhao Dong, Wei Chen, Hujun Bao, Hongxin Zhang, and Qunsheng Peng. *"Real-time Voxelization for Complex Polygonal Models."* In: *Proceedings of the Computer Graphics and Applications, 12th Pacific Conference*. PG '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 43–50.

[Dor+06] Michael Dorr, Martin Böhme, Thomas Martinetz, and Erhardt Barth. *"Gaze-Contingent Spatio-Temporal Filtering in a Head-Mounted Display."* In: ed. by Elisabeth André, Laila Dybkjær, Wolfgang Minker, Heiko Neumann, and Michael Weber. In Proceedings of Perception and Interactive Technologies: International Tutorial and Research Workshop, PIT '06. Springer Berlin Heidelberg, 2006, pp. 205–207.

[Dor+10] Michael Dorr, Thomas Martinetz, Karl R. Gegenfurtner, and Erhardt Barth. *"Variability of eye movements when viewing dynamic natural scenes."* In: *Journal of Vision* 10.10 (2010), pp. 28–44.

[DVB12] Michael Dorr, Eleonora Vig, and Erhardt Barth. *"Eye movement prediction and variability on natural video data sets."* In: *Visual Cognition* 20.4-5 (2012), pp. 495–514.

[Dou+03] R. F. Dougherty, V. M. Koch, A. A. Brewer, B. Fischer, J. Modersitzki, and B. A. Wandell. *"Visual field representations and locations of visual areas V1/2/3 in human visual cortex."* In: *J Vis* 3.10 (2003), pp. 586–598.

[Dre+07] George Drettakis, Nicolas Bonneel, Carsten Dachsbacher, Sylvain Lefebvre, Michael Schwarz, and Isabelle Viaud-Delmon. *"An Interactive Perceptual Rendering Pipeline using Contrast and Spatial Masking."* In: *Proceedings of the 18th Eurographics conference on Rendering Techniques*. ACM EGSR '07. 2007, pp. 297–308.

[Duc+97] Mark Duchaineau, Murray Wolinsky, David E. Sigeti, Mark C. Miller, Charles Aldrich, and Mark B. Mineev-Weinstein. *"ROAMing Terrain: Real-time Optimally Adapting Meshes."* In: *Proceedings of the 8th Conference on Visualization '97*. VIS '97. Phoenix, Arizona, USA: IEEE Computer Society Press, 1997, pp. 81–88. ISBN: 1-58113-011-2.

[Duc07] Andrew T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007. ISBN: 1846286085.

[Duc+14] Andrew T. Duchowski, Donald H. House, Jordan Gestring, Rui I. Wang, Krzysztof Krejtz, Izabela Krejtz, Radoslaw Mantiuk, and Bartosz Bazyluk. *"Reducing Visual Discomfort of 3D Stereoscopic Displays with Gaze-contingent Depth-of-field."* In: *Proceedings of the ACM SAP '14*. Vancouver, British Columbia, Canada: ACM, 2014, pp. 39–46. ISBN: 978-1-4503-3009-1.

[Duc+11] Andrew T. Duchowski, Brandon Pelfrey, Donald H. House, and Rui Wang. *"Measuring Gaze Depth with an Eye Tracker During Stereoscopic Display."* In: *Proceedings of ACM APGV '11*. Toulouse, France: ACM, 2011, pp. 15–22. ISBN: 978-1-4503-0889-2.

[Ebe13] David Eberly. *Distance from a Point to an Ellipse, an Ellipsoid, or a Hyperellipsoid*. Geometric Tools, LLC. 2013. http://www.geometrictools.com/Documentation/DistancePointEllipseEllipsoid.pdf, last visited 21. Nov. 2018.

[Eck+10] Jennifer L. Ecker, Olivia N. Dumitrescu, Kwoon Y. Wong, Nazia M. Alam, Shih-Kuo Chen, Tara LeGates, Jordan M. Renna, Glen T. Prusky, David M. Berson, and Samer Hattar. *"Melanopsin-Expressing Retinal Ganglion-Cell Photoreceptors: Cellular Diversity and Role in Pattern Vision."* In: *Neuron* 67.1 (2010), pp. 49–60. ISSN: 0896-6273.

[Ed17] Pablo Artal (Ed.) *Handbook of Visual Optics, Volume One: Fundamentals and Eye Optics (Volume 1)*. CRC Press, 2017. ISBN: 1482237857.

[Edw09] Keith H. Edwards. *Optometry: Science, Techniques and Clinical Management*. en. Elsevier Health Sciences, 2009. ISBN: 978-0-7506-8778-2.

[Egi+06] Karen Egiazarian, Jaakko Astola, Nikolay Ponomarenko, Vladimir Lukin, Federica Battisti, and Marco Carli. *"New full-reference quality metrics based on HVS."* In: *Second International Workshop on Video Processing and Quality Metrics*. Vol. 4. 2006.

[EMU15] Gabriel Eilertsen, Rafał K. Mantiuk, and Jonas Unger. *"Real-time noise-aware tone mapping."* In: *ACM Transactions on Graphics (TOG), SIGGRAPH Asia '15* 34.6 (2015), pp. 198–222.

[Eil+13] Gabriel Eilertsen, Jonas Unger, Robert Wanat, and Rafał Mantiuk. *"Survey and Evaluation of Tone Mapping Operators for HDR Video."* In: *ACM SIGGRAPH 2013 Talks*. Anaheim, California: ACM, 2013, 11:1–11:1. ISBN: 978-1-4503-2344-4.

[ED08] Elmar Eisemann and Xavier Décoret. *"Single-pass GPU Solid Voxelization for Real-Time Applications."* In: *Proceedings of Graphics Interface 2008* (May 2008), pp. 73–80.

[Eng09] Wolfgang Engel. *"Light Pre-Pass - Deferred Lighting: Latest Development."* In: *Advances in Real-Time Rendering in Games*. Ed. by Rockstar Games. SIGGRAPH Course, 2009. https://advances.realtimerendering.com/s2009/LightPrePass.ppt, last visited 10. Nov. 2019.

[ED00] James T. Enns and Vincent Di Lollo. *"What's new in visual masking?"* In: *Trends in Cognitive Sciences* 4.9 (2000), pp. 345–352.

[ER66] C. Enroth-Cugell and J. G. Robson. *"The contrast sensitivity of retinal ganglion cells of the cat."* In: *J. Physiol. (Lond.)* 187.3 (Dec. 1966), pp. 517–552.

[Eri97] Eric Horvitz and Jed Lengyel. *"Perception, Attention, and Resources: A Decision-Theoretic Approach to Graphics Rendering."* In: *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence, UAI '97*. Morgan Kaufmann, 1997, pp. 238–249.

[EPR04] Kai Essig, Marc Pomplun, and Helge Ritter. *"Application of a Novel Neural Approach to 3D Gaze Tracking: Vergence Eye-Movements in Autostereogram."* In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. 26. 2004.

[Fai05] Mark D. Fairchild. *Color Appearance Models*. en. John Wiley & Sons, 2005. ISBN: 978-0-470-01269-7.

[Fai15] Mark D. Fairchild. *"Seeing, adapting to, and reproducing the appearance of nature."* In: *Applied Optics* 54.4 (2015), pp. 107–116.

[FC00] S. Fang and H. Chen. *"Hardware Accelerated Voxelisation."* In: *Volume Graphics*. Ed. by M. Chen, A.E. Kaufman, and R. Yagel. Springer, London, 2000.

[FP04] Jean-Philippe Farrugia and Bernard Péroche. *"A Progressive Rendering Algorithm Using an Adaptive Perceptually Based Image Metric."* In: *Computer Graphics Forum*. Vol. 23. Eurographics '04 3. 2004, pp. 605–614.

[Fei+07] Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. *"What do we perceive in a glance of a real-world scene?"* In: *Journal of Vision* 7.1 (2007), pp. 1–29.

[FMR08] P. Felzenszwalb, D. McAllester, and D. Ramanan. *"A Discriminatively Trained, Multiscale, Deformable Part Model."* In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR '08*. IEEE. 2008, pp. 1–8.

[Fer+96] James A. Ferwerda, Sumanta N. Pattanaik, Peter Shirley, and Donald P. Greenberg. *"A Model of Visual Adaptation for Realistic Image Synthesis."* In: *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '96. ACM, 1996, pp. 249–258. ISBN: 0-89791-746-4.

[Fer+97] James A. Ferwerda, Peter Shirley, Sumanta N. Pattanaik, and Donald P. Greenberg. *"A model of visual masking for computer graphics."* In: *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '97. ACM, 1997, pp. 143–152.

[FH03] A. P. Field and G. Hole. *How to Design and Report Experiments*. Sage Publications Limited, 2003. ISBN: 9780761973836.

[FWK63] M. C. Flom, F. W. Weymouth, and D. Kahneman. *"Visual Resolution And Contour Interaction."* eng. In: *Journal of the Optical Society of America* 53 (1963), pp. 1026–1032. ISSN: 0030-3941.

[FS05] Tim Foley and Jeremy Sugerman. *"KD-tree Acceleration Structures for a GPU Raytracer."* In: *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS Conference on Graphics Hardware*. HWWS '05. Los Angeles, California: ACM, 2005, pp. 15–22. ISBN: 1-59593-086-8.

[Fov17] Yuka Kojima Fove Inc. *Fove 0 HMD*. Electronic. Nov. 2017. https://www.getfove.com, last visited 13. Nov. 2019.

[FWM15] Simone Frintrop, Thomas Werner, and German Martin Garcia. *"Traditional Saliency Reloaded: A Good Old Model in New Shape."* In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR '15*. 2015, pp. 82–90.

[FRS19] Sebastian Friston, Tobias Ritschel, and Anthony Steed. *"Perceptual Rasterization for Head-mounted Display Image Synthesis."* In: *ACM Trans. Graph.* 38.4 (July 2019), 97:1–97:14. ISSN: 0730-0301.

[Fri+16] Sebastian Friston, Anthony Steed, Simon Tilbury, and Georgi Gaydadjiev. *"Construction and Evaluation of an Ultra Low Latency Frameless Renderer for VR."* In: *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 22.4 (2016), pp. 1377–1386.

[FBP96] Sheng Fu, Hujun Bao, and Qunsheng Peng. *"An accelerated rendering algorithm for stereoscopic display."* In: *Computers & Graphics* 20.2 (1996), pp. 223–229.

[FH14] Masahiro Fujita and Takahiro Harada. *Foveated Real-Time Ray Tracing for Virtual Reality Headset*. 2014. SIGGRAPH Asia '14 - Poster.

[GVC03] R. Gaborski, Vishal S. Vaingankar, and R.L. Canosa. *"Goal directed visual search based on color cues: Cooperative effects of top-down & bottom-up visual attention."* In: *Proceedings of the Artificial Neural Networks in Engineering, Rolla, Missouri* 13 (2003), pp. 613–618.

[GC06] Ran Gal and Daniel Cohen-Or. *"Salient Geometric Features for Partial Shape Matching and Similarity."* In: *ACM Transactions on Graphics (TOG)* 25.1 (Jan. 2006), pp. 130–150.

[GDS14] Steven Galea, Kurt Debattista, and Sandro Spina. *"GPU-Based Selective Sparse Sampling for Interactive High-Fidelity Rendering."* In: *6th International Conference on Games and Virtual Worlds for Serious Applications*. IEEE VS-GAMES '14. 2014, pp. 1–8.

[GH97] Michael Garland and Paul S. Heckbert. *"Surface Simplification Using Quadric Error Metrics."* In: *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '97. New York, NY, USA: ACM Press/Addison–Wesley Publishing Co., 1997, pp. 209–216. ISBN: 0-89791-896-7.

[GO12] Eduardo S. L. Gastal and Manuel M. Oliveira. *"Adaptive Manifolds for Real-Time High-Dimensional Filtering."* In: *ACM TOG* 31.4 (2012), 33:1–33:13. Proceedings of SIGGRAPH 2012.

[Gei84] W. S. Geisler. *"Physical limits of acuity and hyperacuity."* In: *J Opt Soc Am A* 1.7 (July 1984), pp. 775–782.

[GF16] Iliyan Georgiev and Marcos Fajardo. *"Blue-noise Dithered Sampling."* In: *ACM SIGGRAPH 2016 Talks*. SIGGRAPH '16. Anaheim, California: ACM, 2016, 35:1–35:1. ISBN: 978-1-4503-4282-7.

[GWH17] Björn Ludolf Gerdau, Martin Weier, and André Hinkenjann. *"Containerized Distributed Rendering for Interactive Environments."* In: *Virtual Reality and Augmented Reality*. Ed. by Jernej Barbic, Mirabelle D'Cruz, Marc Erich Latoschik, Mel Slater, and Patrick Bourdot. Cham: Springer International Publishing, 2017, pp. 69–86. ISBN: 978-3-319-72323-5.

[GHR84] M. J. Gervais, L. O. Harvey, and J. O. Roberts. *"Identification confusions among letters of the alphabet."* eng. In: *Journal of Experimental Psychology. Human Perception and Performance* 10.5 (Oct. 1984), pp. 655–666. ISSN: 0096-1523.

[GM05] Enrico Gobbetti and Fabio Marton. *"Far Voxels: A Multiresolution Framework for Interactive Rendering of Huge Complex 3D Models on Commodity Graphics Platforms."* In: *ACM SIGGRAPH 2005 Papers*. SIGGRAPH '05. Los Angeles, California: ACM, 2005, pp. 878–885.

[Gol01] E. Bruce Goldstein. *Blackwell Handbook of Perception*. Oxford, UK Malden, Mass., USA: Blackwell, 2001. ISBN: 0631206841.

[Gol10] E. Bruce Goldstein. *Encyclopedia of perception*. SAGE Publications, Inc, 2010. ISBN: 978-1412940818.

[Gol13] E. Bruce Goldstein. *Sensation and Perception*. 9th. Pacific Grove: Wadsworth-Thomson Learning, 2013. ISBN: 1-133-95849-4.

[GW07] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Englisch. 3rd ed. Upper Saddle River, N.J: Prentice Hall, 2007. ISBN: 978-0-13-168728-8.

[Gör15] Jonas Alexander Göransson. *"A Temporal Stable Distance To Eedge Anti-Aliasing Technique For GCN Architexture."* MA thesis. Karlskrona, Sweden: Blekinge Institute of Technology, May 2015. 54 pp. MECS-2015-05.

[Gra69] Yves Le Grand. *Light, Colour and Vision*. Trans. by J. W. T. Walsh translated from the French by R. W. G. Hunt and F. R. W. Hunt. Second Edition. Chapman, Hall, London (U. S. distributor, Barnes, and Noble, New York), 1969.

[GA30] Ragnar Granit and Winona von Ammon. *"Comparative Studies On The Peripheral And Central Retina."* In: *American Journal of Physiology - Legacy Content* 95.1 (1930), pp. 229–241. ISSN: 0002-9513.

[GPB80] D. G. Green, M. K. Powers, and M. S. Banks. *"Depth of focus, eye size and visual acuity."* In: *Vision Research* 20.10 (1980), pp. 827–835.

[Gre70] Daniel G. Green. *"Regional variations in the visual acuity for interference fringes on the retina."* In: *The Journal of Physiology* 207.2 (1970), pp. 351–356.

[Gre+09] David Grelaud, Nicolas Bonneel, Michael Wimmer, Manuel Asselot, and George Drettakis. *"Efficient and Practical Audio-visual Rendering for Games Using Crossmodal Perception."* In: *Proceedings of the 2009 Symposium on Interactive 3D Graphics and Games*. I3D '09. ACM, 2009, pp. 177–182.

[Gro05] Herbert Gross. *Handbook of Optical Systems*. Vol. 4. Wiley-VCH, 2005. ISBN: 978-3527403806. Volume 4: Survey of Optical Instruments.

[GP07] Markus Gross and Hanspeter Pfister. *Point-Based Graphics*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007. ISBN: 0123706041.

[Gru+06] Mark Grundland, Rahul Vohra, Gareth P. Williams, and Neil A. Dodgson. *"Cross Dissolve Without Cross Fade: Preserving Contrast, Color and Salience in Image Compositing."* In: *Computer Graphics Forum* (2006).

[Gue+12] Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. *"Foveated 3D Graphics."* In: *SIGGRAPH Asia '12, ACM Transactions on Graphics (TOG)* 31.6 (Nov. 2012), 164:1–164:10.

[Gun98] Steve R. Gunn. *Support Vector Machines for Classification and Regression*. Tech. rep. University of Southhampton, May 1998.

[Guo98] Baining Guo. *"Progressive Radiance Evaluation Using Directional Coherence Maps."* In: *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '98. ACM, 1998, pp. 255–266.

[Guo+15] Jinjiang Guo, Vincent Vidal, Atilla Baskurt, and Guillaume Lavoué. *"Evaluating the Local Visibility of Geometric Artifacts."* In: *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception*. SAP '15. ACM, 2015, pp. 91–98. ISBN: 978-1-4503-3812-7.

[Gut+05] Diego Gutierrez, Oscar Anson, Adolfo Munoz, and Francisco Seron. *"Perception-based Rendering: Eyes Wide Bleached."* In: *Eurographics '05, Short Presentations* 5 (2005), pp. 49–52.

[Hab+01] Jörg Haber, Karol Myszkowski, Hitoshi Yamauchi, and Hans-Peter Seidel. *"Perceptually guided corrective splatting."* In: *The European Association for Computer Graphics 22th Annual Conference, Computer Graphics Forum*. Vol. 20. Eurographics '01 3. 2001, pp. 142–153.

[HT00] Nouchine Hadjikhani and Roger B. H. Tootell. *"Projection of rods and cones within human visual cortex."* In: *Human Brain Mapping* 9.1 (2000), pp. 55–63.

[HA90] Paul Haeberli and Kurt Akeley. *"The Accumulation Buffer: Hardware Support for High-quality Rendering."* In: *Proceedings of the 17th Annual Conference on Computer Graphics and Interactive Techniques*. Vol. 24. SIGGRAPH '90 4. Dallas, TX, USA, 1990, pp. 309–318.

[HPG09] Thorsten Hansen, Lars Pracejus, and Karl R. Gegenfurtner. *"Color perception in the intermediate periphery of the visual field."* In: *Journal of Vision* 9.4 (2009), 26:1–26:12.

[HKP07] Jonathan Harel, Christof Koch, and Pietro Perona. *"Graph-Based Visual Saliency."* In: *Advances in Neural Information Processing Systems 19*. 2007, pp. 545–552.

[HH04] Shawn Hargreaves and Mark Harris. NVIDIA Cooperation. Sept. 2004. URL: http://download.nvidia.com/developer/presentations/2004/6800_Leagues/6800_Leagues_Deferred_Shading.pdf. last visited 10. Nov. 2019.

[Har+16] Carlo Harvey, Kurt Debattista, Thomas Bashford-Rogers, and Alan Chalmers. *"Multi-Modal Perception for Selective Rendering."* In: *Computer Graphics Forum*. 2016.

[HCS10] Jasminka Hasic, Alan Chalmers, and Elena Sikudova. *"Perceptually Guided High-fidelity Rendering Exploiting Movement Bias in Visual Attention."* In: *ACM Transactions on Applied Perception (TAP)* 8.1 (2010), 6:1–6:19. ISSN: 1544-3558.

[HAO05] Jon Hasselgreen, Tomas Akenine-Möller, and Lennart Ohlsson. *"42. Conservative Rasterization."* In: *GPU Gems 2*. NVIDIA, Addison-Wesley, 2005, pp. 677–690.

[HGF14] Yong He, Yan Gu, and Kayvon Fatahalian. *"Extending the Graphics Pipeline with Adaptive, Multi-rate Shading."* In: *ACM Transactions on Graphics (TOG)*. Vol. 33. SIGGRAPH '14 4. ACM, 2014, 142:1–142:12.

[HSD13] Daniel Heck, Thomas Schlömer, and Oliver Deussen. *"Blue Noise Sampling with Controlled Aliasing."* In: *ACM Trans. Graph.* 32.3 (July 2013), 25:1–25:12.

[Hee91] D. W. Heeley. *"Spatial frequency difference thresholds depend on stimulus area."* In: *Spat Vis* 5.3 (1991), pp. 205–217.

[Hel+10] Robert T. Held, Emily A. Cooper, James F. O'Brien, and Martin S. Banks. *"Using Blur to Affect Perceived Distance and Size."* In: *ACM Trans. Graph.* 29.2 (Apr. 2010), 19:1–19:16. ISSN: 0730-0301.

[Hel24] H. Helmholtz. *Treatise on Physiological Optics*. Vol. 1. New York: Dover, 1924.

[Hel67] Hermann Ludwig Ferdinand von Helmholtz. *Handbuch der physiologischen Optik*. ger. Leopold Voss, 1867.

[Hen+07] John M. Henderson, James R. Brockmole, Monica S. Castelhano, and Michael Mack. *"Visual saliency does not account for eye movements during visual search in real-world scenes."* In: *Eye movements: A window on mind and brain* (2007), pp. 537–562.

[HH99] John M. Henderson and Andrew Hollingworth. *"High-Level Scene Perception."* In: *Annual Review of Psychology* 50.1 (Feb. 1999), pp. 243–271.

[Hen+13] John M. Henderson, Steven G. Luke, Joseph Schmidt, and John E. Richards. *"Co-registration of eye movements and event-related potentials in connected-text paragraph reading."* In: *Frontiers in Systems Neuroscience* 7 (2013), 28:1–28:13.

[Hen05] Anita Hendrickson. *"Organization of the Adult Primate Fovea."* In: *Macular Degeneration*. Ed. by Philip L. Penfold and Jan M. Provis. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 1–23.

[HC98] Stewart H. C. Hendry and David J. Calkins. *"Neuronal chemistry and functional organization in the primate visual system."* In: *Trends in Neurosciences* 21.8 (1998), pp. 344–349. ISSN: 0166-2236.

[Her20] Ewald Hering. *Grundzüge der Lehre vom Lichtsinn*. Springer, 1920.

[Her+12] Robert Herzog, Martin Čadík, Tunç O. Aydčin, Kwang In Kim, Karol Myszkowski, and Hans-P. Seidel. *"NoRM: No-Reference Image Quality Metric for Realistic Image Synthesis."* In: *Computer Graphics Forum*. Vol. 31. 2012, pp. 545–554.

[Hil+07] Sébastien Hillaire, Anatole Lécuyer, Rémi Cozot, and Géry Casiez. *"Depth-of-field Blur Effects for First-person Navigation in Virtual Environments."* In: *Proceedings of the 2007 ACM Symposium on Virtual Reality Software and Technology*. VRST '07. Newport Beach, California, 2007, pp. 203–206. ISBN: 978-1-59593-863-3.

[HK97] I. Hontsch and L. J. Karam. *"APIC: adaptive perceptual image coding based on subband decomposition with locally adaptive perceptual weighting."* In: *Proceedings of the International Conference on Image Processing, 1997*. Vol. 1. Oct. 1997, pp. 37–40.

[Hop96] Hugues Hoppe. *"Progressive Meshes."* In: *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '96. New York, NY, USA: ACM, 1996, pp. 99–108. ISBN: 0-89791-746-4.

[Hop97] Hugues Hoppe. *"View-dependent Refinement of Progressive Meshes."* In: *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '97. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1997, pp. 189–198. ISBN: 0-89791-896-7.

[HH91] J. C. Horton and W. F. Hoyt. *"The representation of the visual field in human striate cortex. A revision of the classic Holmes map."* In: *Archives of Ophthalmology* 109.6 (1991), pp. 816–824.

[How12] Ian P. Howard. *Perceiving in Depth, Volume 1: Basic Mechanisms*. Oxford University Press, 2012. ISBN: 019976414X.

[HR95] Ian P. Howard and Brian J. Rogers. *Binocular vision and stereopsis*. Oxford University Press, USA, 1995. ISBN: 0195084764.

[HHO04] Sarah Howlett, John Hamill, and Carol O'Sullivan. *"An Experimental Approach to Predicting Saliency for Simplified Polygonal Models."* In: *Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization*. APGV '04. Los Angeles, California, USA: ACM, 2004, pp. 57–64. ISBN: 1-58113-914-4.

[HHO05] Sarah Howlett, John Hamill, and Carol O'Sullivan. *"Predicting and Evaluating Saliency for Simplified Polygonal Models."* In: *ACM Transactions on Applied Perception (TAP)* 2.3 (July 2005), pp. 286–308. ISSN: 1544-3558.

[HSH10] Liang Hu, Pedro V. Sander, and Hugues Hoppe. *"Parallel View-Dependent Level-of-Detail Control."* In: *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 16.5 (2010), pp. 718–728.

[Hua65] T. Huang. *"The subjective effect of two-dimensional pictorial noise."* In: *IEEE Transactions on Information Theory* 11.1 (Jan. 1965), pp. 43–53.

[Hub88] David Hubel. *Eye, brain, and vision*. New York: Scientific American Library Distributed by W. H. Freeman, 1988. ISBN: 0716760096.

[HW04] David H. Hubel and Torsten N. Wiesel. *Brain and Visual Perception: The Story of a 25-Year Collaboration*. Oxford University Press, USA. 2004. ISBN: 9780195176186.

[Hul+09] Vedad Hulusić, Gabriela Czanner, Kurt Debattista, Elena Sikudova, Piotr Dubla, and Alan Chalmers. *"Investigation of the Beat Rate Effect on Frame Rate for Animated Content."* In: *Proceedings of the 25th Spring Conference on Computer Graphics*. SCCG '09. Budmerice, Slovakia: ACM, 2009, pp. 151–159. ISBN: 978-1-4503-0769-7.

[Hul+12] Vedad Hulusic, Carlo Harvey, Kurt Debattista, Nicolas Tsingos, Steve Walker, David Howard, and Alan Chalmers. *"Acoustic Rendering and Auditory-Visual Cross-Modal Perception and Interaction."* In: *Computer Graphics Forum* 31.1 (2012), pp. 102–131.

[Hun91] R. W. G. Hunt. *"Revised colour-appearance model for related and unrelated colours."* In: *Color Research & Application* 16.3 (1991), pp. 146–165.

[Hun94] R. W. G. Hunt. *"An Improved Predictor of Colourfulness in a Model of Colour Vision."* In: *Color Research & Application* 19.1 (1994), pp. 23–26.

[Hun15] Warren Hunt. *Virtual Reality: The Next Great Graphics Revolution*. Keynote Talk HPG. 2015. URL: http://www.highperformancegraphics.org/wp-content/uploads/2015/Keynote1/WarrenHuntHPGKeynote.pptx.

[Hun17] Warren Hunt. *"Real-Time Ray Casting for Virtual Reality."* In: *HPG 2017, Keynote*. Oculus Research. ACM, July 2017. http://www.highperformancegraphics.org/wp-content/uploads/2017/Hot3D/HPG2017_RealTimeRayCasting.pptx, last visited 21. Nov. 2018.

[HMN18] Warren Hunt, Michael Mara, and Alex Nankervis. *"Hierarchical Visibility for Virtual Reality."* In: *Proc. ACM Comput. Graph. Interact. Tech.* 1.1 (July 2018), 8:1–8:18. ISSN: 2577-6193.

[HJ57] Leo M. Hurvich and Dorothea Jameson. *"An opponent-process theory of color vision."* In: *Psychological review* 64.1(6) (1957), pp. 384–404.

[HG08] Q. Huynh-Thu and M. Ghanbari. *"Scope of validity of PSNR in image/video quality assessment."* In: *Electronics Letters* 44.13 (June 2008), pp. 800–801.

[Ige99] Homan Igehy. *"Tracing Ray Differentials."* In: *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '99. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1999, pp. 179–186. ISBN: 0-201-48560-5.

[II91] Robert J. Marks II. *Introduction to Shannon Sampling and Interpolation Theory*. New York: Springer-Verlag, 1991. ISBN 0-387-97391-5.

[Ima19] Imagination Technologies. *Ray Tracing White Paper - Shining a Light on Ray Tracing*. Electronic. May 2019. https://cdn2.imgtec.com/whitepapers/powervr/ray-tracing/powervr-shining-a-light-on-ray-tracing.pdf, last visited 29. Oct. 2019.

[IYP09] Konstantine Iourcha, Jason C. Yang, and Andrew Pomianowski. *"A Directionally Adaptive Edge Anti-aliasing Filter."* In: *Proceedings of the Conference on High Performance Graphics 2009*. HPG '09. New Orleans, Louisiana: ACM, 2009, pp. 127–133. ISBN: 978-1-60558-603-8.

[IK99] T. Isshiki and H. Kunieda. *"Efficient anti-aliasing algorithm for computer generated images."* In: *Circuits and Systems, 1999. ISCAS '99. Proceedings of the 1999 IEEE International Symposium on*. July 1999, 532–535 Vol.4.

[IK01] Laurent Itti and Christof Koch. *"Computational modelling of visual attention."* In: *Nature Reviews Neuroscience* 2.3 (2001), pp. 194–203.

[IKN98] Laurent Itti, Christof Koch, and Ernst Niebur. *"A Model of Saliency-Based Visual Attention for Rapid Scene Analysis."* In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 20.11 (1998), pp. 1254–1259.

[Jac+11] Skyler L. Jackman, Norbert Babai, James J. Chambers, Wallace B. Thoreson, and Richard H. Kramer. *"A Positive Feedback Synapse from Retinal Horizontal Cells to Cone Photoreceptors."* In: *PLOS Biology* 9.5 (2011). Ed. by Fred Rieke.

[Jac+15] David E. Jacobs, Orazio Gallo, Emily A. Cooper, Kari Pulli, and Marc Levoy. *"Simulating the Visual Experience of Very Bright and Very Dark Scenes."* In: *ACM Trans. Graph.* 34.3 (May 2015), 25:1–25:15. ISSN: 0730-0301.

[JOK09] Lina Jansen, Selim Onat, and Peter König. *"Influence of disparity on fixation and saccades in free viewing of natural scenes."* In: *Journal of Vision* 9.1 (2009), pp. 1–19.

[Jan01] Ruud Janssen. *Computational image quality.* Bellingham, Wash., USA: SPIE Press, 2001. ISBN: 9780819441324.

[Jar+12] Adrian Jarabo, Tom Van Eyck, Veronica Sundstedt, Kavita Bala, Diego Gutierrez, and Carol O'Sullivan. *"Crowd Light: Evaluating the Perceived Fidelity of Illuminated Dynamic Scenes."* In: *Computer Graphics Forum, Eurographics '12* 31.2, 4 (May 2012), pp. 565–574. ISSN: 0167-7055.

[Jim17] Jorge Jimenez. *"Dynamic Temporal Antialiasing in Call of Duty: Infinite Warfare."* In: *Advances in Real-Time Rendering in Games.* Ed. by Activision. SIGGRAPH Course, 2017.

[Jim+12] Jorge Jimenez, Jose I. Echevarria, Tiago Sousa, and Diego Gutierrez. *"SMAA: Enhanced Subpixel Morphological Antialiasing."* In: *Computer Graphics Forum.* Vol. 31. Eurographics '12. 2012, pp. 355–364.

[Jim+11] Jorge Jimenez et al. *"Filtering Approaches for Real-Time Anti-Aliasing."* In: *SIGGRAPH Courses* 2.3 (2011), p. 4.

[Jin+09] Bongjun Jin, Insung Ihm, Byungjoon Chang, Chanmin Park, Wonjong Lee, and Seokyoon Jung. *"Selective and Adaptive Supersampling for Real-Time Ray Tracing."* In: *Proceedings of the Conference on High Performance Graphics.* HPG '09. ACM, 2009, pp. 117–125.

[JFN98] Elaine W. Jin, X.-F. Feng, and John Newell. *"The development of a color visual difference model (CVDM)."* In: *International Technical Conference on Digital Image Capture and Associated System, Reproduction and Image Quality Technologies.* IS&T's Pics Conference '98. Society for Imaging Science & Technology, 1998, pp. 154–158.

[JSF17] J. Jo, J. Seo, and J. Fekete. *"A progressive k-d tree for approximate k-nearest neighbors."* In: *2017 IEEE Workshop on Data Systems for Interactive Analysis (DSIA).* Oct. 2017, pp. 1–5.

[JF02] Garrett M. Johnson and Mark D. Fairchild. *"On Contrast Sensitivity in an Image Difference Model."* In: *International Technical Conference on Digital Image Capture and Associated System, Reproduction and Image Quality Technologies.* IS&T's Pics Conference '02. Society for Imaging Science & Technology, 2002, pp. 18–23.

[JDT12] Tilke Judd, Frédo Durand, and Antonio Torralba. *"A Benchmark of Computational Models of Saliency to Predict Human Fixations."* In: *MIT Computer Science and Artificial Intelligence Laboratory Technical Report* (2012). MIT-CSAIL-TR-2012-001.

[K90] Nakayama K. *"Properties of early motion processing: Implications for the sensing of ego motion."* In: *The Perception and Control of Self Motion.* Ed. by R. Warren and A. H. Wertheim. Hillsdale, NJ: Lawrence Erlbaum, 1990, pp. 69–80.

[Kai09] P. K. Kaiser. *"Prospective evaluation of visual acuity assessment: a comparison of snellen versus ETDRS charts in clinical practice (An AOS Thesis)."* In: *Trans Am Ophthalmol Soc* 107 (Dec. 2009), pp. 311–324.

[Kak+07] Masanori Kakimoto, Tomoaki Tatsukawa, Yukiteru Mukai, and Tomoyuki Nishita. *"Interactive Simulation of the Human Eye Depth of Field and Its Correction by Spectacle Lenses."* In: *Computer Graphics Forum* (2007). ISSN: 1467-8659.

[KBS15] Nima Khademi Kalantari, Steve Bako, and Pradeep Sen. *"A Machine Learning Approach for Filtering Monte Carlo Noise."* In: *ACM Trans. Graph.* 34.4 (July 2015), 122:1–122:12. ISSN: 0730-0301.

[KSA13] Viktor Kämpe, Erik Sintorn, and Ulf Assarsson. *"High Resolution Sparse Voxel DAGs."* In: *ACM Trans. Graph.* 32.4 (July 2013), 101:1–101:13.

[Kan+14] Le Kang, Peng Ye, Yi Li, and David Doermann. *"Convolutional neural networks for no-reference image quality assessment."* In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1733–1740.

[Kar14] Brian Karis. *"High-Quality Temporal Supersampling."* In: *Advances in Real-Time Rendering in Games*. Ed. by Inc. Epic Games. SIGGRAPH Course, 2014.

[KK95] Milton Katz and Philip B. Kruger. *"The Human Eye as an Optical System."* In: *Duane's Clinical Ophthalmology*. Ed. by Tasman W. Jaeger. Philidelphia: Lippincott and Raven, 1995. Chap. 33.

[KW17] Matthew Kay and Jacob O. Wobbrock. *ARTool: Aligned Rank Transform*. Electronic. Apr. 2017. https://cran.r-project.org/web/packages/ARTool/index.html, last visited 23. May 2018.

[Kel+15] Petr Kellnhofer, Tobias Ritschel, Karol Myszkowski, and Hans-Peter Seidel. *"A Transformation-Aware Perceptual Image Metric."* In: *SPIE/IS&T Electronic Imaging*. International Society for Optics and Photonics, 2015, pp. 939408, 1–14.

[Kel61] D. H. Kelly. *"Visual Responses to Time-Dependent Stimuli."* en. In: vol. 51. 7. Journal of the Optical Society of America, July 1961. Chap. II Single-Channel Model of the Photopic Visual System, p. 747.

[Kel79] D. H. Kelly. *"Motion and vision."* In: vol. 69. 10. Journal of the Optical Society of America, Oct. 1979. Chap. II. Stabilized spatio-temporal threshold surface, pp. 1340–1349.

[Khu08] A. K. Khurana. *Theory And Practice Of Optics And Refraction 2nd Edition*. Elsevier, 2008. ISBN: 8131211320.

[KRK11] Min H. Kim, Tobias Ritschel, and Jan Kautz. *"Edge-aware Color Appearance."* In: *ACM Trans. Graph.* 30.2 (Apr. 2011), 13:1–13:9. ISSN: 0730-0301.

[Kim+10] Youngmin Kim, Amitabh Varshney, David W. Jacobs, and François Guimbretière. *"Mesh Saliency and Human Eye Fixations."* In: *ACM Transactions on Applied Perception (TAP)* 7.2 (2010), 12:1–12:13.

[Kis+14] Naohiro Kishishita, Kiyoshi Kiyokawa, Ernst Kruijff, Jason Orlosky, Tomohiro Mashita, and Haruo Takemura. *"Analysing the Effects of a Wide Field of View Augmented Reality Display on Search Performance in Divided Attention Tasks."* In: *IEEE International Symposium on Mixed and Augmented Reality*. ISMAR '14. IEEE, 2014, pp. 177–186. ISBN: 978-1-4799-6184-9.

[Koh+05] Chin Chye Koh, Sanjit K. Mitra, John M. Foley, and Ingrid E. J. Heynderickx. *"Annoyance of individual artifacts in MPEG-2 compressed video and their relation to overall annoyance."* In: vol. 5666. 2005, pp. 5666 - 5666 - 12.

[KMH95] Craig Kolb, Don Mitchell, and Pat Hanrahan. *"A Realistic Camera Model for Computer Graphics."* In: *Proceedings of the 22d Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '95. ACM, 1995, pp. 317–324. ISBN: 0897917014.

[Kos+17] Matias Koskela, Kalle Immonen, Timo Viitanen, Pekka Jääskeläinen, Joonas Multanen, and Jarmo Takala. *"Foveated Instant Preview for Progressive Rendering."* In: *SIGGRAPH Asia Technical Briefs 2017*. ACM, Nov. 2017.

[Kou+14a] George Alex Koulieris, George Drettakis, Douglas Cunningham, and Katerina Mania. "*An Automated High-Level Saliency Predictor for Smart Game Balancing*." In: *ACM Transactions on Applied Perception (TAP)* 11.4 (Dec. 2014), 17:1–17:21.

[Kou+14b] George Alex Koulieris, George Drettakis, Douglas Cunningham, and Katerina Mania. "*C-LOD: Context-aware Material Level-of-Detail applied to Mobile Graphics*." In: *Computer Graphics Forum, Eurographics Symposium on Rendering, EGSR '14* 33.4 (2014), pp. 41–49.

[Kow11] Eileen Kowler. "*Eye movements: The past 25 years*." In: *Vision Research* 51.13 (2011), pp. 1457–1483.

[Kri+05] J. Krivanek, P. Gautron, S. Pattanaik, and K. Bouatouch. "*Radiance caching for efficient global illumination computation*." In: *IEEE Transactions on Visualization and Computer Graphics* 11.5 (Sept. 2005), pp. 550–561.

[Kři+10] Jaroslav Křivánek, Marcos Fajardo, Per H. Christensen, Eric Tabellion, Michael Bunnell, David Larsson, and Anton Kaplanyan. "*Global Illumination Across Industries*." In: *ACM SIGGRAPH 2010 Courses*. SIGGRAPH '10. Los Angeles, California: ACM, 2010. ISBN: 978-1-4503-0395-8.

[KSW06] J. Krüger, J. Schneider, and R. Westermann. "*ClearView: An Interactive Context Preserving Hotspot Visualization Technique*." In: *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 12.5 (2006), pp. 941–948. ISSN: 1077-2626.

[KAB15] Srinivas S. S. Kruthiventi, Kumar Ayush, and R. Venkatesh Babu. "*DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations*." In: *arXiv* (2015). 1510.02927.

[Kuh17] Max Kuhn. *The caret Package, 15 Variable Importance*. Electronic. Apr. 2017. http://topepo.github.io/caret/variable-importance.html, last visited 21. Nov. 2017.

[KTB14] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. "*Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet*." In: *arXiv:1411.1045* (2014).

[KWB14] Matthias Kümmerer, Thomas Wallis, and Matthias Bethge. "*How close are we to understanding image-based saliency?*" In: *arXiv:1409.7686* (2014).

[Lab18] Livenda Labs. *Cinematic Advanced Temporal Anti-Aliasing (CTAA)*. 2018. URL: http://www.livenda.com/ctaa/. last visited 20. July 2018.

[Lag09] A. Lagae. *Wang Tiles in Computer Graphics*. Morgan & Claypool, 2009.

[LK11] S. Laine and T. Karras. "*Efficient Sparse Voxel Octrees*." In: *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 17 (2011), pp. 1048–1059. ISSN: 1077-2626.

[LHJ01] E. LaMar, B. Hamann, and K. I. Joy. "*A magnification lens for interactive volume visualization*." In: *Ninth Pacific Conference on Computer Graphics and Applications*. 2001, pp. 223–232.

[Lan+19] Matteo P. Lanaro, Hélène Perrier, David Coeurjolly, Victor Ostromoukhov, and Alessandro Rizzi. "*Blue-noise sampling for human retinal cone spatial distribution modeling*." June 2019. https://hal.archives-ouvertes.fr/hal-02155785, last visited 20. Nov. 2019.

[Lan+16] Eike Langbehn, Tino Raupp, Gerd Bruder, Frank Steinicke, Benjamin Bolte, and Markus Lappe. "*Visual Blur in Immersive Virtual Environments: Does Depth of Field or Motion Blur Affect Distance and Speed Estimation?*" In: *Proceedings of the 22Nd ACM Conference on Virtual Reality Software and Technology*. VRST '16. Munich, Germany: ACM, 2016, pp. 241–250. ISBN: 978-1-4503-4491-3.

[LaV+14] S. M. LaValle, A. Yershova, M. Katsev, and M. Antonov. *"Head tracking for the Oculus Rift."* In: *IEEE International Conference on Robotics and Automation*. ICRA '14. May 2014, pp. 187–194.

[LLV16] G. Lavoué, M. C. Larabi, and L. Váša. *"On the Efficiency of Image Metrics for Evaluating the Visual Quality of 3D Models."* In: *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 22.8 (Aug. 2016), pp. 1987–1999.

[Lav07] Guillaume Lavoué. *"A Roughness Measure for 3D Mesh Visual Masking."* In: *Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization*. APGV '07. ACM. 2007, pp. 57–60.

[LC16] Olivier Le Meur and Antoine Coutrot. *"Introducing context-dependent and spatially-variant viewing biases in saccadic models."* In: *Vision Research* 121 (2016), pp. 72–84.

[LSC04] Patrick Ledda, Luis Paulo Santos, and Alan Chalmers. *"A local model of eye adaptation for high dynamic range images."* In: *Proceedings of the 3rd International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*. Afrigraph '04. Stellenbosch, South Africa: ACM, 2004, pp. 151–160. ISBN: 1-58113-863-6.

[LTJ86] S. J. Lederman, G. Thorne, and B. Jones. *"Perception of texture by vision and touch: multidimensionality and intersensory integration."* eng. In: *Journal of Experimental Psychology. Human Perception and Performance* 12.2 (1986), pp. 169–180. ISSN: 0096-1523.

[LVJ05] Chang Ha Lee, Amitabh Varshney, and David W. Jacobs. *"Mesh Saliency."* In: *ACM Transactions on Graphics (TOG)*. Vol. 24. ACM, 2005, pp. 659–666.

[LF80] Gordon E. Legge and John M. Foley. *"Contrast masking in human vision."* In: *Journal of the Optical Society of America* 70.12 (1980), pp. 1458–1471.

[LJ19] Dmitry Zhdan Lei Yang and Matthew Johnson. *"NVIDIA Adaptive Shading Overview."* In: *Game Developers Conference: GDC*. 2019.

[LS97] Jed Lengyel and John Snyder. *"Rendering with Coherent Layers."* In: *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '97. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1997, pp. 233–242. ISBN: 0-89791-896-7.

[LE13] Peter Lenkic and James Enns. *"Apparent Motion Can Impair and Enhance Target Visibility: The Role of Shape in Predicting and Postdicting Object Continuity."* In: *Frontiers in Psychology* 4 (2013), p. 35.

[LKA85] D. M. Levi, S. A. Klein, and A. P. Aitsebaomo. *"Vernier acuity, crowding and cortical magnification."* In: *Vision Research* 25.7 (1985), pp. 963–977.

[LW90] M. Levoy and R. Whitaker. *"Gaze-directed Volume Rendering."* In: *Proceedings of the 1990 Symposium on Interactive 3D Graphics*. I3D '90. Snowbird, Utah, USA: ACM, 1990, pp. 217–223. ISBN: 0-89791-351-5.

[LMK01] Bei Li, Gary W. Meyer, and R. Victor Klassen. *"A Comparison of Two Image Quality Models."* In: *Proceedings of SPIE*. Ed. by Bernice E. Rogowitz and Thrasyvoulos N. Pappas. The International Society for Optical Engineering, 2001, pp. 98–109.

[Lim+13] Max Limper, Yvonne Jung, Johannes Behr, and Marc Alexa. *"The POP Buffer: Rapid Progressive Clustering by Geometry Quantization."* In: *Computer Graphics Forum* (2013).

[Lin16] Tim Lindeberg. *"Concealing Rendering Simplifications using Gaze Contingent Depth of Field."* MA thesis. Sweden: KTH Royal Institute of Technology School of Computer Science and Communication, 2016. https://kth.diva-portal.org/smash/get/diva2:947325/FULLTEXT01.pdf.

[LN07] Robert W. Lindeman and Haruo Noma. *"A Classification Scheme for Multi-sensory Augmented Reality."* In: *Proceedings of the 2007 ACM Symposium on Virtual Reality Software and Technology*. VRST '07. ACM, 2007, pp. 175–178. ISBN: 9781595938633.

[Lin00] Peter Lindstrom. *"Out-of-core Simplification of Large Polygonal Models."* In: *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '00. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 2000, pp. 259–262. ISBN: 1-58113-208-5.

[LT00] Peter Lindstrom and Greg Turk. *"Image-driven Simplification."* In: *ACM Transactions on Graphics (TOG)* 19.3 (July 2000), pp. 204–241. ISSN: 0730-0301.

[LB97] H. L. Liou and N. A. Brennan. *"Anatomically accurate, finite model eye for optical modeling."* In: *J Opt Soc Am A Opt Image Sci Vis* 14.8 (Aug. 1997), pp. 1684–1695.

[Liu+13] Yiming Liu, Jue Wang, Sunghyun Cho, Adam Finkelstein, and Szymon Rusinkiewicz. *"A No-reference Metric for Evaluating the Quality of Motion Deblurring."* In: *ACM Transactions on Graphics (TOG)* 32.6 (2013), 175:1–175:12.

[LF03] Christine A. Livingston and Stacey Rickert Fedder. *"Visual-Ocular Motor Activity in the Macaque Pregeniculate Complex."* In: *Journal of Neurophysiology* 90.1 (July 2003), pp. 226–244.

[Lo+10] Cheng-Hung Lo, Chih-Hsing Chu, Kurt Debattista, and Alan Chalmers. *"Selective rendering for efficient ray traced stereoscopic images."* In: *The Visual Computer* 26.2 (2010), pp. 97–107.

[LS10] Thurmon E. Lockhart and Wen Shi. *"Effects of age on dynamic accommodation."* In: *Ergonomics* 53.7 (July 2010), pp. 892–903.

[Loe99] Irene E. Loewenfeld. *"The reaction to near vision."* In: *The Pupil: Anatomy, physiology, and clinical applications*. Vol. 1. Oxford: Butterworth-Heinemann, 1999.

[LDC06] Peter Longhurst, Kurt Debattista, and Alan Chalmers. *"A GPU Based Saliency Map for High-fidelity Selective Rendering."* In: *Proceedings of the 4th International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*. AFRIGRAPH '06. Cape Town, South Africa: ACM, 2006, pp. 21–29. ISBN: 1-59593-288-7.

[LMS10] Francisco Lopez, Ramon Molla, and Veronica Sundstedt. *"Exploring Peripheral LOD Change Detections during Interactive Gaming Tasks."* In: *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*. APGV '10. ACM, 2010, pp. 73–80.

[LM00] Lester C. Loschky and George W. McConkie. *"User Performance With Gaze Contingent Multiresolutional Displays."* In: *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*. ETRA '00. ACM. 2000, pp. 97–103.

[LW07] Lester C. Loschky and Gary S. Wolverton. *"How Late Can You Update Gaze-Contingent Multiresolutional Displays Without Detection?"* In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 3.4 (2007), p. 7.

[Lot11] Timothy Lottes. *FXAA*. 2011. http://developer.download.nvidia.com/assets/gamedev/files/sdk/11/FXAA_WhitePaper.pdf, last visited 5. Sep. 2018.

[Lub95] Jeffrey Lubin. *"A visual discrimination model for imaging system design and evaluation."* In: *Vision models for target detection and recognition* 2 (1995), pp. 245–357.

[LBR08] Marcel Lucassen, P. Bijl, and Jolanda Roelofsen. *"The perception of static colored noise: Detection and masking described by CIE94."* In: *Color Research & Application* 33 (June 2008), pp. 178–191.

[LH01] David P. Luebke and Benjamin Hallen. *"Perceptually-Driven Simplification for Interactive Rendering."* In: *Proceedings of the 12th Eurographics Workshop on Rendering Techniques*. EGWR '01. Springer-Verlag, 2001, pp. 223–234. ISBN: 3-211-83709-4.

[Lue+03] David P. Luebke, Martin Reddy, Jonathan Cohen, Amitabh Varshney, Benjamin Watson, and Robert Huebner. *Level of Detail for 3D Graphics*. en. Morgan Kaufmann Publishers, 2003. ISBN: 978-1-55860-838-2.

[Lue+00] David Luebke, Benjamin Hallen, Dale Newfield, and Benjamin Watson. *Perceptually Driven Simplification Using Gaze-Directed Rendering*. Tech. rep. University of Virginia, 2000. CS-2000-04.

[Luk12] Rastislav Lukac. *Perceptual Digital Imaging: Methods and Applications*. CRC Press, 2012. ISBN: 1439868565.

[Luk99] H. D. Luke. "*The origins of the sampling theorem*." In: *IEEE Communications Magazine* 37.4 (Apr. 1999), pp. 106–108. ISSN: 0163-6804.

[LLK96] M. Ronnier Luo, Mei-Chun Lo, and Wen-Guey Kuo. "*The LLAB (l:c) colour model*." In: *Color Research & Application* 21.6 (1996), pp. 412–429.

[MRW96] Sabira K. Mannan, Keith H. Ruddock, and David S. Wooding. "*The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images*." In: *Spatial Vision* 10.3 (1996), pp. 165–188.

[MS74] J. Mannos and D. Sakrison. "*The Effects of a Visual Fidelity Criterion on the Encoding of Images*." In: *IEEE Transactions on Information Theory (TIT)* 20.4 (Sept. 1974), pp. 525–536. ISSN: 0018-9448.

[MBM13] Radoslaw Mantiuk, Bartosz Bazyluk, and Rafał K. Mantiuk. "*Gaze-driven Object Tracking for Real Time Rendering*." In: *Computer Graphics Forum*. Vol. 32. 2pt2. 2013, pp. 163–173.

[MBT11] Radoslaw Mantiuk, Bartosz Bazyluk, and Anna Tomaszewska. "*Gaze-Dependent Depth-of-field Effect Rendering in Virtual Environments*." In: *Proceedings of the SGDA'11*. Lisbon, Portugal: Springer-Verlag, 2011, pp. 1–12. ISBN: 978-3-642-23833-8.

[Man+05] Rafał Mantiuk, Scott J. Daly, Karol Myszkowski, and Hans-Peter Seidel. "*Predicting Visible Differences in High Dynamic Range Images: Model and its Calibration*." In: *Electronic Imaging 2005*. International Society for Optics and Photonics, 2005, pp. 204–214.

[Man+11] 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*." In: *ACM Transactions on Graphics (TOG)*. Vol. 30. 4. ACM. ACM, 2011, pp. 40–52.

[MMS15] Rafał Mantiuk, Karol Myszkowski, and Hans-Peter Seidel. *High Dynamic Range Imaging*. Wiley Encyclopedia of Electrical and Electronics Engineering, 2015. ISBN: 012374914X.

[Mar09] Jonathan Marbach. "*GPU Acceleration of Stereoscopic and Multi-view Rendering for Virtual Reality Applications*." In: *Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology*. VRST '09. ACM, 2009, pp. 103–110. ISBN: 978-1-60558-869-8.

[Mar+17] Iván Marín-Franch, Antonio Del Águila-Carrasco, Paula Bernal, J. J. Esteve-Taboada, Norberto López-Gil, R. Montés-Micó, and Philip Kruger. "*There is more to accommodation of the eye than simply minimizing retinal blur*." In: 8 (Oct. 2017), p. 4717.

[MMB97] William R. Mark, Leonard McMillan, and Gary Bishop. "*Post-rendering 3D Warping*." In: *Proceedings of the 1997 Symposium on Interactive 3D Graphics*. I3D '97. Providence, Rhode Island, USA: ACM, 1997, pp. 7–16. ISBN: 0-89791-884-3.

[MKC07] Ricardo Marroquim, Martin Kraus, and Paulo Roma Cavalcanti. "*Efficient Point-Based Rendering Using Image Reconstruction*." In: *The Eurographics Association on Point-Based Graphics* (2007).

[Mar+18] Adam Marrs, Josef Spjut, Holger Gruen, Rahul Sathe, and Morgan McGuire. "*Adaptive Temporal Antialiasing*." In: *Proceedings of the Conference on High-Performance Graphics*. HPG '18. Vancouver, British Columbia, Canada: ACM, 2018, 1:1–1:4. ISBN: 978-1-4503-5896-5.

[MMP05] M. Martelli, N. J. Majaj, and D. G. Pelli. *"Are faces processed like words? A diagnostic test for recognition by parts."* In: *Journal of Vision* 5.1 (2005), pp. 58–70.

[MFN16] 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*. CHI '16. ACM, 2016, pp. 5191–5202.

[Mau+12] Marilena Maule, Joao LD Comba, Rafael Torchelsen, and Rui Bastos. *"Transparency and Anti-Aliasing Techniques for Real-Time Rendering."* In: *25th Conference on Graphics, Patterns and Images , Tutorials*. SIBGRAPI-T. IEEE, 2012, pp. 50–59.

[MRD12] L. McIntosh, Bernhard E. Riecke, and Steve DiPaola. *"Efficiently Simulating the Bokeh of Polygonal Apertures in a Post-Process Depth of Field Shader."* In: *Computer Graphics Forum*. Vol. 31. 6. 2012, pp. 1810–1822.

[MN84] S. McKee and K. Nakayama. *"The detection of motion in the peripheral visual field."* In: *Vision Research* 24.1 (1984), pp. 25–32. ISSN: 0042-6989.

[McM97] Leonard McMillan Jr. *"An Image-Based Approach to Three-Dimensional Computer Graphics."* PhD thesis. University of North Carolina at Chapel Hill, 1997.

[Men+18] Xiaoxu Meng, Ruofei Du, Matthias Zwicker, and Amitabh Varshney. *"Kernel Foveated Rendering."* In: *ACM I3D* 1.1 (May 2018), pp. 124–133.

[MG10] Nicolas Menzel and Michael Guthe. *"Towards Perceptual Simplification of Models with Arbitrary Materials."* In: *Computer Graphics Forum*. Vol. 29. Pacific Graphics '10. 2010, pp. 2261–2270.

[Mer+17] Olivier Mercier, Yusufu Sulai, Kevin Mackenzie, Marina Zannoli, James Hillis, Derek Nowrouzezahrai, and Douglas Lanman. *"Fast Gaze-contingent Optimal Decompositions for Multifocal Displays."* In: *ACM TOG* 36.6 (Nov. 2017), 237:1–237:15. ISSN: 0730-0301.

[MM93] W. H. Merigan and J. H. Maunsell. *"How parallel are the primate visual pathways?"* In: *Annu. Rev. Neurosci.* 16 (1993), pp. 369–402.

[MPE16] Olivier Le Meur, Sumanta N. Pattanaik, and Jain Eakta. *"Visual Attention from a Graphics Point of View."* In: *Eurographics '16 Tutorial*. 2016. URL: http://jainlab.cise.ufl.edu/visual-attention-graphics-pov.html. last visited 02. Oct. 2017.

[ML92] Gary W. Meyer and Aihua Liu. *"Color spatial acuity control of a screen subdivision image synthesis algorithm."* In: *Human Vision, Visual Processing, and Digital Display III*. Ed. by Bernice E. Rogowitz. SPIE '92. The International Society for Optical Engineering, 1992, pp. 387–399.

[Mik+13] Michihiro Mikamo, Marcos Slomp, Bisser Raytchev, Toru Tamaki, and Kazufumi Kaneda. *"Technical Section: Perceptually Inspired Afterimage Synthesis."* In: *Computer & Graphics* 37.4 (2013), pp. 247–255.

[MR06] Erik Millan and Isaac Rudomin. *"Impostors and Pseudo-instancing for GPU Crowd Rendering."* In: *Proceedings of the 4th International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia*. GRAPHITE '06. Kuala Lumpur, Malaysia: ACM, 2006, pp. 49–55. ISBN: 1-59593-564-9.

[MUM83] Mortimer Mishkin, Leslie G. Ungerleider, and Kathleen A. Macko. *"Object vision and spatial vision: two cortical pathways."* In: *Trends in Neurosciences* 6 (1983), pp. 414–417. ISSN: 0166-2236.

[Mit87] Don P. Mitchell. *"Generating Antialiased Images at Low Sampling Densities."* In: *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*. Vol. 21. SIGGRAPH '87 4. ACM, 1987, pp. 65–72.

[Mit12] Martin Mittring. *"The Technology Behind the Unreal Engine 4 Elemental Demo."* In: *Siggraph 2012*. Epic Games Inc., 2012. URL: http://advances.realtimerendering.com/s2012/Epic/The%20Technology%20Behind%20the%20Elemental%20Demo%2016x9.pptx. last visited 4. May 2018.

[Miu86] T. Miura. *"Coping with situational demands: A study of eye movements and peripheral vision performance."* In: *Vision in Vehicles* (1986), pp. 206–216.

[Moo+15] Bochang Moon, Jose A. Iglesias-Guitian, Sung-Eui Yoon, and Kenny Mitchell. *"Adaptive Rendering with Linear Predictions."* In: *ACM Trans. Graph.* 34.4 (July 2015), 121:1–121:11. ISSN: 0730-0301.

[MB10] A. K. Moorthy and A. C. Bovik. *"A Two-Step Framework for Constructing Blind Image Quality Indices."* In: *IEEE Signal Processing Letters* 17.5 (2010), pp. 513–516.

[MC04] J. A. Mordi and K. J. Ciuffreda. *"Dynamic aspects of accommodation: age and presbyopia."* In: *Vision Research* 44.6 (Mar. 2004), pp. 591–601.

[MK88] J. A. Movshon and L. Kiorpes. *"Analysis of the development of spatial contrast sensitivity in monkey and human infants."* In: *J Opt Soc Am A* 5.12 (Dec. 1988), pp. 2166–2172.

[ML00] Jurriaan D. Mulder and Robert van Liere. *"Fast Perception-based Depth of Field Rendering."* In: *Proceedings of the ACM VRST' 00*. VRST '00. ACM, 2000, pp. 129–133. ISBN: 1-58113-316-2.

[Mul85] K. T. Mullen. *"The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings."* In: *The Journal of Physiology* 359 (Feb. 1985), pp. 381–400. ISSN: 0022-3751.

[MS99] K. T. Mullen and M. J. Sankeralli. *"Evidence for the stochastic independence of the blue-yellow, red-green and luminance detection mechanisms revealed by subthreshold summation."* eng. In: *Vision Research* 39.4 (1999), pp. 733–745. ISSN: 0042-6989.

[Mur78] B. J. Murphy. *"Pattern thresholds for moving and stationary gratings during smooth eye movement."* In: *Vision Research* 18.5 (1978), pp. 521–530.

[MD01] Hunter Murphy and Andrew T. Duchowski. *"Gaze-contingent level of detail rendering."* In: *EuroGraphics 2001* (2001).

[MD07] Hunter Murphy and Andrew T. Duchowski. *"Hybrid Image-/Model-based Gaze-contingent Rendering."* In: *4th Symposium on Applied Perception in Graphics and Visualization*. Tübingen, Germany: ACM, 2007, pp. 107–114. ISBN: 978-1-59593-670-7.

[Mys98] Karol Myszkowski. *"The visible differences predictor: applications to global illumination problems."* In: *Rendering Techniques '98*. Springer, 1998, pp. 223–236.

[Mys02] Karol Myszkowski. *"Perception-based Global Illumination, Rendering, and Animation Techniques."* In: *Proceedings of the 18th Spring Conference on Computer Graphics*. SCCG '02. Budmerice, Slovakia: ACM, 2002, pp. 13–24. ISBN: 1-58113-608-0.

[Nad+16a] Georges Nader, Kai Wang, Franck Hetroy-Wheeler, and Florent Dupont. *"Visual Contrast Sensitivity and Discrimination for 3D Meshes and their Applications."* In: *Computer Graphics Forum (CGF), Pacific Graphics '16* (2016).

[Nad+16b] Georges Nader, Kai Wang, Franck Hétroy-Wheeler, and Florent Dupont. *"Just Noticeable Distortion Profile for Flat-Shaded 3D Mesh Surfaces."* In: *IEEE Transactions on Visualization & Computer Graphics (TVCG)* 22.11 (2016), pp. 2423–2436. ISSN: 1077-2626.

[Nai98] Avi C. Naiman. *"Jagged Edges: When is Filtering Needed?"* In: *ACM Trans. Graph.* 17.4 (Oct. 1998), pp. 238–258. ISSN: 0730-0301.

[NY88] K. Nakatani and K. W. Yau. *"Calcium and light adaptation in retinal rods and cones."* In: *Nature* 334.6177 (July 1988), pp. 69–71.

[Nar+10] Takuji Narumi, Takashi Kajinami, Tomohiro Tanikawa, and Michitaka Hirose. *"Meta Cookie."* In: *ACM SIGGRAPH '10 Emerging Technologies*. SIGGRAPH '10. New York, NY, USA: ACM, 2010, 18:1–18:1. ISBN: 978-1-4503-0392-7.

[NDL15] Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet. *"HDR-VQM: An objective quality measure for high dynamic range video."* In: *Signal Processing: Image Communication* 35 (2015), pp. 46–60.

[Nar+14] Manish Narwaria, Matthieu Perreira Da Silva, Patrick Le Callet, and Romuald Pépion. *"On Improving the Pooling in HDR-VDP-2 towards Better HDR Perceptual Quality Assessment."* In: *Human Vision and Electronic Imaging 2014*. San Francisco, United States, Feb. 2014, pp. 1–6.

[NSI06] Diego Nehab, Pedro V. Sander, and John R. Isidoro. *"The Real-time Reprojection Cache."* In: *ACM SIGGRAPH 2006 Sketches*. SIGGRAPH '06. Boston, Massachusetts: ACM, 2006, p. 185. ISBN: 1-59593-364-6.

[Nik+04] Stavri G. Nikolov, Timothy D. Newman, Dave R. Bull, Nishan C. Canagarajah, Michael G. Jones, and Iain D. Gilchrist. *"Gaze-contingent display using texture mapping and OpenGL: System and Applications."* In: *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*. ETRA '04. ACM. 2004, pp. 11–18.

[Noo+83] Cornelis Noorlander, Jan J. Koenderink, Ron J. Den Olden, and B. Wigbold Edens. *"Sensitivity to spatiotemporal colour contrast in the peripheral visual field."* In: *Vision Research* 23.1 (1983), pp. 1–11.

[NE15] Antje Nuthmann and Wolfgang Einhäuser. *"A new approach to modeling the influence of image features on fixation selection in scenes."* In: *Annals of the New York Academy of Sciences* 1339.1 (2015), pp. 82–96.

[Nut+10] Antje Nuthmann, Tim J. Smith, Ralf Engbert, and John M. Henderson. *"CRISP: a computational model of fixation durations in scene viewing."* In: *Psychological Review* 117.2 (2010), pp. 382–405.

[OYT96] Toshikazu Ohshima, Hiroyuki Yamamoto, and Hideyulu Tamura. *"Gaze-directed adaptive rendering for interacting with virtual space."* In: *Proceedings of the IEEE Virtual Reality Annual International Symposium*. IEEE VR '96. IEEE. 1996, pp. 103–110.

[Oli+03] Aude Oliva, Antonio Torralba, Monica S. Castelhano, and John M. Henderson. *"Top-down control of visual attention in object detection."* In: *Proceedings of the International Conference on Image Processing*. Vol. 1. ICIP '03. IEEE. 2003, pp. 253–256.

[Opr+09] Cristina Oprea, Ionut Pirnog, Constantin Paleologu, and Mihnea Udrea. *"Perceptual video quality assessment based on salient region detection."* In: *Fifth Advanced International Conference on Telecommunications*. AICT'09. IEEE, 2009, pp. 232–236.

[Ove92] C. W. A. M. van Overveld. *"Application of morphological filters to tackle discretisation artefacts."* In: *The Visual Computer* 8.4 (July 1992), pp. 217–232.

[Pad+17] 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."* In: *Proceedings of the National Academy of Sciences* 114.9 (Feb. 2017), pp. 2183–2188.

[Pai05] Dinesh K. Pai. *"Multisensory Interaction: Real and Virtual."* In: *The Eleventh International Symposium on Robotics Research*. Springer, 2005, pp. 489–498.

[PS89] J. Painter and K. Sloan. *"Antialiased Ray Tracing by Adaptive Progressive Refinement."* In: *Proceedings of the 16th Annual Conference on Computer Graphics and Interactive Techniques*. Vol. 23. SIGGRAPH '89 3. ACM, 1989, pp. 281–288.

[Pal99] Stephen E. Palmer. *Vision Science: Photons to Phenomenology*. MIT Press, 1999. ISBN: 0262161834.

[PFD05] Hao Pan, Xiao-Fan Feng, and S. Daly. *"LCD motion blur modeling and analysis."* In: *IEEE International Conference on Image Processing 2005.* Vol. 2. Sept. 2005, pp. II-21-4.

[Pan11] J. Pantaleoni. *"VoxelPipe: A Programmable Pipeline for 3D Voxelization."* In: *Proceedings of the ACM SIGGRAPH Symposium on High Performance Graphics.* Vancouver, British Columbia, Canada: ACM, 2011, pp. 99–106. ISBN: 978-1-4503-0896-0.

[Pan14] Alexey Panteleev. *"Practical Real-Time Voxel-based Global Illumination For Current GPUs."* In: *GPU Technology Conference.* GTC' 14. NVIDIA. 2014. http://on-demand.gputechconf.com/gtc/2014/presentations/S4552-rt-voxel-based-global-illumination-gpus.pdf, last visited 6. July 2018.

[PSS14] Alexey Panteleev, Rahul Sathe, and Marco Salvi. *"Advances in Real-Time Voxel-Based GI."* In: *Game Developes Conference.* GDC' 18. NVIDIA. 2014. http://schedule.gdconf.com/session/advances-in-real-time-voxel-based-gi-temporal-super-resolution-presented-by-nvidia/856294, last visited 7. June 2018.

[PK13] C. Papadopoulos and A. E. Kaufman. *"Acuity-Driven Gigapixel Visualization."* In: *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 19.12 (2013), pp. 2886–2895. ISSN: 1077-2626.

[PLN01] Derrick Parkhurst, Irwin Law, and Ernst Niebur. *Evaluating Gaze-Contingent Level of Detail Rendering of Virtual Environments using Visual Search.* 2001. http://cnslab.mb.jhu.edu/publications/Parkhurst_etal01c.pdf, last visited 04 Nov. 2018.

[Pat+16a] Anjul Patney, Joohwan Kim, Marco Salvi, Anton Kaplanyan, Chris Wyman, Nir Benty, Aaron Lefohn, and David Luebke. *"Perceptually-based Foveated Virtual Reality."* In: *ACM SIGGRAPH 2016 Emerging Technologies.* ACM, 2016, 17:1–17:2. ISBN: 978-1-4503-4372-5.

[Pat+16b] 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."* In: *ACM TOG, SIGGRAPH Asia '16* 35.6 (Nov. 2016), 179:1–179:12. ISSN: 0730-0301.

[Pat+16c] 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 - Supplementary Material."* In: *ACM Transactions on Graphics (TOG), SIGGRAPH Asia '16* 35.6 (Nov. 2016).

[Ped16] Lasse Jon Fuglsang Pedersen. *"Temporal Reprojection Anti-Aliasing in INSIDE."* In: *GDC.* 2016. http://twvideo01.ubm-us.net/o1/vault/gdc2016/Presentations/Pedersen_LasseJonFuglsang_TemporalReprojectionAntiAliasing.pdf, last visited 13. Nov. 2019.

[PT08] D. G. Pelli and K. A. Tillman. *"The uncrowded window of object recognition."* In: *Nature Neuroscience* 11.10 (2008), pp. 1129–1135.

[Pen+11] Chao Peng, Seung In Park, Yong Cao, and Jie Tian. *"A Real-time System for Crowd Rendering: Parallel LOD and Texture-preserving Approach on GPU."* In: *Proceedings of the 4th International Conference on Motion in Games.* MIG'11. Edinburgh, UK: Springer-Verlag, 2011, pp. 27–38. ISBN: 978-3-642-25089-7.

[PKS17] Arsène Pérard-Gayot, Javor Kalojanov, and Philipp Slusallek. *"GPU Ray Tracing Using Irregular Grids."* In: *Comput. Graph. Forum* 36.2 (May 2017), pp. 477–486. ISSN: 0167-7055.

[Pér+17] Arsène Pérard-Gayot, Martin Weier, Richard Membarth, Philipp Slusallek, Roland Leißa, and Sebastian Hack. *"RaTrace: Simple and Efficient Abstractions for BVH Ray Traversal Algorithms."* In: *Proceedings of the 16th International Conference on Generative Programming: Concepts & Experiences (GPCE).* ACM. Vancouver, BC, Canada, Oct. 2017, pp. 157–168.

[PC85] V. Hugh Perry and Alan Cowey. *"The ganglion cell and cone distributions in the monkey's retina: Implications for central magnification factors."* In: *Vision Research* 25.12 (1985), pp. 1795–1810. ISSN: 0042-6989.

[Per07] Emil Persson. *"Selective Supersampling."* In: *ShaderX5 : Advanced Rendering Techniques*. Ed. by Wolfgang Engel. 5. Boston MA: Charles River Media, 2007.

[Per11] Emil Persson. *Geometric Post-process Anti-Aliasing*. Mar. 2011. URL: http://www.humus.name/index.php?page=3D&ID=86. last visited 16. April 2018.

[PM13] Josselin Petit and Rafał K Mantiuk. *"Assessment of video tone-mapping: Are cameras' S-shaped tone-curves good enough?"* In: *Journal of Visual Communication and Image Representation* 24.7 (2013), pp. 1020–1030.

[Pet+04] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama. *"Digital Photography with Flash and No-flash Image Pairs."* In: *ACM SIGGRAPH 2004 Papers*. SIGGRAPH '04. Los Angeles, California: ACM, 2004, pp. 664–672.

[Pet15] Matt Pettineo. *"Rendering The Alternate History of The Order: 1886."* In: *SIGGRAPH Advances in Real-Time Rendering in Games Course*. 2015. http://advances.realtimerendering.com/s2015/, last visited 05. Jan. 2018.

[Pfi+00] Hanspeter Pfister, Matthias Zwicker, Jeroen van Baar, and Markus Gross. *"Surfels: Surface Elements As Rendering Primitives."* In: *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '00. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 2000, pp. 335–342. ISBN: 1-58113-208-5.

[PH04] M. Pharr and G. Humphreys. *Physically Based Rendering: From Theory to Implementation, First Ed.* Morgan Kaufmann series in interactive 3D technology. Elsevier Science, 2004. ISBN: 0-12-553180-X.

[Pha18] Matt Pharr. *"Real-time rendering's next frontier: adopting lessons from offline ray tracing to practical real-time ray tracing pipelines."* In: ACM Siggraph. 2018. Course.

[PCR00] M. Piana, M. Canfora, and M. Riani. *"Role of noise in image processing by the human perceptive system."* In: *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 62.1 Pt B (July 2000), pp. 1104–1109.

[PKC15] Jakub Pietrzak, Krzysztof Kacperski, and Marek Cieślar. *"NVIDIA OptiX ray-tracing engine as a new tool for modelling medical imaging systems."* In: *SPIE Medical Imaging*. International Society for Optics and Photonics. 2015, 94122P.

[PZB16] D. Pohl, X. Zhang, and A. Bulling. *"Combining eye tracking with optimizations for lens astigmatism in modern wide-angle HMDs."* In: *2016 IEEE Virtual Reality (VR)*. Mar. 2016, pp. 269–270.

[Poh+15] Daniel Pohl, Timo Bolkart, Stefan Nickels, and Oliver Grau. *"Using astigmatism in wide angle HMDs to improve rendering."* In: *Annual International Symposium on Virtual Reality*. IEEE VR '15. 2015, pp. 263–264.

[PRC84] A. Pollatsek, K. Rayner, and W. E. Collins. *"Integrating pictorial information across eye movements."* In: *Journal of Experimental Psychology: General* 113.3 (1984), pp. 426–442.

[PRH90] A. Pollatsek, K. Rayner, and J. M. Henderson. *"Role of spatial location in integration of pictorial information across saccades."* In: *Journal of Experimental Psychology: Human Perception and Performance* 16.1 (1990), pp. 199–210.

[PSS16] Nicholas F. Polys, Ankit Singh, and Peter Sforza. *"A Novel Level-of-detail Technique for Virtual City Environments."* In: *Proceedings of the 21st International Conference on Web3D Technology*. Web3D '16. Anaheim, California: ACM, 2016, pp. 183–184. ISBN: 978-1-4503-4428-9.

[Pon+11] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, and M. Carli. *"Modified image visual quality metrics for contrast change and mean shift accounting."* In: *2011 11th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM)*. Feb. 2011, pp. 305–311.

[Por02] T. C. Porter. *"Contributions to the Study of Flicker. Paper II."* In: *Proceedings of the Royal Society of London* 70 (1902), pp. 313–329. ISSN: 0370-1662.

[PB71] Michael I. Posner and Stephen J. Boies. *"Components of attention."* In: *Psychological Review* 78.5 (1971), pp. 391–408.

[PP99] Jan Prikryl and Werner Purgathofer. *"Perceptually-driven termination for stochastic radiosity."* In: *Seventh International Conference in Central Europe on Computer Graphics and Visualization (Winter School on Computer Graphics)*. WSCG '99. 1999.

[PR98] S. J. Prince and B. J. Rogers. *"Sensitivity to disparity corrugations in peripheral vision."* eng. In: *Vision Research* 38.17 (1998), pp. 2533–2537. ISSN: 0042-6989.

[Qu+00] Huamin Qu, Ming Wan, Jiafa Qin, and Arie Kaufman. *"Image Based Rendering with Stable Frame Rates."* In: *Proceedings of the Conference on Visualization '00*. VIS '00. Salt Lake City, Utah, USA: IEEE Computer Society Press, 2000, pp. 251–258.

[QM08] Lijun Qu and Gary W. Meyer. *"Perceptually Guided Polygon Reduction."* In: *IEEE Transactions on Visualization and Computer Graphics*. Vol. 14. 5. 2008, pp. 1015–1029.

[RLH14] N. Radeva, L. Levy, and J. Hahn. *"Generalized Temporal Focus Context Framework for Improved Medical Data Exploration."* In: *Journal of Digital Imaging* 27.2 (2014), pp. 207–219.

[RR07] Ali Rahimi and Benjamin Recht. *"Random Features for Large-scale Kernel Machines."* In: *Proceedings of NIPS'07*. Vancouver, British Columbia, Canada: Curran Associates Inc., 2007, pp. 1177–1184.

[Ram+07] Ganesh Ramanarayanan, James Ferwerda, Bruce Walter, and Kavita Bala. *"Visual Equivalence: Towards a New Standard for Image Fidelity."* In: *ACM Transaction on Graphics (TOG), SIGGRAPH '07* 26.3 (2007), 76:1–76:12.

[RPG99] Mahesh Ramasubramanian, Sumanta N. Pattanaik, and Donald P. Greenberg. *"A perceptually based physical error metric for realistic image synthesis."* In: *Proceedings of the 26th annual conference on Computer graphics and Interactive Techniques*. SIGGRAPH '99. ACM, 1999, pp. 73–82.

[RW17] Waseem Rawat and Zenghui Wang. *"Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review."* In: *Neural Computation* 29.9 (2017), pp. 2352–2449.

[Red97] Martin Reddy. *"Perceptually Modulated Level of Detail for Virtual Environments."* PhD thesis. University of Edinburgh, 1997.

[Red01] Martin Reddy. *"Perceptually Optimized 3D Graphics."* In: *IEEE Computer Graphics and Applications* 21.5 (Sept. 2001), pp. 68–75.

[Ree15] Nathan Reed. *Gameworks VR*. Technical Slides. NVIDIA, 2015. https://developer.nvidia.com/sites/default/files/akamai/gameworks/vr/GameWorks_VR_2015_Final_handouts.pdf, last visited 30. Jan. 2018.

[Rei+12] Florian Reichl, Matthäus G. Chajdas, Kai Bürger, and Rüdiger Westermann. *"Hybrid Sample-based Surface Rendering."* In: *Proceedings of VMV*. 2012, pp. 47–54.

[Rei+10] Erik Reinhard, Greg Ward, Sumanta Pattanaik, Paul Debevec, Wolfgang Heidrich, and Karol Myszkowski. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. The Morgan Kaufmann Series in Computer Graphics. San Francisco: Morgan Kaufmann, 2010. ISBN: 978-0-12-585263-0.

[Rel11] AMD Developer Relations. *EQAA Modes for AMD 6900 Series Graphics Cards*. Tech. rep. Advanced Micro Devices, 2011. http://developer.amd.com/wordpress/media/2012/10/EQAAModesforAMDHD6900SeriesCards.pdf, last visited 4. April 2018.

[Res09] Alexander Reshetov. "*Morphological Antialiasing*." In: *Proceedings of the Conference on High Performance Graphics 2009*. HPG '09. New Orleans, Louisiana: ACM, 2009, pp. 109–116. ISBN: 978-1-60558-603-8.

[Rin+14] Ryan V. Ringer, Aaron P. Johnson, John G. Gaspar, Mark B. Neider, James Crowell, Arthur F. Kramer, and Lester C. Loschky. "*Creating a New Dynamic Measure of the Useful Field of View Using Gaze-Contingent Displays*." In: *Proceedings of the Symposium on Eye Tracking Research and Applications*. ETRA '14. ACM. 2014, pp. 59–66.

[Ris07] Eric Risser. *GPU Gems 3*. Ed. by Hubert Nguyen. First. Vol. 3. 21. Addison-Wesley Professional, 2007. Chap. True Impostors. ISBN: 9780321545428.

[RE12] Tobias Ritschel and Elmar Eisemann. "*A Computational Model of Afterimages*." In: *Comput. Graph. Forum* 31.2pt4 (May 2012), pp. 529–534. ISSN: 0167-7055.

[Rit02] Jörg Ritter. "*Wavelet based image compression using FPGAs*." PhD thesis. Mathematisch-Naturwissenschaftlich-Technische Fakultät der Martin-Luther-Universität Halle-Wittenberg, 2002.

[RS09] Austin Robison and Peter Shirley. "*Image Space Gathering*." In: *Proceedings of the Conference on High Performance Graphics 2009*. HPG '09. New Orleans, Louisiana: ACM, 2009, pp. 91–98.

[Rod98] R. W. Rodieck. *The first steps in seeing*. Sunderland MA: Sinauer Associates, 1998. ISBN: 0-87893-757-9.

[RW99] Austin Roorda and David R. Williams. "*The arrangement of the three cone classes in the living human eye*." In: *Nature* 397.6719 (Feb. 1999), pp. 520–522.

[RLN07] R. Rosenholtz, Y. Li, and L. Nakano. "*Measuring visual clutter*." In: *Journal of Vision* 7.2 (2007), pp. 11–22.

[Ros+01] John Ross, M. Concetta Morrone, Michael E Goldberg, and David C. Burr. "*Changes in visual perception at the time of saccades*." In: *Trends in Neurosciences* 24.2 (2001), pp. 113–121.

[Ros14] RossDaBoss. *Space Shooting Range*. 3D Warehouse. Mar. 2014. https://3dwarehouse.sketchup.com/model//space-shooting-range, last visited 13. Nov. 2019.

[Rot+17] Thorsten Roth, Martin Weier, André Hinkenjann, Yongmin Li, and Philipp Slusallek. "*A Quality-Centered Analysis of Eye Tracking Data in Foveated Rendering*." In: *Journal of Eye Movement Research (JEMR)* 10.5 (2017).

[Rot+15] Thorsten Roth, Martin Weier, Jens Maiero, André Hinkenjann, and Yongmin Li. "*Guided High-Quality Rendering*." In: *11th International Symposium on Visual Computing (ISVC)*. 2015.

[RR88] J. Rovamo and A. 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*." eng. In: *Vision Research* 28.7 (1988), pp. 785–790. ISSN: 0042-6989.

[RV79] J. Rovamo and V. Virsu. "*An estimation and application of the human cortical magnification factor*." eng. In: *Experimental Brain Research* 37.3 (1979), pp. 495–510. ISSN: 0014-4819.

[RVN78] Jyrki Rovamo, Veijo Virsu, and Risto Näsänen. "*Cortical magnification factor predicts the photopic contrast sensitivity of peripheral vision*." en. In: *Nature* 271.5640 (Jan. 1978), pp. 54–56. ISSN: 0028-0836.

[RL00] Szymon Rusinkiewicz and Marc Levoy. *"QSplat: A Multiresolution Point Rendering System for Large Meshes."* In: *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '00. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 2000, pp. 343–352. ISBN: 1-58113-208-5.

[SP04] Miguel Sainz and Renato Pajarola. *"Point-based rendering techniques."* In: *Computers & Graphics* 28.6 (2004), pp. 869–879. ISSN: 0097-8493.

[SPL04] Miguel Sainz, Renato Pajarola, and Roberto Lario. *"Points Reloaded: Point-based Rendering Revisited."* In: *Proceedings of the First Eurographics Conference on Point-Based Graphics*. SPBG'04. Switzerland: Eurographics Association, 2004, pp. 121–128. ISBN: 3-905673-09-6.

[San+01] Pedro V. Sander, John Snyder, Steven J. Gortler, and Hugues Hoppe. *"Texture Mapping Progressive Meshes."* In: *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '01. New York, NY, USA: ACM, 2001, pp. 409–416. ISBN: 1-58113-374-X.

[San+07] Fabrizio Santini, Gabriel Redner, Ramon Iovin, and Michele Rucci. *"EyeRIS: a general-purpose system for eye-movement-contingent display control."* In: *Behavior Research Methods* 39.3 (2007), pp. 350–364.

[SW14] Daniel R. Saunders and Russell L. Woods. *"Direct measurement of the system latency of gaze-contingent displays."* In: *Behavior Research Methods* 46.2 (2014), pp. 439–447.

[Sch56] Otto H. Schade. *"Optical and Photoelectric Analog of the Eye."* In: *Journal of the Optical Society of America* 46.9 (1956), pp. 721–739.

[SJW07] Daniel Scherzer, Stefan Jeschke, and Michael Wimmer. *"Pixel-correct Shadow Maps with Temporal Reprojection and Shadow Test Confidence."* In: *Proceedings of the 18th Eurographics Conference on Rendering Techniques*. EGSR'07. Grenoble, France: Eurographics Association, 2007, pp. 45–50. ISBN: 978-3-905673-52-4.

[SYM10] Daniel Scherzer, Lei Yang, and Oliver Mattausch. *"Exploiting temporal coherence in real-time rendering."* In: *ACM SIGGRAPH ASIA 2010 Courses*. ACM, 2010, p. 24.

[Sch+12] Daniel Scherzer, Lei Yang, Oliver Mattausch, Diego Nehab, Pedro V. Sander, Michael Wimmer, and Elmar Eisemann. *"Temporal Coherence Methods in Real-Time Rendering."* en. In: *Computer Graphics Forum* 31.8 (2012), pp. 2378–2408. ISSN: 01677055. Survey.

[Sch+17a] Christoph Schied, Anton Kaplanyan, Chris Wyman, Anjul Patney, Chakravarty R. Alla Chaitanya, John Burgess, Shiqiu Liu, Carsten Dachsbacher, Aaron Lefohn, and Marco Salvi. *"Spatiotemporal Variance-guided Filtering: Real-time Reconstruction for Path-traced Global Illumination."* In: *Proceedings of High Performance Graphics*. HPG '17. Los Angeles, California: ACM, 2017, 2:1–2:12. ISBN: 978-1-4503-5101-0.

[Sch01] Brian J. Scholl. *"Objects and attention: the state of the art."* In: *Cognition* 80.1-2 (2001), pp. 1–46. ISSN: 0010-0277. Objects and Attention.

[Sch+17b] A. Schollmeyer, S. Schneegans, S. Beck, A. Steed, and B. Froehlich. *"Efficient Hybrid Image Warping for High Frame-Rate Stereoscopic Rendering."* In: *IEEE Transactions on Visualization and Computer Graphics* 23.4 (Apr. 2017), pp. 1332–1341.

[SB85] C. M. Schor and D. R. Badcock. *"A comparison of stereo and vernier acuity within spatial channels as a function of distance from fixation."* In: *Vision Research* 25.8 (1985), pp. 1113–1119.

[SS10] M. Schwarz and H.-P. Seidel. *"Fast Parallel Surface and Solid Voxelization on GPUs."* In: *ACM Transaction on Graphics* 29.6 (2010), 179:1–179:10. ISSN: 0730-0301.

[Sch13] K. Schwenk. *"Filtering Techniques for Low-noise Previews of Interactive Stochastic Ray Tracing."* PhD thesis. TU Darmstadt, Germany, 2013. URL: http://tuprints.ulb.tu-darmstadt.de/3590/. last visited 13. Nov. 2019.

[Sch04] Jim Schwiegerling. *"Arizona Eye Model."* In: *Field Guide to Visual and Ophthalmic Optics*. SPIE, 2004.

[Sci10] Kara Rogers Senior Editor Biomedical Sciences. *The Eye - The Physiology of Human Perception*. The Rosen Publishing Group, Inc, 2010. ISBN: 978-1-615-30116-4.

[SMM00] Randy K. Scoggins, Robert J. Moorhead, and Raghu Machiraju. *"Enabling level-of-detail matching for exterior scene synthesis."* In: *Proceedings of the IEEE Conference on Visualization*. IEEE VIS '00. 2000, pp. 171–178.

[SSL97] Robert Sekuler, Allison B. Sekuler, and Renee Lau. *"Sound alters visual motion perception."* en. In: *Nature* 385.6614 (1997), pp. 308–308. ISSN: 0028-0836.

[Sha+98] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. *"Layered Depth Images."* In: *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '98. New York, NY, USA: ACM, 1998, pp. 231–242. ISBN: 0-89791-999-8.

[Sha+96] Jonathan Shade, Dani Lischinski, David H. Salesin, Tony DeRose, and John Snyder. *"Hierarchical Image Caching for Accelerated Walkthroughs of Complex Environments."* In: *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '96. New York, NY, USA: ACM, 1996, pp. 75–82. ISBN: 0-89791-746-4.

[Sha49] Claude Elwood Shannon. *"Communication in the Presence of Noise."* In: *Proceedings of the Institute of Radio Engineers* 37.1 (1949), pp. 10–21.

[Sha+90] Robert Shapley, T. Caelli, S. Grossberg, M. Morgan, and I. Rentschler. *"Computational theories of visual perception."* English (US). In: *The neural basis for visual perception*. Ed. by L. Spillman and J. Werner. Academic Press, 1990.

[SB05] H. R. Sheikh and A. C. Bovik. *"A visual information fidelity approach to video quality assessment."* In: *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*. 2005, pp. 23–25.

[SI10] John Shen and Laurent Itti. *"Gender differences in visual attention during listening as measured by neuromorphic saliency: What women (and Men) watch."* In: *Journal of Vision* 10.7 (2010), pp. 159–159.

[SLR10] Maxim Shevtsov, Mikhail Letavin, and Alexey Rukhlinskiy. *"Low Cost Adaptive Anti-Aliasing for Real-Time Ray-Tracing."* In: *Proceedings of the 20th International Conference on Computer Graphics and Vision*. GraphiCon '10. St. Petersburg, Russia, 2010, pp. 45–49.

[Shi+11] Takashi Shibata, Joohwan Kim, David M. Hoffman, and Martin S. Banks. *"Visual discomfort with stereo displays: effects of viewing distance and direction of vergence-accommodation conflict."* In: *Proceeding of SPIE*. Ed. by Andrew J. Woods, Nicolas S. Holliman, and Neil A. Dodgson. Feb. 2011, 78630P.

[SS01] Shinsuke Shimojo and Ladan Shams. *"Sensory modalities are not separate modalities: plasticity and interactions."* In: *Current Opinion in Neurobiology* 11.4 (2001), pp. 505–509.

[SK13a] M. Shohara and K. Kotani. *"The visual perception sensitivity for achromatic noise and chromatic noise."* In: *IEEE International Conference on Image Processing*. Sept. 2013, pp. 127–131.

[Sie+13] Christian Siegl, Quirin Meyer, Gerd Sußner, and Marc Stamminger. *"Solving aliasing from shading with selective shader supersampling."* In: *Computers & Graphics* 37.8 (2013), pp. 955–962. ISSN: 0097-8493.

[SS00] Maryann Simmons and Carlo H. Séquin. *"Tapestry: A Dynamic Mesh-based Display Representation for Interactive Rendering."* In: *Eurographics Workshop on Rendering Techniques*. 2000, pp. 329–340. ISBN: 3-211-83535-0.

[Sit+08a] Pitchaya Sitthi-amorn, Jason Lawrence, Lei Yang, Pedro V. Sander, and Diego Nehab. *"An Improved Shading Cache for Modern GPUs."* In: *Proceedings of the 23rd ACM SIGGRAPH/EUROGRAPHICS Symposium on Graphics Hardware*. Sarajevo, Bosnia and Herzegovina: Eurographics Association, 2008, pp. 95–101.

[Sit+08b] Pitchaya Sitthi-amorn, Jason Lawrence, Lei Yang, Pedro V. Sander, Diego Nehab, and Jiahe Xi. *"Automated Reprojection-based Pixel Shader Optimization."* In: *ACM SIGGRAPH Asia 2008 Papers*. SIGGRAPH Asia '08. Singapore: ACM, 2008, 127:1–127:11. ISBN: 978-1-4503-1831-0.

[SK13b] Edward E. Smith and Stephen M. Kosslyn. *Cognitive Psychology: Mind and Brain*. Englisch. Pearson New International. 2013. ISBN: 978-1-292-02235-2.

[STT12] Robert Snowden, Peter Thompson, and Tom Troscianko. *Basic Vision: An Introduction To Visual Perception*. Englisch. 2nd edition. Oxford: Oxford University Press, 2012. ISBN: 978-0-19-957202-1.

[SS03] Charles Spence and Sarah Squire. *"Multisensory Integration: Maintaining the Perception of Synchrony."* In: *Current Biology* 13.13 (July 2003), pp. 519–521. ISSN: 0960-9822.

[Sta+06] O. G. Staadt, B. A. Ahlborn, O. Kreylos, and B. Hamann. *"A Foveal Inset for Large Display Environments."* In: *Proceedings of the 2006 ACM International Conference on Virtual Reality Continuum and Its Applications*. VRCIA '06. Hong Kong, China: ACM, 2006, pp. 281–288. ISBN: 1-59593-324-7.

[Ste+15] Michael Stengel, Steve Grogorick, Martin Eisemann, Elmar Eisemann, and Marcus Magnor. *"An Affordable Solution for Binocular Eye Tracking and Calibration in Head-mounted Displays."* In: *Proceedings of the 23rd ACM international conference on Multimedia 2015*. MM' 15. 2015, pp. 15–24.

[Ste+16] Michael Stengel, Steve Grogorick, Martin Eisemann, and Marcus Magnor. *"Adaptive Image-Space Sampling for Gaze-Contingent Real-time Rendering."* In: *Proceedings of the Eurographics Symposium on Rendering* 35.4 (2016). Ed. by E. Eisemann and E. Fiume. EGSR '16.

[Sti18] Martin Stich. *Real-time Ray Tracing with NVIDIA RTX*. Electronic. Oct. 2018. http://on-demand.gputechconf.com/gtc-eu/2018/pdf/e8527-real-time-ray-tracing-with-nvidia-rtx.pdf, last visited 28. Oct. 2019.

[SFD09] Martin Stich, Heiko Friedrich, and Andreas Dietrich. *"Spatial Splits in Bounding Volume Hierarchies."* In: *High Performance Graphics 2009*. New Orleans, Louisiana: ACM, 2009, pp. 7–13. ISBN: 978-1-60558-603-8.

[Sto+04] William A. Stokes, James A. Ferwerda, Bruce Walter, and Donald P. Greenberg. *"Perceptual Illumination Components: A New Approach to Efficient, High Quality Global Illumination Rendering."* In: *ACM Transactions on Graphics (TOG)*. Vol. 23. ACM, 2004, pp. 742–749.

[Str18] Rita Strack. *"Deep learning advances super-resolution imaging."* In: *Nature Methods* 15.6 (2018), pp. 403–403.

[SRJ11] Hans Strasburger, Ingo Rentschler, and Martin Jüttner. *"Peripheral vision and pattern recognition: a review."* In: *Journal of Vision* 11.5 (2011). ISSN: 1534-7362.

[Str14] Filip Strugar. *Conservative Morphological Anti-Aliasing (CMAA)*. Ed. by Intel. 2014. URL: https://software.intel.com/en-us/articles/conservative-morphological-anti-aliasing-cmaa-update. last visited 15. Feb. 2017.

[Suf07] Kevin Suffern. *Ray Tracing from the Ground Up*. Natick, MA, USA: A K Peters, Ltd., 2007. ISBN: 1568812728.

[Sun+17] Qi Sun, Fu-Chung Huang, Joohwan Kim, Li-Yi Wei, David Luebke, and Arie Kaufman. *"Perceptually-guided Foveation for Light Field Displays."* In: *ACM Trans. Graph.* 36.6 (Nov. 2017), 192:1–192:13.

[Sun+05] Veronica Sundstedt, Kurt Debattista, Peter Longhurst, Alan Chalmers, and Tom Troscianko. *"Visual Attention for Efficient High-Fidelity Graphics."* In: *Proceedings of the 21st Spring Conference on Computer graphics*. SCCG '05. ACM. 2005, pp. 169–175.

[Sut02] Alistair Sutcliffe. *Multimedia and Virtual Reality: Designing Usable Multisensory User Interfaces*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 2002. ISBN: 978-0-8058-3950-0.

[SSS74] Ivan E. Sutherland, Robert F. Sproull, and Robert A. Schumacker. *"A Characterization of Ten Hidden-Surface Algorithms."* In: *ACM Comput. Surv.* 6.1 (Mar. 1974), pp. 1–55. ISSN: 0360-0300.

[SCM15] Nicholas T. Swafford, Darren Cosker, and Kenny Mitchell. *"Latency Aware Foveated Rendering in Unreal Engine 4."* In: *Proceedings of the 12th European Conference on Visual Media Production*. CVMP '15. ACM, 2015, 17:1–17:1.

[Swa+16] Nicholas T. Swafford, José A. Iglesias-Guitian, Charalampos Koniaris, Bochang Moon, Darren Cosker, and Kenny Mitchell. *"User, Metric, and Computational Evaluation of Foveated Rendering Methods."* In: *Proceedings of the ACM Symposium on Applied Perception*. SAP '16. ACM, 2016, pp. 7–14. ISBN: 978-1-4503-4383-1.

[TV05] Tatsuto Takeuchi and Karen K. De Valois. *"Sharpening image motion based on the spatio-temporal characteristics of human vision."* In: *Proc.SPIE*. Vol. 5666. 2005, p. 12.

[Tat+16] Natalya Tatarchuk, J. T. Hooker, Stephen Hill, Eric Heitz, Carlos Gonzalez-Ochoa, Louis Bavoil, Cyril Crassin, and et al. *Advances in Real-Time Rendering in 3D Graphics and Games - Course*. 2016. URL: http://advances.realtimerendering.com/s2016/index.html. last visited 20. Sep. 2018.

[Tau+98] Gabriel Taubin, André Guéziec, William Horn, and Francis Lazarus. *"Progressive Forest Split Compression."* In: *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '98. New York, NY, USA: ACM, 1998, pp. 123–132. ISBN: 0-89791-999-8.

[TM89] L. A. Temme and A. Morris. *"Speed of accommodation and age."* In: *Optom Vis Sci* 66.2 (Feb. 1989), pp. 106–112.

[Tem+11] Krzysztof Templin, Piotr Didyk, Tobias Ritschel, Elmar Eisemann, Karol Myszkowski, and Hans-Peter Seidel. *"Apparent Resolution Enhancement for Animations."* In: *Proc. of the 27th Spring Conference on Computer Graphics*. Vinicne, Slovakia, 2011, pp. 85–92.

[TL05] Rui Wang Tenghui Zhu and David Luebke. *"A GPU-Accelerated Render Cache."* In: *Pacific Graphics 2005 (short paper)*. Macao, China, Oct. 2005.

[TG02] Jan Theeuwes and Richard Godijn. *"Irrelevant singletons capture attention: Evidence from inhibition of return."* In: *Perception & Psychophysics* 64.5 (2002), pp. 764–770.

[TW01] L. C. Thomas and C. D. Wickens. *"Visual Displays and Cognitive Tunneling: Frames of Reference Effects on Spatial Judgments and Change Detection."* In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 45.4 (Oct. 2001), pp. 336–340.

[TW06] Lisa C. Thomas and Christopher D. Wickens. *"Effects of battlefield display frames of reference on navigation tasks, spatial judgements, and change detection."* In: *Ergonomics* 49.12-13 (2006), pp. 1154–1173.

[Tho+01] S. J. Thorpe, K. R. Gegenfurtner, M. Fabre-Thorpe, and H. H. Bülthoff. *"Detection of animals in natural images using far peripheral vision."* In: *European Journal of Neuroscience* 14.5 (2001), pp. 869–876.

[To+11] Michelle P. S. To, Ian D. Gilchrist, Tom Troscianko, and David J. Tolhurst. *"Discrimination of natural scenes in central and peripheral vision."* In: *Vision Research* 51.14 (2011), pp. 1686–1698. ISSN: 0042-6989.

[Toc16] Yvonne Toczek. *"The Influence of Visual Realism on the Sense of Presence in Virtual Environments."* MA thesis. Eindhoven University of Technology, Department of Industrial Engineering & Innovation Sciences, 2016.

[Tol+02] Parag Tole, Fabio Pellacini, Bruce Walter, and Donald P. Greenberg. *"Interactive Global Illumination in Dynamic Scenes."* In: *ACM Transactions on Graphics (TOG)* 21.3 (2002), pp. 537–546. ISSN: 0730-0301.

[Tol+05] David J. Tolhurst, Caterina Ripamonti, C. Alejandro Párraga, P. George Lovell, and Tom Troscianko. *"A Multiresolution Color Model for Visual Difference Prediction."* In: *Proceedings of the 2nd Symposium on Applied Perception in Graphics and Visualization*. APGV '05. ACM, 2005, pp. 135–138. ISBN: 978-1-59593-139-9.

[Tra17] Tobii Eye Tracking. *Tobii Eye Tracker 4C - This is how your head and eye gives gaming*. electronic. Mar. 2017. https://help.tobii.com/hc/en-us/article_attachments/115007616885/Introduction_to_the_Tobii_Eye_Tracker_4C.pdf, last visited 28. Oct. 2019.

[Tre88] Anne Treisman. *"Features and objects: The fourteenth Bartlett memorial lecture."* In: *The Quarterly Journal of Experimental Psychology* 40.2 (1988), pp. 201–237.

[TG80] Anne M. Treisman and Garry Gelade. *"A feature-integration theory of attention."* In: *Cognitive Psychology* 12.1 (1980), pp. 97–136.

[TDM16] M. Trevino, B. De la Torre-Valdovinos, and E. Manjarrez. *"Noise Improves Visual Motion Discrimination via a Stochastic Resonance-Like Phenomenon."* In: *Front Hum Neurosci* 10 (2016), p. 572.

[TE14] Hans A. Trukenbrod and Ralf Engbert. *"ICAT: A computational model for the adaptive control of fixation durations."* In: *Psychonomic bulletin & review* 21.4 (2014), pp. 907–934.

[Tur+19] Okan Tarhan Tursun, Elena Arabadzhiyska-Koleva, Marek Wernikowski, Radoslaw Mantiuk, Hans-Peter Seidel, Karol Myszkowski, and Piotr Didyk. *"Luminance-contrast-aware Foveated Rendering."* In: *ACM Trans. Graph.* 38.4 (July 2019), 98:1–98:14.

[Tyl85] C. W. Tyler. *"Analysis of visual modulation sensitivity. II. Peripheral retina and the role of photoreceptor dimensions."* In: *J Opt Soc Am A* 2.3 (Mar. 1985), pp. 393–398.

[Vai+14] Karthik Vaidyanathan, Marco Salvi, Robert Toth, Foley, et al. *"Coarse Pixel Shading."* In: *ACM HPG '14*. 2014, pp. 9–18.

[Val+14] Giuseppe Valenzise, Francesca De Simone, Paul Lauga, and Frederic Dufaux. *"Performance evaluation of objective quality metrics for HDR image compression."* In: *SPIE Optical Engineering+ Applications, 92170C*. International Society for Optics and Photonics. 2014.

[Val14] Michal Valient. *"Taking Killzone Shadow Fall Image Quality into the Next Generation."* In: *GDC 2014*. Guerrilla-Games, 2014. URL: https://www.guerrilla-games.com/read/taking-killzone-shadow-fall-image-quality-into-the-next-generation-1. last visited 21. Nov. 2018.

[Van+11] Peter Vangorp, Gaurav Chaurasia, Pierre-Yves Laffont, Roland W. Fleming, and George Drettakis. *"Perception of Visual Artifacts in Image-based Rendering of FaçAdes."* In: *Proceedings of the Twenty-second Eurographics Conference on Rendering*. EGSR '11. Prague, Czech Republic: Eurographics Association, 2011, pp. 1241–1250.

[VCD09] Suzane Vassallo, Sian L. Cooper, and Jacinta M. Douglas. *"Visual scanning in the recognition of facial affect: Is there an observer sex difference?"* In: *Journal of Vision* 9.3 (2009), pp. 1–11.

[Vat+11] Dmitriy Vatolin, Sergey Grishin, Kumok Boris, and Sheludko Victor. *VirtualDub MSU Noise Estimation Filter*. Electronic. Apr. 2011. MSU Graphics & Media Lab (Video Group), http://www.compression.ru/video/noise_estimation/index_en.html, last visited 20. Nov. 2018.

[Vel+06] Edgar Velázquez-Armendáriz, Eugene Lee, Kavita Bala, and Bruce Walter. *"Implementing the Render Cache and the Edge-and-point Image on Graphics Hardware."* In: *Proceedings of Graphics Interface 2006*. GI '06. Quebec, Canada: Canadian Information Processing Society, 2006, pp. 211–217. ISBN: 1-56881-308-2.

[VLA15] Yannick Verdie, Florent Lafarge, and Pierre Alliez. *"LOD Generation for Urban Scenes."* In: *ACM Trans. Graph.* 34.3 (May 2015), 30:1–30:14. ISSN: 0730-0301.

[VDC14] Eleonora Vig, Michael Dorr, and David Cox. *"Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images."* In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. CVPR '14. 2014, pp. 2798–2805.

[Vig+12] Eleonora Vig, Michael Dorr, Thomas Martinetz, and Erhardt Barth. *"Intrinsic Dimensionality Predicts the Saliency of Natural Dynamic Scenes."* In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 34.6 (2012), pp. 1080–1091.

[VAF16] M. Vinnikov, R. S. Allison, and S. Fernandes. *"Impact of depth of field simulation on visual fatigue: Who are impacted? and how?"* In: *International Journal of Human-Computer Studies* 91 (2016), pp. 37–51.

[VJ04] P. Viola and M. J. Jones. *"Robust Real-Time Face Detection."* In: *International Journal of Computer Vision* (2004), pp. 137–154.

[Vla16] Alex Vlachos. *"Advanced VR Rendering Performance."* In: *Game Developer Conference 2016 - Slides*. GDC '16. 2016.

[VMA17] Netflix's VMAF. *vmaf: Perceptual video quality assessment based on multi-method fusion*. Electronic. Apr. 2017. https://github.com/Netflix/vmaf, last visited 26. Sep. 2018.

[Vol+78] Frances C. Volkmann, Lorrin A. Riggs, Keith D. White, and Robert K. Moore. *"Contrast sensitivity during saccadic eye movements."* In: *Vision Research* 18.9 (1978), pp. 1193–1199.

[W3C16] World Wide Web Consortium (W3C). *SVG 1.1 (Second Edition) - Standard*. 2016. URL: https://www.w3.org/TR/SVG11/filters.html#feGaussianBlur. last visited 9. Nov. 2019.

[Wal98] Bruce Walter. *"Density Estimation Techniques for Global Illumination."* PhD thesis. Cornell University, Aug. 1998.

[WDP99] Bruce Walter, George Drettakis, and Steven Parker. *"Interactive Rendering using the Render Cache."* In: *Rendering Techniques (Proceedings of the Eurographics Workshop on Rendering)*. Ed. by D. Lischinski and G. W. Larson. Vol. 10. Springer-Verlag, June 1999, pp. 235–246.

[WPG02] Bruce Walter, Sumanta N. Pattanaik, and Donald P. Greenberg. *"Using Perceptual Texture Masking for Efficient Image Synthesis."* In: *Computer Graphics Forum* 21.3 (2002).

[WNR00] Jörg Walter, Claudia Nölker, and Helge Ritter. *"The PSOM Algorithm and Applications."* In: *Proceedings of the International Symposium on Neural Computation*. 2000, pp. 758–764.

[Wan95] Brian A. Wandell. *Foundations of Vision*. Stanford University, 1995. ISBN: 0878938532.

[Wan+12] Rui I. Wang, Brandon Pelfrey, Andrew T. Duchowski, and Donald H. House. *"Online Gaze Disparity via Bioncular Eye Tracking on Stereoscopic Displays."* In: *Proceedings of 3DIMPVT '12*. IEEE, 2012, pp. 184–191. ISBN: 978-0-7695-4873-9.

[Wan+14] Rui I. Wang, Brandon Pelfrey, Andrew T. Duchowski, and Donald H. House. *"Online 3D Gaze Localization on Stereoscopic Displays."* In: *ACM Trans. Appl. Percept.* 11.1 (Apr. 2014), 3:1–3:21. ISSN: 1544-3558.

[WFG02] Yigang Wang, Bernd Fröhlich, and Martin Göbel. *"Fast Normal Map Generation for Simplified Meshes."* In: *J. Graph. Tools* 7.4 (Dec. 2002), pp. 69–82. ISSN: 1086-7651.

[Wan06] Yubing Wang. *Survey of Objective Video Quality Measurements*. Tech. rep. EMC Corporation Hopkinton, MA 01748, USA: Worcester Polytechnic Institute, Jan. 2006.

[Wan+15] Yuxiang Wang, Chris Wyman, Yong He, and Pradeep Sen. *"Decoupled Coverage Antialiasing."* In: *Proceedings of the 7th Conference on High-Performance Graphics*. HPG '15. New York, NY, USA: ACM, 2015, pp. 33–42. ISBN: 978-1-4503-3707-6.

[Wan+04] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. *"Image quality assessment: from error visibility to structural similarity."* In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[WSB03] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. *"Multiscale structural similarity for image quality assessment."* In: *Conference Record of the 37th Asilomar Conference on Signals, Systems and Computers*. Vol. 2. IEEE, 2003, pp. 1398–1402.

[WRC88] Gregory J. Ward, Francis M. Rubinstein, and Robert D. Clear. *"A Ray Tracing Solution for Diffuse Interreflection."* In: *SIGGRAPH Comput. Graph.* 22.4 (June 1988), pp. 85–92. ISSN: 0097-8930.

[WS99] Gregory Ward and Maryann Simmons. *"The Holodeck Ray Cache: An Interactive Rendering System for Global Illumination in Nondiffuse Environments."* In: *ACM Transaction on Graphics (TOG)* 18.4 (1999), pp. 361–368. ISSN: 0730-0301.

[Wat93] Andrew B. Watson. *"DCT quantization matrices visually optimized for individual images."* In: *Human Vision, Visual Processing, and Digital Display IV*. Vol. 1913. SPIE '93. 1993, pp. 202–216.

[Wei05] Qingqing Wei. *"Converting 2D to 3D: A survey."* In: *Information and Communication Theory Group (ICT)*. Ed. by E. A. Hendriks and P. A. Redert. Faculty of Engineering, Mathematics and Computer science, Delft University of Technology, the Netherlands, 2005, p. 43.

[Wei+13] Martin Weier, André Hinkenjann, Georg Demme, and Philipp Slusallek. *"Generating and Rendering Large Scale Tiled Plant Populations."* In: *JVRB - Journal of Virtual Reality and Broadcasting* 10.1 (2013).

[WHS15] Martin Weier, André Hinkenjann, and Philipp Slusallek. *"A Unified Triangle/Voxel Structure for GPUs and its Applications."* In: *Journal of WSCG*. WSCG 24.No. 1-2 (2015), pp. 83–90. ISSN: 2464-4617.

[Wei+14a] Martin Weier, Jens Maiero, Thorsten Roth, André Hinkenjann, and Philipp Slusallek. *"Enhancing Rendering Performance with View-Direction-Based Rendering Techniques for Large, High Resolution Multi-Display Systems."* In: *11. Workshop Virtuelle Realität und Augmented Reality der GI-Fachgruppe VR/AR*. Sept. 2014.

[Wei+14b] Martin Weier, Jens Maiero, Thorsten Roth, André Hinkenjann, and Philipp Slusallek. *Lazy Details for Large High-Resolution Displays*. SIGGRAPH Asia. 2014. Poster.

[Wei+18a] Martin Weier, Thorsten Roth, André Hinkenjann, and Philipp Slusallek. *"Foveated Depth-of-Field Filtering in Head-Mounted Displays."* In: *ACM Transactions on Applied Perception (TAP)*. Vancouver, Canada, Aug. 2018. Best Paper Award, invited article.

[Wei+18b] Martin Weier, Thorsten Roth, André Hinkenjann, and Philipp Slusallek. *"Predicting the Gaze Depth in Head-mounted Displays using Multiple Feature Regression."* In: *Proceedings of the ACM Symposium on Eye Tracking Research and Applications (ETRA)*. Warsaw, Poland, June 2018.

[Wei+16] Martin Weier, Thorsten Roth, Ernst Kruijff, André Hinkenjann, Arsène Pérard-Gayot, Philipp Slusallek, and Yongmin Li. *"Foveated Real-Time Ray Tracing for Head-Mounted Displays."* In: *Computer Graphics Forum (Proceedings of Pacific Graphics '16)*. Oct. 2016.

[Wei+17] Martin Weier et al. *"Perception-driven Accelerated Rendering."* In: *Computer Graphics Forum (Proceedings of Eurographics)* 36.2 (Apr. 2017).

[WP04] Janet M. Weisenberger and Gayla L. Poling. *"Multisensory roughness perception of virtual surfaces: effects of correlated cues."* In: *Proceedings 12th International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*. HAPTICS '04. IEEE, 2004, pp. 161–168.

[WJB] Du Wenjun, Feng Jieqing, and Yang Baoguang. *"Sub-Pixel Anti-Aliasing Via Triangle-Based Geometry Reconstruction."* In: *Computer Graphics Forum* 33.7 (), pp. 81–90.

[WM78] G. Westheimer and S. P. McKee. *"Steroscopic acuity for moving retinal images."* eng. In: *Journal of the Optical Society of America* 68.4 (Apr. 1978), pp. 450–455. ISSN: 0030-3941.

[Wes12] Gerald Westheimer. *"Optical superresolution and visual hyperacuity."* In: *Progress in Retinal and Eye Research* 31.5 (2012), pp. 467–480. ISSN: 1350-9462.

[Wey58] Frank W. Weymouth. *"Visual sensory units and the minimal angle of resolution."* In: *American Journal of Ophthalmology* 46.1 (1958), pp. 102–113.

[WA09] C. D. Wickens and A. L. Alexander. *"Attentional tunneling and task management in synthetic vision displays."* In: *The International Journal of Aviation Psychology* 19.2 (2009), pp. 182–199.

[WC97] Christopher D. Wickens and C. Melody Carswell. *"Information Processing."* In: *Handbook of Human Factors and Ergonomics*. John Wiley & Sons, Inc., 1997, pp. 130–149.

[WDW99] A. Mark Williams, Keith Davids, and John Garrett Pascoe Williams. *Visual Perception & Action in Sport*. Taylor & Francis, 1999. ISBN: 041918290X.

[Wil+94] D. R. Williams, D. H. Brainard, M. J. McMahon, and R. Navarro. *"Double-pass and interferometric measures of the optical quality of the eye."* In: *J Opt Soc Am A Opt Image Sci Vis* 11.12 (Dec. 1994), pp. 3123–3135.

[WC83] David R. Williams and Robert Collier. *"Consequences of Spatial Sampling by a Human Photoreceptor Mosaic."* In: *Science* 221.4608 (1983), pp. 385–387. ISSN: 00368075, 10959203.

[Wil+03] Nathaniel Williams, David Luebke, Jonathan D. Cohen, Michael Kelley, and Brenden Schubert. *"Perceptually Guided Simplification of Lit, Textured Meshes."* In: *Proceedings of the 2003 Symposium on Interactive 3D Graphics*. I3D '03. Monterey, California: ACM, 2003, pp. 113–121. ISBN: 1-58113-645-5.

[WM04] Jered E. Windsheimer and Gary W. Meyer. *"Implementation of a visual difference metric using commodity graphics hardware."* In: *Electronic Imaging 2004*. International Society for Optics and Photonics, 2004, pp. 150–161.

[Win12] Stefan Winkler. *"Characteristics of Human Vision."* In: *Perceptual Digital Imaging: Methods and Application*. Ed. by Rastislav Lukac. CRC Press, 2012.

[WR80] Nyron L. Wolbarsht and James Ringo. *Visual acuity and the balance between receptor density and ganglion cell receptive field overlap*. Tech. rep. N00019-79-C-0-370. Duke University Eye Center, 1980, p. 47.

[Wol94] Jeremy M. Wolfe. *"Guided Search 2.0 A revised model of visual search."* In: *Psychonomic Bulletin & Review* 1.2 (1994), pp. 202–238.

[WB97] Jeremy M. Wolfe and Sara C. Bennett. *"Preattentive Object Files: Shapeless Bundles of Basic Features."* In: *Vision Research* 37.1 (1997), pp. 25–43.

[Wol+19] Krzysztof Wolski, Daniele Giunchi, Shinichi Kinuwaki, Piotr Didyk, Karol Myszkowski, Rafal K. Mantiuk, and Steed Anthony. *"Selecting Texture Resolution Using a Task-specific Visibility Metric."* In: *Computer Graphics Forum (Proc. Pacific Graphics)* 38.7 (2019).

[Wol+18] Krzysztof Wolski, Daniele Giunchi, Nanyang Ye, Piotr Didyk, Karol Myszkowski, Radoslaw Mantiuk, Hans-Peter Seidel, Anthony Steed, and Rafał K. Mantiuk. *"Dataset and Metrics for Predicting Local Visible Differences."* In: *ACM Transactions on Graphics* (2018). ISSN: 0730-0301.

[WDB05] Kwoon Y. Wong, Felice A. Dunn, and David M. Berson. *"Photoreceptor Adaptation in Intrinsically Photosensitive Retinal Ganglion Cells."* In: *Neuron* 48.6 (2005), pp. 1001–1010. ISSN: 0896-6273.

[Wu+13] Jinliang Wu, Xiaoyong Shen, Wei Zhu, and Ligang Liu. *"Mesh saliency with global rarity."* In: *Graphical Models* 75.5 (2013), pp. 255–264.

[Xav+01] Decoret Xavier, Sillion Franĉois, Schaufler Gernot, and Dorsey Julie. *"Multi-layered impostors for accelerated rendering."* In: *Computer Graphics Forum* 18.3 (2001), pp. 61–73.

[Yan+15] Dong-Ming Yan, Jian-Wei Guo, Bin Wang, Xiao-Peng Zhang, and Peter Wonka. *"A Survey of Blue-Noise Sampling and Its Applications."* In: *Journal of Computer Science and Technology* 30.3 (May 2015), pp. 439–452. DOI: 10.1007/s11390-015-1535-0. URL: https://doi.org/10.1007%2Fs11390-015-1535-0.

[Yan+16] Bailin Yang, Frederick W. B. Li, Xun Wang, Mingliang Xu, Xiaohui Liang, Zhaoyi Jiang, and Yanhui Jiang. *"Visual saliency guided textured model simplification."* In: *The Visual Computer* 32.11 (2016), pp. 1415–1432.

[Yan+09] Lei Yang, Diego F. Nehab, Pedro V. Sander, Pitchaya Sitthi-amorn, Jason Lawrence, and Hugues Hoppe. *"Amortized supersampling."* In: *ACM Transactions on Graphics (TOG)* 28.5 (2009), 135:1–135:12.

[Yan+11] Lei Yang, Yu-Chiu Tse, Pedro V. Sander, Jason Lawrence, Diego Nehab, Hugues Hoppe, and Clara L. Wilkins. *"Image-based Bidirectional Scene Reprojection."* In: *ACM Trans. Graph.* 30.6 (Dec. 2011), 150:1–150:10. ISSN: 0730-0301.

[Yan+18] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, and Jing-Hao Xue. *Deep Learning for Single Image Super-Resolution: A Brief Review.* 2018. eprint: arXiv:1808.03344.

[Yan+12] Xuan S. Yang, Linling Zhang, Tien-Tsin Wong, and Pheng-Ann Heng. *"Binocular Tone Mapping."* In: *ACM Transactions on Graphics (TOG), SIGGRAPH '12* 31.4 (2012), 93:1–93:10.

[Yan95] S. Yantis. *"Perceived Continuity of Occluded Visual Objects."* In: *Psychological Science* 6.3 (1995), pp. 182–186.

[YPG01] Hector Yee, Sumanita Pattanaik, and Donald P. Greenberg. *"Spatiotemporal Sensitivity and Visual Attention for Efficient Rendering of Dynamic Environments."* In: *ACM Transactions on Graphics (TOG)* 20.1 (2001), pp. 39–65. ISSN: 0730-0301.

[Yel83] John I. Yellott. *"Spectral consequences of photoreceptor sampling in the rhesus retina."* In: *Science* 221.4608 (1983), pp. 382–385.

[You06] Peter Young. *CSAA (Coverage Sampling Antialiasing).* Tech. rep. NVIDIA, Developer Technology Group, 2006. URL: http://www.nvidia.com/object/coverage-sampled-aa.html.

[Yu+09] Insu Yu, Andrew Cox, Min H. Kim, Tobias Ritschel, Thorsten Grosch, Carsten Dachsbacher, and Jan Kautz. *"Perceptual Influence of Approximate Visibility in Indirect Illumination."* In: *ACM Transactions on Applied Perception (TAP)* 6.4 (Oct. 2009), 24:1–24:14. ISSN: 1544-3558.

[Yu+17] Jinhui Yu, Kailin Wu, Kang Zhang, and Xianjun Sam Zheng. *A Computational Model of Afterimages based on Simultaneous and Successive Contrasts*. 2017. arXiv:1709.04550.

[ZK15] Sergey Zagoruyko and Nikos Komodakis. *"Learning to compare image patches via convolutional neural networks."* In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 4353–4361.

[ZC01] D. Zavagno and G. Caputo. *"The glare effect and the perception of luminosity."* In: *Perception* 30.2 (2001), pp. 209–222.

[ZDL02] Wenjun Zeng, Scott Daly, and Shawmin Lei. *"An overview of the visual optimization tools in JPEG 2000."* In: *Signal Processing: Image Communication*. JPEG 2000 17.1 (Jan. 2002), pp. 85–104. ISSN: 0923-5965.

[Zha+07] Long Zhang, Wei Chen, David S. Ebert, and Qunsheng Peng. *"Conservative Voxelization."* In: *Vis. Comput.* 23.9 (Aug. 2007), pp. 783–792.

[ZBJ06] Xiaopeng Zhang, Frédéric Blaise, and Marc Jaeger. *"Multiresolution Plant Models with Complex Organs."* In: *Proceedings of the 2006 ACM International Conference on Virtual Reality Continuum and Its Applications*. VRCIA '06. Hong Kong, China: ACM, 2006, pp. 331–334. ISBN: 1-59593-324-7.

[ZK13] Qi Zhao and Christof Koch. *"Learning saliency-based visual attention: A review."* In: *Signal Processing: Special issue on Machine Learning in Intelligent Image Processing* 93.6 (2013), pp. 1401–1407.

[Zou+17] G. Zoulinakis, J. J. Esteve-Taboada, T. Ferrer-Blasco, D. Madrid-Costa, and R. Montes-Mico. *"Accommodation in human eye models: a comparison between the optical designs of Navarro, Arizona and Liou-Brennan."* In: *Int J Ophthalmol* 10.1 (2017), pp. 43–50.

[ZWF16] Lingxuan Zuo, Hanli Wang, and Jie Fu. *"Screen content image quality assessment via convolutional neural network."* In: *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE. 2016, pp. 2082–2086.

[Zwi+15] Matthias Zwicker, Wojciech Jarosz, Jaakko Lehtinen, Bochang Moon, Ravi Ramamoorthi, Fabrice Rousselle, Pradeep Sen, Cyril Soler, and Sung-Eui Yoon. *"Recent Advances in Adaptive Sampling and Reconstruction for Monte Carlo Rendering."* In: *Computer Graphics Forum (Proceedings of Eurographics)* 34.2 (May 2015), pp. 667–681.

GLOSSARY AND ABBREVIATIONS

2AFC Two Alternatives Forced Choice. 118,

AA Anti-Aliasing. 69, 109, 139,

Accommodation Mechanical ability of the eye to compress and relax the lens, enabling the eye to maintain focus on an object, so that a sharp image appears on the *retina*.

Adaptation Automatically triggered and a time-dependent process of tuning sensitivity of the photosensitive *retina* to the amount of incoming light, also includes the *pupillary light reflex*.

ANOVA Analysis Of Variance. 144, 183, 192,

AQM Animation Quality Metric. 75,

AR Augmented Reality. 4, 77, 192,

ATAA Adaptive Temporal Anti-aliasing. 73, 136,

AXAA Adaptive Approximate Anti-Aliasing. 90,

CDF Cumulative Distribution Function. 147,

Central Vision Part of the visual field that is projected onto the *fovea*, *parafovea* and *perifovea*, i.e., up to an eccentricity of up to 17° .

CFF Critical Flicker Frequency. 59,

CIE Commission Internationale de l'Eclairage. 43,

CLOD Continuous Level-of-Detail. 62,

CMAA Conservative Morphological Anti-Aliasing. 90,

CMF Cortical Magnification Factor. 39, 67,

CoC Circle-of-Confusion. 167,

Cones Cone-shaped *photoreceptors* on the *retina* responsible for(*Photopic Vision*). They are tightly packed in the *fovea centralis* with their density decreasing quickly towards the periphery. Cones can be subdivided into **L**ong, **M**edium and **S**hort-Cones according to the band of the visual spectrum they are sensitive to.

Contrast Sensitivity Sensitivity to the difference in the light intensities of two adjacent areas [Gol13, p. 411].

Contrast Sensitivity Function (CSF) A function defined over spatial frequency of a *sinusoidal grating pattern* yielding a subject's *contrast sensitivity*.

Cortical Magnification Factor (CMF) The linear extent of the visual striate cortex to which each degree of the *retina* projects. It is directly proportional to visual acuity [CR74].

cpd Cycle per Degree. 24, 36, 202,

CPU Central Processing Unit.

Critical Flicker Frequency (CFF) The frame rate at which a sequentially presented series of images appears as continuous, or is perceptually fused. Measured in Hertz (Hz) .

Cross-modal Interaction Effects between various perceptual channels, e.g. visual stimuli might be missed when an auditory distractor is active .

CRT Cathode Ray Tube.

CSF Contrast Sensitivity Function. 26, 40, 64, 111, 194,

CVA Comfortable Viewing Angle. 29, 149, 174,

Cycle per Degree (cpd) A unit to describe the spatial frequency, defined as one period of a *sinusoidal grating pattern* at the projected size of 1 degree of the visual field.

DAEAA Directionally Adaptive Edge Anti-Aliasing.

DAG Directed Acyclic Graph. 126,

DEAA Distance-to-edge Anti-Aliasing.

Depth Cues Strategies such as eye convergence (binocular depth cue), motion parallax (monocular depth cue) and perspective for estimating the distance of an object.

Depth-of-Field (DoF) Is a property of an optical system and describes the nearest and the farthest distance in which objects are sharply imaged and considered to be in focus. Objects outside those range are imaged blurred. .

DGI Detail-guide image. 115,

DLSS Deep Learning Super Sampling. 91,

DoF Depth-of-Field. 5, 30, 67, 155, 157, 192,

Eccentricity Angular deviation from the center of the *fovea*.

Field of View (FoV) A measure describing the extent of the world observable by an optical system at one specific point in time, given in degrees. Using both eyes and looking straight ahead humans have an almost 180° horizontal field of view. If the eyeball rotation is included (and with the temporal restriction being relaxed) the horizontal field of view extends to 270°.

Fixation Gazing at a point of the scene or display for a certain time (fixation duration).

FoV Field of View. 3, 23, 80, 98, 129, 169, 201,

Fovea (Centralis) The area of the retina that is able to perceive and resolve visual information at the highest possible detail from approx. 5.2° around the central optical axis.

FPS Frames-per-Second. 3, 46, 110, 139,

FRC Foveal Region Configuration. 134, 161,

FSAA Full-Scene Anti-Aliasing. 68,

FSV Field-of-Sharp-Vision. 115,

FXAA Fast Approximate Anti-Aliasing. 89,

Gaze-contingency Paradigm A generic term for devices and methods that adapt their function depending on the user's gaze. Usually, the user's gaze is determined with an eye tracker .

GBAA Geometry-Buffer Anti-Aliasing.

GI Global Illumination. 68, 122, 139, 172, 193,

GLSL OpenGL Shading Language. 98,

GPAA Geometric Post-process Anti-Aliasing.

GPU Graphics Processing Unit. 49, 63, 98, 172, 191,

HDR High Dynamic Range. 28, 44,

High-level Perception A field concerned with how known objects are recognized. The "top-down" processing of the human visual system.

HMD Head-Mounted Display. 4, 76, 129, 159, 191, 201,

HSVO Hybrid Sparse Voxel Octree. 98, 191, 198,

Human Visual System (HVS) A model that describes the entire system that enables humans to perceive and process visual input including the eyes, visual pathways, visual cortex, and deeper neural processing.

HVS Human Visual System. 3, 9, 33, 59, 114, 150, 157, 191,

HVVR Hierarchical Visibility for Virtual Reality. 81,

Hyperacuity Perception of features that exceed the *visual acuity* .

Inattentional Blindness Effect A psychological lack of attention in which an individual fails to recognize an unexpected stimulus that is in plain sight.

Interpupillary Distance (IPD) The distance between the optical centers of the pupils.

IOD Interocular Distance. 24, 165,

IPD Interpupillary Distance.

JND Just Noticeable Difference. 45, 65,

Just Noticeable Difference (JND) A psycho-physical measure of how much a stimulus has to be changed

in order for a difference to be perceivable in at least 50% of the cases.

kNN k Nearest-Neighbors. 81,

LCD Liquid Crystal Display.
LDR Low Dynamic Range. 44,
LGN Lateral Geniculate Nucleus. 17, 48,
LGOV Leave-Group-Out-Validation. 176,
LHRDW Large High-Resolution Display Wall.
LMS Color Space Represents colors, separated by their distribution into **L**ong, **M**edium and **S**hort wavelengths, corresponding to the *cone* types in the human eye.
LoD Level-of-Detail. 5, 60, 97, 129, 191,
LOSO Leave-One-Scene-Out. 181,
Low-level Perception The "bottom-level" processing in the early stages of the human visual system. Models allow *saliency* estimation.
Luminance A photometric measure of the intensity per unit area of light emitted in a specific direction.

M-Scaling Hypothesis States that visual performance degradation with increasing eccentricity can be canceled out by spatial scaling of stimuli, by the inverse of the *CMF* .
MAR Minimum Angle of Resolution. 24, 36, 114, 167, 201,
MCPD Mean Co-Located Pixel Difference. 48,
Mesopic Vision A combination of photopic and scotopic vision occurring at dim light levels where both rods and cones are active.
Minimum Angular Resolution (MAR) Property to describe the resolution of an optical system. Resolution is expressed as the minimum angle allowing for the distinction of two points. For the eye and with normal vision this corresponds to about 1° when mapped to the *fovea* and decreases with increasing *eccentricity*.
MLAA Morphological Anti-Aliasing. 89,
MOS Mean Opinion Score. 50,
Motion Sickness Over time conflicting visual and motion cues can result in motion sickness.
MSAA Multisampling Anti-Aliasing. 70, 127,
MSE Mean Squared Error. 47, 176,
MSSSIM Multi-Scale SSIM. 48, 111,
MTF Modulation Transfer Function. 35,

NAS NVIDIA Adaptive Shading. 76,
NPR Non-Photorealistic Rendering.

OBB Oriented Bounding Box. 124,
Object of Interest (OoI) An object or part of a scene the user is looking at. It can be either measured by using active *eye tracking* or approximated by *saliency* analysis.
OoI Object of Interest. 17,

Parafovea The area of the retina from approx. 5.2° to 9° around the central optical axis.
PDF Probability Density Function. 38, 82,
Perifovea The area of the retina from approx. 9° to 17° around the central optical axis.
Peripheral Vision Visual stimuli that are not within *central vision*.
Photopic Vision Color vision using the cone receptors under normal lighting conditions (daylight). *Rods* are permanently saturated and therefore deactivated under these conditions.
Photoreceptor Retinal cells (rods and cones) that convert light received at the *retina* into nerve signals. *Rods* are achromatic and sensitive to motion, while *cones* provide color sensitivity.
Point-of-Regard (PoI) The point the user is looking at in image space, obtained by the eye tracker.
PoR Point-of-Regard. 56, 61, 129, 158, 194, 200,
PP-Interpolation Pull-Push Interpolation. 161,

PPI pixels-per-inch. 25,

PSNR Peak Signal-to-Noise Ratio. 47, 111,

PSOM Periodic Self-Organizing Map. 179,

Pupillary Light Reflex The process of adjusting the pupil's diameter to the amount of incoming light as a part of *adaptation*.

Receptive Field A particular part of the sensory space in which a stimulus triggers a neuron. The receptive field of a *photoreceptor* can be described as a cone-shaped volume representing the directions in which light can trigger a response. For the *retina* it is the entire *visual field*.

Retina Photosensitive layer of the eye containing *photoreceptors*.

Retinal Ganglion Cells The output neurons containing circular *receptive fields* in order to encode and transmit information from the eye to the brain.

RGSS Rotated Grid Supersampling.

Rods Rod-shaped achromatic *photoreceptors* in the *retina* that are especially important in dim lighting conditions (*scotopic vision*).

Saccade A small rapid movement of the eye that occurs during the scanning of a scene and fixation changes.

Saccadic Suppression The effect that the visual system seems to shut down to some degree during *saccades*. That is, even though the point of fixation moves at very high velocities during a *saccade*, blurred vision is not experienced.

Saliency The perceptual importance of parts in a scene and their likelihood to capture attention .

Scan Path A description for captured gaze behavior usually including spatial fixation locations and fixation durations .

Scene Schema Hypothesis States that objects that are unexpected/unusual in a specific context have a high *saliency* [HH99].

Scotopic Vision Monochromatic vision under low light-level conditions making use of the *rod receptors* exclusively.

Simultaneous Contrasts The effect that two colors when viewed side-by-side interact with each other and can lead to a different visual sensation.

Simultaneous Masking see *Visual Masking* .

Singleton Hypothesis States that the viewer's attention is drawn by stimuli that are locally unique and globally rare [TG02].

Sinusoidal Grating Pattern An alternating pattern of bright and dark areas at a specific or increasing frequency of a sine function. Used to measure a subject's contrast sensitivity.

SMAA Subpixel Morphological Anti-Aliasing. 89, 136,

Smooth Pursuit Eye Motion (SPEM) Smooth movement of the eyes when following a moving object, stands contrary to saccadic movements. Smooth pursuit and saccadic movements may occur in conjunction when an object is moving fast, so catch-up saccades may be required .

SPEM Smooth Pursuit Eye Motion. 28, 88, 150,

spp sample-per-pixel. 68, 74, 110, 139, 191, 198,

SR Stochastic Resonance. 69,

SRAA Subpixel Reconstruction Anti-Aliasing.

SSAA Screen-space Anti-Aliasing.

SSIM Structural Similarity. 47,

Stereo Vision Describes the human ability to combine two visual streams (*Stereopsis*) to improve visual performance, e.g., depth perception.

Stereopsis Process that fuses the visual input from both eyes to allow for *stereo vision*.

Supra-threshold A term to describe a stimulus large enough to produce a response. This can be an action potential in a sensory cell or even just a perceivable difference of a stimulus (see *Just Noticeable Difference*).

SVM Support Vector Machine. 52, 164,

SVO Sparse Voxel Octree. 63, 98,

SVR Support Vector Regression. 166,

TAA Temporal Anti-Aliasing. 78, 127, 137, 157,

TBAA Triangle-based Anti-Aliasing.

TC Temporal Coherence. 59, 73, 132, 159, 192,

TMLAA Topological Reconstruction Anti-Aliasing. 89,

TMO Tone Mapping Operator. 28, 42,

Tone Mapping Operator (TMO) A computational method to compute *Tone Mapping*. This includes methods for compressing the dynamic range of a high-dynamic-range image in order to display it on a low-dynamic-range device such as a typical computer screen.

VDB View-Direction Based.

VDM Sarnoff Visual Discrimination Metric. 45, 49, 64,

VDP Daly's Visible Differences Predictor. 45, 48, 74,

VEP Visual Equivalence Predictor. 50, 66,

Vergence Describes the process that is required to simultaneously rotate both eyes into opposite directions to fixate an object.

Vergence-accommodation conflict Describes the discomforting situation when stereo images are generated that convey depth information, which needs a conflicting *vergence* and *accommodation* to the one given by the actual screen's focal distance [Shi+11] .

Vestibular System The mechanism in the ear to monitor the body's acceleration, equilibrium and its relationship with the earth's gravitational field.

vestibular-ocular reflex Keeps the orientation of the eyes aligned with the current *OOI*, based on acceleration information from the vestibular system, amount of head rotation and retinal velocity.

VIF Visual Information Fidelity. 48,

Visual Acuity The ability to resolve small detail under ideal illumination conditions, i.e., the ability to detect and distinguish two points close to each other.

Visual Cortex The main part of the brain concerned with the sense of sight and the processing of visual information .

Visual Cues Signals or prompts derived from visual input. Such cues are preattentive by providing information from the environment subconsciously. Moreover, they might bring knowledge from previous experiences to mind.

Visual Difference Predictor (VDP) Daly's Visible Differences Predictor [Dal93] introduces a psycho-physical computational model of the HVS to compare two input images and derive a measure of perceivable differences. VDP processes images in the frequency domain. In contrast to *VDM*, it is particularly sensitive to differences near the visibility threshold.

Visual Discrimination Metric (VDM) The Sarnoff Visual Discrimination Metric [Lub95] introduces a psycho-physical computational model of the HVS to compare two input images. VDM derives a single *JND* value and a difference map. VDM processes images by convolution and down-sampling. In contrast to *VDP*, it is designed to generate a response above the *supra-threshold* at the expense of precision loss, when

near threshold differences need to be judged.

Visual Equivalence Predictor (VEP) The VEP metric by Ramanarayanan et al. [Ram+07] introduces a psychophysical computational model with the goal of measuring the visual equivalency of input images. Visual equivalency means the same impression of scene appearance is conveyed even though there can be measurable perceptual differences.

Visual Field see *Field of View*.

Visual Masking The reduction or elimination of a stimulus (target) by the presentation of a second stimulus (mask). The detection threshold of the target can be affected by the interfering masking stimulus when closely coupled in space and time.

VMAF Video Multi-Method Assessment Fusion. 48, 111,

VNSR Visual Signal-to-Noise Ratio. 50,

VR Virtual Reality. 4, 31, 77, 118, 129, 192,

VRS Variable Rate Shading. 76,