

Perceptual modeling for stereoscopic 3D

A thesis for obtaining the title of
Doctor of Engineering (Dr.-Ing.)
of the Faculties of Natural Science and Technology
of Saarland University

by

PETR KELLNHOFER

Saarbrücken, Germany
July 2016



Supervisors

Prof. Dr.-Ing. Karol Myszkowski
Prof. Dr. Hans-Peter Seidel

Dean

Prof. Dr. Frank-Olaf Schreyer

Colloquium

Date

4 November, AD 2016

Chair

Prof. Dr.-Ing. Philipp Slusallek

Reviewers

Prof. Dr. Hans-Peter Seidel
Prof. Dr.-Ing. Karol Myszkowski
Prof. Dr. Belen Masia
Prof. Dr. Wojciech Matusik

Academic assistant

Dr. Shida Beigpour

Declaration on Oath I hereby certify under penalty of perjury that I have done this work independently and without using any resources other than the ones specified. Such data and concepts that were acquired indirectly from other sources are marked and their respective source is indicated. This work has never been submitted in Germany or any other country in the same or similar form in order to obtain an academic degree.

Eidesstattliche Versicherung Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, 4 November 2016

Abstract

Virtual and Augmented Reality applications typically rely on both stereoscopic presentation and involve intensive object and observer motion. A combination of high dynamic range and stereoscopic capabilities become popular for consumer displays, and is a desirable functionality of head mounted displays to come. The thesis is focused on complex interactions between all these visual cues on digital displays.

The first part investigates challenges of the stereoscopic 3D and motion combination. We consider an interaction between the continuous motion presented as discrete frames. Then, we discuss a disparity processing for accurate reproduction of objects moving in the depth direction. Finally, we investigate the depth perception as a function of motion parallax and eye fixation changes by means of saccadic motion.

The second part focuses on the role of high dynamic range imaging for stereoscopic displays. We go beyond the current display capabilities by considering the full perceivable luminance range and we simulate the real world experience in such adaptation conditions. In particular, we address the problems of disparity retargeting across such wide luminance ranges and reflective/refractive surface rendering.

The core of our research methodology is perceptual modeling supported by our own experimental studies to overcome limitations of current display technologies and improve the viewer experience by enhancing perceived depth, reducing visual artifacts or improving viewing comfort.

Kurzfassung

Anwendungen von virtueller und erweiterter Realität verwenden in der Regel eine stereoskopische Darstellung und schließen intensive Objekt- und Betrachterbewegung ein. Die Integration von hohen Dynamikumfangen stellt eine weitere erstrebenswerte Funktionalität dar. Diese Arbeit beschäftigt sich mit den komplexen Wechselwirkungen zwischen all diesen visuellen Wahrnehmungselementen. Wir beschreiben die Einschränkungen aktueller Bildschirmtechnologien und überwinden sie, indem wir Wahrnehmungsmodelle zusammen mit unseren eigenen Studien verwenden, um den Betrachterkomfort zu steigern, die wahrgenommene Tiefe zu verstärken und visuelle Artefakte zu reduzieren.

Der erste Teil untersucht die Herausforderungen, die entstehen, wenn stereoskopisches 3D mit Bewegung kombiniert wird. Wir betrachten Wechselwirkungen zwischen kontinuierlicher Bewegung, die in Form diskreter Einzelbilder dargestellt wird. Weiterhin untersuchen wir die Tiefenwahrnehmung sowohl von Objekten, die sich in die Tiefenrichtung bewegen, als auch bei Bewegungsparallaxe und Fixationsveränderungen des Auges mittels sakkadischer Bewegung.

Der zweite Teil beschäftigt sich mit der Rolle von Bildern mit hohem Dynamikumfang. Indem wir den kompletten wahrnehmbaren Luminanzumfang betrachten, überschreiten wir die Möglichkeiten aktueller Bildschirme und analysieren die Anpassung von Disparitäten und das Rendern von reflektierenden Oberflächen in solchen unterschiedlichen Bedingungen.

Summary

Stereoscopic 3D has already established itself as a mainstream feature in digital imaging, gaming, and film production. This in turn has also triggered significant research efforts to improve the overall depth perception experience. A better understanding of 3D display limitations and the human perception has opened ways for smarter content processing. Traditionally, stereoscopic 3D as presented on current devices can suffer from some problems. The discomfort caused by rivalry or excessive disparity in junction with so called vergence-accomodation conflict can easily diminish the advantage of binocular perception, cause fatigue and consequently increase preference for legacy 2D viewing. Perceptual modeling of the human visual system (HVS) is a way how to both optimize the content for best viewing experience by general audience but also how to account for particular properties of a device or personalize for individual observers. In this thesis we employ perceptual frameworks to tackle new challenges in modern multimedia systems. Head mounted displays (MHD) for virtual reality (VR) and Augmented reality (AR) are an example of such a device. They utilize a stereoscopic display together with user tracking for direct motion interaction. High dynamic range (HDR) displays are another exciting innovation arriving to the customer market and they can potentially be implemented in HMD as well. This thesis investigates interactions of stereoscopic 3D with both the motion and HDR and it is organized into two corresponding parts. We show how properties of HVS change under presence of motion or wide range of adaptation luminance, and how we can utilize such behavior to improve perceptual characteristics of stereoscopic displays. This way we achieve better depth reproduction, reduce visual artifacts and improve viewing comfort.

The research methodology is consistent across the whole thesis for both the motion and HDR stereoscopic topics. Our techniques are deeply rooted in perceptual research that we often extend by our own measurements to fit needs imposed by our applications. We generalize the observations into a perceptual model that is in turn adapted into a computational model suitable for our input data. Such model can be either directly applied as a metric or more often it is integrated into an optimization framework that seeks to improve a desired image quality such as depth reproduction or viewing comfort. Finally, the loop gets closed by a validation study which verifies that properties originally observed for simple stimuli hold also when the model is applied to complex images.

Part I: Stereoscopic 3D and motion

In the first part of this thesis, we focus on motion in stereoscopic applications. Such motion originates both from the content itself, as movies or interactive applications contain motion naturally, but also from the viewer self-motion as emulated by HMDs with the head and gaze tracking capability that support immersion into virtual worlds and novel ways of human-computer interaction. Beyond a typical focus on achieving

temporal coherence, we study temporal limits of the HVS and the way motion contributes to our understanding of the world. In particular we investigate the following aspects of motion:

Optimizing disparity for screen plane motion

Due to a discretization in both space and time the stereoscopic motion reproduced on display devices undergoes a number of constraints with respect to the inherently continuous real world. We describe, how content-adaptive capture protocols can reduce false motion in depth sensation for time sequential displays. Another motion distortion originates from a combination of a display design and specific limitation of the HVS itself. We study such behavior for an anaglyph display and propose a temporal compensation for the resulting Pulfrich effect.

Optimizing disparity for motion in depth

The perception of dynamic stereo content relies on reproducing the full disparity-time volume that a scene point undergoes in motion. This volume can be strongly distorted during disparity manipulation, which is only concerned with changing disparity at each frame, even if the temporal coherence of that change is maintained. We experimentally measure how sensitive a human observer is to different types of disparity distortion. Then we propose an optimization to preserve stereo motion of content that was subject to an arbitrary disparity manipulation, based on a perceptual model of temporal disparity changes. Furthermore, we introduce a novel 3D warping technique to create stereo image pairs that conform to this optimized disparity map. User studies show that our method improves both the viewing comfort and performance in depth estimation tasks.

Gaze-driven disparity manipulations

In many scenarios, the range of depth that can be reproduced by the disparity cue from stereoscopic viewing is greatly limited and typically fixed due to device constraints. In this chapter, we show that this problem can be significantly alleviated when the eye fixation regions can be roughly estimated. We propose a new method for stereoscopic depth adjustment that utilizes eye tracking or other gaze prediction information. Unlike previous work we apply gradual unnoticeable depth adjustments during eye fixation. We measure the speed limits of disparity changes in various depth adjustment scenarios, and formulate a new model that can guide such seamless stereoscopic content processing. Based on this model, we propose a latency-immune real-time controller that applies local manipulations to stereoscopic content to find the optimum between depth reproduction and visual comfort. We also demonstrate benefits of our model in off-line applications, such as pre-processing for stereoscopic movie production. A validation study shows significant improvements in depth perception without sacrificing the visual quality when our techniques are applied.

Motion parallax as a depth cue

Current displays, most notably automultiscopic screens, can only reproduce limited disparity depth. In this chapter, we explore motion parallax which is a relatively strong depth cue, but it is purely 2D, and therefore, its reproduction is not limited. In many practical scenarios, the depth from motion parallax can compensate for an aggressive disparity compression. We conduct psychovisual experiments that measure the influence

of motion parallax on depth perception and relate it to the depth resulting from binocular disparity. Our joint disparity-parallax computational model predicts apparent depth resulting from both cues. We then present new disparity manipulation techniques, which first quantify depth obtained from motion parallax, and then, adjust binocular disparity information accordingly. A user study demonstrates that allocating the depth budget according to the strength of motion parallax improves the overall depth reproduction.

Part II: Stereoscopic 3D and HDR

In the second part of this thesis, we for the first time investigate a stereoscopic content presentation on high dynamic range (HDR) displays. We study the effect of both very bright and very dark luminance levels on disparity perception. Through careful simulation we cover the entire luminance range that can be perceived by the human vision which is still beyond the capabilities of commercially available displays.

Disparity perception in photopic vision

Bright day-like images are often associated with shiny reflective or refractive surfaces. Regardless how appealing such view dependent effects are, they create a challenge for their representation in stereoscopic 3D. The transfer of light through optical interfaces can easily introduce uncomfortable extensive disparity or deform resulting images. That prevents their fusion and causes an unpleasant rivalry. Although such effects occur in real world, we argue that they are not desirable on current displays, as the absence of correct accommodation and inability to avoid uncomfortable viewpoint by moving one's head make the experience significantly differ. We propose an optimization scheme that modifies camera parameters for each pixel in order to maintain visually pleasing and realistic disparities avoid annoying rivalry. We validate our approach in a user study where it achieves a better viewing comfort and at the same time higher realism than competitors.

Disparity perception in scotopic vision

The appearance of a scotopic low-light night scene on a photopic display ("day-for-night") can be simulated by color desaturation, acuity loss, and the Purkinje shift towards blue colors. We argue that faithful stereo reproduction of night scenes on photopic stereo displays requires manipulation of not only color but also binocular disparity. To this end, we performed a psychophysical experiment to devise a model of disparity at scotopic luminance levels. Using this model, we can match binocular disparity of a scotopic stereo content displayed on a photopic monitor to the disparity that would be perceived if the scene was actually scotopic. The model allows for real-time processing for interactive applications such as simulators or computer games.

Luminance perception at absolute threshold

When human luminance perception operates close to its lower limit the stereoscopic vision is no longer possible and the appearance changes substantially compared to common photopic or scotopic vision. Most observers report perceiving temporally-varying noise due to quantum noise (due to the low absolute number of photons) and spontaneous photochemical reactions. Previously, static noise with a normal distribution that does not adapt to absolute luminance intensity, was used to simulate the scotopic appearance on a photopic display for movies and interactive applications. Our perceptually-calibrated

computational model reproduces the experimentally derived distribution and dynamics of “scotopic noise” for a given luminance level and supports animated imagery. The real time simulation favorably compares to simpler alternatives in a perceptual experiment.

Zusammenfassung

Stereoskopisches 3D hat sich inzwischen als ein wichtiger Bestandteil in der digitalen Bilderzeugung, in Computerspielen und in der Filmproduktion etabliert. Dies wiederum hat auch intensive Forschungsbemühungen hervorgerufen, den Eindruck der Tiefenwahrnehmung zu verbessern. Ein besseres Verständnis der Einschränkungen von 3D Bildschirmen und dem menschlichen Wahrnehmungsvermögen hat den Weg zu einer intelligenteren Verarbeitung der Inhalte geebnet. Üblicherweise bereitet stereoskopisches 3D einige Probleme, wenn es auf aktuellen Geräten dargestellt wird. Das Unbehagen, das durch Rivalität der beiden Ansichten oder übermäßig große Disparitäten zusammen mit dem sogenannten Vergenz-Akkommodation-Konflikt ausgelöst wird, kann leicht die Vorteile der binokularen Wahrnehmung schmälern, Ermüdungserscheinungen hervorrufen und folglich dazu führen, dass das Betrachten in althergebrachtem 2D bevorzugt wird. Die Modellierung der Wahrnehmung des menschlichen Sehsystems (HVS) ist eine Möglichkeit, um sowohl Inhalte im Hinblick auf die beste Betrachtererfahrung für ein allgemeines Publikum zu optimieren, als auch spezifische Eigenschaften eines Geräts zu berücksichtigen oder auf individuelle Betrachter abzustimmen. In dieser Arbeit verwenden wir Wahrnehmungsstrukturen, um neue Herausforderungen in modernen Multimedia-Systemen anzugehen. Head-Mounted Displays (HMD) für virtuelle Realität (VR) und erweiterte Realität (AR) sind ein Beispiel für solche Geräte. Sie verwenden ein stereoskopisches Display zusammen mit einem Tracking des Benutzers für direkte Bewegungsinteraktion. Bildschirme mit hohem Dynamikumfang (HDR) sind eine weitere spannende Innovation, die den Verbrauchermarkt erreicht, und können potentiell ebenfalls in HMDs eingebaut werden. Diese Arbeit untersucht Wechselwirkungen von stereoskopischem 3D sowohl mit Bewegung als auch mit HDR, und ist entsprechend in zwei Teile gegliedert. Wir zeigen, wie sich die Eigenschaften des HVS unter Bewegung oder einem hohen Dynamikumfang der Adaptionsluminanz verändern und wie wir dieses Verhalten nutzen können, um die Wahrnehmungscharakteristiken bei stereoskopischen Bildschirmen zu verbessern. Auf diese Weise erreichen wir eine bessere Tiefenreproduktion, schwächen visuelle Artefakte ab und verbessern den Betrachterkomfort.

Die Forschungsmethodik ist in der gesamten Arbeit einheitlich, sowohl für den Teil über Bewegung als auch für den Teil über HDR Stereoskopie. Unsere Methoden sind fest in der Wahrnehmungsforschung verankert und erweitern diese oft durch eigene Messungen in Fällen, bei denen spezieller Bedarf aufgrund unserer Anwendungen entsteht. Wir verallgemeinern die Beobachtungen in ein Wahrnehmungsmodell, das wiederum in ein Berechnungsmodell umgewandelt wird, um unsere Eingabedaten zu verarbeiten. Dieses Modell kann entweder direkt als Metrik verwendet, oder - häufiger - in den Rahmen einer Optimierung integriert werden, welche eine Verbesserung der angestrebten Bildqualität wie Tiefenwiedergabe oder Betrachterkomfort zu erreichen

versucht. Am Ende wird der Kreis durch eine Validierungsstudie geschlossen, die bestätigt, dass Eigenschaften, die ursprünglich für einfache Stimuli beobachtet wurden, auch gelten, wenn das Modell auf komplexe Bilder angewendet wird.

Teil I: Stereoskopisches 3D und Bewegung

Im ersten Teil dieser Arbeit konzentrieren wir uns auf Bewegung in stereoskopischen Anwendungen. Diese Bewegung stammt sowohl vom Inhalt selbst, da Filme oder interaktive Anwendungen naturgemäß Bewegung beinhalten, als auch von der Eigenbewegung des Betrachters, wie sie durch HMDs mit Kopf- und Blick-Tracking nachgebildet wird und dadurch das Eintauchen in virtuelle Welten und neue Wege in der Mensch-Maschine-Interaktion erlauben. Über den üblichen Schwerpunkt zeitliche Kohärenz zu erreichen hinausgehend, studieren wir die zeitlichen Grenzen des HVS und wie Bewegung uns dabei unterstützt, die Welt zu verstehen. Im Einzelnen untersuchen wir die folgenden Aspekte von Bewegung:

Optimierung von Disparität für Bewegung in der Bildebene

Aufgrund einer Diskretisierung sowohl im Raum als auch in der Zeit unterliegt stereoskopische Bewegung, die auf einem Bildschirm abgebildet wird, einer Reihe von Einschränkungen im Hinblick auf die von Natur aus kontinuierliche Welt. Wir beschreiben, wie Aufnahmeprotokolle, die sich an den Inhalt anpassen, bei zeitsequentiellen Bildschirmen unechte Bewegungen in der Tiefenwahrnehmung reduzieren können. Eine andere Bewegungsverzerrung stammt von einer Kombination aus Bildschirmdesign und einer speziellen Einschränkung des HVS selbst. Wir untersuchen dieses Verhalten für einen anaplyphen Bildschirm und schlagen eine zeitliche Ausgleicheung des entstehenden Pulfrich-Effekts vor.

Optimierung von Disparität für Bewegung in die Tiefe

Die Wahrnehmung von dynamischem Stereoinhalt beruht darauf, das komplette Disparität-Zeit-Volumen, das ein bewegter Szenenpunkt durchläuft, wiederzugeben. Dieses Volumen kann während der Disparitätsmanipulation, die lediglich Disparitätsveränderungen in jedem Einzelbild berücksichtigt, stark verzerrt werden, selbst wenn die zeitliche Kohärenz dieser Veränderung beibehalten wird. Wir führen experimentelle Messungen durch um herauszufinden, wie empfindlich ein menschlicher Betrachter auf verschiedene Disparitätsverzerrungen reagiert. Daraufhin schlagen wir eine auf Wahrnehmungsmodellen zeitlicher Disparitätsveränderungen basierende Optimierung vor, die Stereobewegung von Inhalten nach einer beliebigen Manipulation der Disparität erhält. Des Weiteren führen wir ein neues 3D-Warping-Verfahren ein, das Stereobildpaare erzeugt, die dieser optimierten Disparität entsprechen. Nutzerstudien zeigen, dass unser Verfahren sowohl den Betrachterkomfort als auch die Leistungsfähigkeit in Aufgaben zur Tiefenbestimmung verbessert.

Blickgesteuerte Disparitätsmanipulationen

In vielen Szenarien ist der Tiefenumfang, der durch das Wahrnehmungselement der Disparität mittels Stereoskopie wiedergegeben werden kann, stark eingeschränkt und in der Regel durch technische Bedingungen festgelegt. In diesem Kapitel zeigen wir, dass dieses Problem erheblich verringert werden kann, wenn die Fixationsbereiche des Auges grob abgeschätzt werden können. Wir schlagen ein Verfahren zur stereoskopischen

Tiefenanpassung vor, das Augen-Tracking oder andere Informationen zur Vorhersage von Blickrichtungen verwendet. Anders als frühere Arbeiten wenden wir sukzessive, nicht wahrnehmbare Tiefenanpassungen während der Augenfixation an. Wir messen die Geschwindigkeitsbegrenzungen von Disparitätsveränderungen in verschiedenen Szenarios bei denen Tiefe angepasst wird, und formulieren ein neues Modell, das solch eine nahtlose Verarbeitung von stereoskopischen Inhalten steuert. Auf der Grundlage dieses Modells schlagen wir eine latenzunempfindliche Echtzeitsteuerung vor, die lokale Manipulationen an stereoskopischen Inhalten vornimmt, um das Optimum zwischen Tiefenwiedergabe und Betrachtungskomfort zu finden. Außerdem zeigen wir den Nutzen unseres Modells in Offline-Anwendungen, zum Beispiel als Vorverarbeitung in der stereoskopischen Filmproduktion. Eine Validierungsstudie belegt erhebliche Verbesserungen der Tiefenwahrnehmung, ohne dass bei der Anwendung unserer Methoden die visuelle Qualität leidet.

Bewegungsparallaxe als ein Tiefenwahrnehmungselement

Aktuelle Bildschirme, insbesondere automultiskopische, können lediglich einen beschränkten Disparitätsumfang wiedergeben. In diesem Kapitel untersuchen wir Bewegungsparallaxe, ein verhältnismäßig starkes Tiefenwahrnehmungselement, das jedoch völlig zweidimensional und deshalb in seiner Wiedergabe nicht beschränkt ist. In vielen praktischen Szenarien kann Tiefe durch Bewegungsparallaxe eine aggressive Kompression der Disparität kompensieren. Wir führen psychovisuelle Experimente durch, welche den Einfluss von Bewegungsparallaxe auf die Tiefenwahrnehmung messen, und setzen ihn mit der Tiefe, die von binokularer Disparität herrührt, in Beziehung. Unser gesamtheitliches Disparität-Parallaxe-Berechnungsmodell prognostiziert die empfundene Tiefe, die von beiden Wahrnehmungselementen stammt. Daraufhin stellen wir neue Methoden zur Disparitätsmanipulation vor, die zuerst die aus der Bewegungsparallaxe gewonnene Tiefe messen und dann entsprechend die binokulare Disparität anpassen. Eine Nutzerstudie zeigt, dass Zuordnungen des Tiefenbudgets in Abhängigkeit von der Stärke der Bewegungsparallaxe die Tiefenwahrnehmung insgesamt verbessert.

Teil II: Stereoskopisches 3D und HDR

Im zweiten Teil dieser Arbeit untersuchen wir zum ersten Mal die Wiedergabe von stereoskopischen Inhalten auf Bildschirmen mit hohem Dynamikumfang (HDR). Wir erforschen den Effekt von sowohl sehr hellen als auch sehr dunklen Luminanzniveaus auf die Disparitätswahrnehmung. Mit Hilfe von genauen Simulationen betrachten wir den gesamten Luminanzumfang, der vom menschlichen Sehsinn wahrgenommen werden kann und weiterhin die Möglichkeiten handelsüblicher Bildschirme überschreitet.

Wahrnehmung von Disparität bei photopischem Sehen

Helle Bilder, die bei Tageslicht entstanden sind, werden oft mit glänzend reflektierenden oder lichtbrechenden Oberflächen in Verbindung gebracht. Ungeachtet des Reizes, den diese blickwinkelabhängigen Effekte ausüben, verursachen sie bei ihrer Darstellung in stereoskopischem 3D doch eine Herausforderung. Lichttransport durch optische Schnittstellen kann leicht unangenehm große Disparitäten oder Verzerrungen der entstehenden Bilder zur Folge haben. Dies verhindert die Fusion der Bilder und verursacht unangenehme Rivalität. Obwohl solche Effekte auch in der Realität vorkommen, behaupten wir,

dass sie auf aktuellen Bildschirmen unerwünscht sind, da das Fehlen korrekter Akkommodation und die fehlende Möglichkeit, den unangenehmen Blickwinkel durch eine Kopfbewegung zu verändern, den Sinneseindruck erheblich verändern. Wir schlagen ein Optimierungsschema vor, das die Kameraparameter für jeden Pixel so modifiziert, dass visuell angenehme und realistische Disparitäten beibehalten und störende Rivalitäten verhindert werden. Wir validieren unseren Ansatz in einer Nutzerstudie, in der unsere Methode einen größeren Betrachtungskomfort und gleichzeitig einen höheren Realitätsgrad im Vergleich zu konkurrierenden Arbeiten erreicht.

Wahrnehmung von Disparität bei skotopischem Sehen

Das Erscheinungsbild einer skotopischen Nachtszene bei wenig Licht kann auf einem photopischen Bildschirm durch Farbentsättigung, verringerte Sehschärfe und die Purkinje-Blauverschiebung simuliert werden ("Day-for-Night"). Wir behaupten, dass eine wirklichkeitsgetreue Stereoabbildung von Nachtszenen auf photopischen Stereobildschirmen nicht nur eine Manipulation der Farben, sondern auch eine der binokularen Disparität erfordert. Zu diesem Zweck führen wir ein psychophysisches Experiment durch, um ein Disparitätsmodell für skotopische Luminanzniveaus zu konstruieren. Mit Hilfe dieses Modells können wir die binokulare Disparität von skotopischen Stereoinhalten, die auf einem photopischen Bildschirm dargestellt werden, an die Disparitäten anpassen, die wahrgenommen würden, wenn die Szene tatsächlich skotopisch wäre. Das Modell erlaubt eine Echtzeitverarbeitung für interaktive Anwendungen, wie zum Beispiel Simulationen oder Computerspiele.

Wahrnehmung von Luminanz an der Minimalreizschwelle

Wenn die menschliche Lichtmengenwahrnehmung nah an der Minimalreizschwelle arbeitet, ist stereoskopisches Sehen nicht mehr möglich und das Erscheinungsbild im Vergleich zum üblichen photopischen oder skotopischen Sehen verändert sich wesentlich. Die meisten Beobachter berichten von zeitlich variierendem Rauschen aufgrund von Quantenrauschen (bedingt durch die geringe Anzahl der Photonen) und spontanen photochemischen Reaktionen. Bisher wurde statisches normalverteiltes Rauschen, das sich nicht an die absolute Lichtmenge anpasst, benutzt, um ein skotopisches Erscheinungsbild auf photopischen Bildschirmen für Filme und interaktive Anwendungen zu simulieren. Unser wahrnehmungskalibriertes Berechnungsmodell bildet die experimentell hergeleitete Verteilung und Dynamik von "skotopischem Rauschen" bei einem gegebenen Luminanzniveau nach und unterstützt animierte Bilder. Die Echtzeitsimulation wird beim Vergleich zu einfacheren Alternativen in einem Wahrnehmungsexperiment bevorzugt.

Acknowledgments

I would like to thank everybody who supported me during 4 years of my PhD studies.

I thank to both my supervisors Prof. Karol Myszkowski and Prof. Hans-Peter Seidel for their advice and guidance. Karol was the person who invited me to join his research group at MPI Informatik. He greatly helped me to get myself familiarized with the role of perception in computer graphics as this was mostly a new field to me at the time. His research ideas, vast knowledge and patient guiding was one of the key aspects that allowed me to successfully finish my research projects and summarize their results into this thesis. Hans-Peter has made my study possible by creating and maintaining such an inspirational and welcoming environment in our computer graphics group.

Major part of my dissertation was done thanks to a close collaboration with Tobias Ritschel. He was my advisor on many of our projects and motivated me by his focus on achieving stated goals. He also educated me in issues of GPU computing. The software framework “Plexus” developed by him and his group has vastly improved efficiency of my work.

I thank to Prof. Wojciech Matusik who was my supervisor during the internship in his group at MIT CSAIL. I am grateful for the support and advice that he provided me. I also thank the rest of his group for making me feel welcome during this visit.

Also Piotr Didyk’s contribution to this thesis is very significant. His expertise in application of perception and stereoscopic 3D were key ingredients to success of many of our papers. His help was especially important during my internship as he was a connecting bridge for me between both MPI and MIT which was essential for a productive collaboration.

My thanks go to Łukasz Dąbala, Thomas Leimkühler, Peter Vangorp and all other co-authors of my publications during the PhD. Working with them was very inspiring and often also great fun. I cannot forget to mention Junaid Ali and all other students helpers who together greatly contributed to our work by assisting during many of our perceptual experiments.

A special thank goes to my colleges and office mates Krzysztof Templin and Yulia Gryaditskaya for an exceptionally pleasant and friendly atmosphere, help and fruitful discussion on both research and leisure topics both at work and in free time. Yulia has been a great support to me whenever I needed and I am very grateful for that. A similar can also be said about the rest of people that shared the time with me at MPI Informatik. Notably I want to express my gratitude to Sabine Budde and Ellen Fries who were always very kind and helpful while providing all the vital support in the administration and beyond.

Last but not least, I thank my parents Vladimír and Jitka as well as my sister Lenka for their support and care during the PhD as well as in the rest of my life.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Novel contributions	2
1.3	Overview	5
2	Background	7
2.1	Binocular vision	7
2.1.1	Stereopsis	7
2.1.2	Modeling	8
2.1.3	Visual discomfort in stereoscopic images	8
2.1.4	Role of display	9
2.2	Motion in depth	11
2.2.1	Role in depth judgment	11
2.2.2	Human sensitivity	11
2.2.3	Interactions	12
2.3	Eye motion	13
2.3.1	Eye vergence	13
2.3.2	Saccadic suppression	14
2.4	Motion parallax	14
2.4.1	Modeling	14
2.4.2	Depth from motion parallax	15
2.4.3	Disparity vs. motion parallax	16
2.4.4	Kinetic depth effect	17
2.5	Luminance perception	17
2.5.1	Scotopic, mesopic, and photopic vision	17
2.5.2	Photon quantum perception	18
2.5.3	Perception of luminance and depth	18
3	Previous work	21
3.1	Disparity processing	21
3.1.1	Disparity mapping	21
3.1.2	Temporal considerations	21
3.1.3	Viewing comfort enhancement	22
3.2	Gaze-driven applications	23
3.2.1	Gaze-driven disparity manipulation	23
3.2.2	Other gaze-driven applications	24
3.3	HDR processing	25
3.3.1	Tone mapping: Night scene depiction	25

3.3.2	Image noise	26
I	Stereoscopic 3D and motion	27
4	Optimizing disparity for screen plane motion	29
4.1	Correction for the Pulfrich effect	30
4.2	Correction for time-sequential presentation	32
4.3	Conclusions	37
5	Optimizing disparity for motion in depth	39
5.1	Experiment: Disparity distortion visibility	40
5.1.1	Description	41
5.1.2	Results	45
5.2	Our approach	49
5.2.1	Cost function	50
5.2.2	Perceived disparity velocity changes	52
5.2.3	Minimization	52
5.2.4	Upsampling	53
5.2.5	Implementation	54
5.2.6	3D warping	54
5.3	Validation	57
5.4	Conclusion	59
6	Gaze-driven disparity manipulations	61
6.1	Overview	62
6.2	Model for sensitivity to disparity manipulations	63
6.2.1	Experiment 1: Disparity shifting	63
6.2.2	Experiment 2: Disparity scaling	65
6.3	Our approach	66
6.3.1	Visible disparity change predictor	66
6.3.2	Seamless transition to target disparity	67
6.4	Applications	68
6.4.1	Real-time gaze-driven retargeting	69
6.4.2	Seamless disparity mapping in preprocessing	72
6.4.3	Scene cut optimization	73
6.4.4	Visibility visualization	74
6.5	Validation	75
6.5.1	Perceptual study	75
6.5.2	Limitations	77
6.6	Conclusions	78
7	Motion parallax as a depth cue	79
7.1	Joint motion parallax and disparity model	80
7.1.1	Methods	81
7.1.2	Data analysis and model fitting	82
7.1.3	Discussion	85
7.1.4	Definition for points and pixels	86
7.2	Our approach	87
7.2.1	Overview	88

7.2.2	Parallax map	88
7.2.3	Disparity scaling	89
7.2.4	Curve construction	89
7.2.5	Application to autostereoscopic displays	91
7.3	Validation	91
7.3.1	Perceptual studies	91
7.3.2	Discussion and limitations	94
7.4	Conclusions	95
 II Stereoscopic 3D and HDR		97
 8 Disparity perception in photopic vision		99
8.1	Our approach	100
8.1.1	Disparity model	101
8.1.2	Disparity extraction	101
8.1.3	Cost function	101
8.1.4	Optimization	102
8.2	Validation	103
8.3	Conclusions	103
 9 Disparity perception in scotopic vision		107
9.1	Overview	108
9.2	Experiments	108
9.2.1	Methods	109
9.2.2	Experiment 1: Optimal RSS frequency selection	110
9.2.3	Experiment 2: Luminance contrast detection threshold	111
9.2.4	Experiment 3: Disparity detection threshold	111
9.3	Model of wide-luminance range stereo sensitivity	112
9.4	Our approach	113
9.4.1	Disparity processing	115
9.4.2	Luminance processing	118
9.5	Validation	120
9.5.1	Results and discussion	120
9.5.2	User study	121
9.6	Conclusions	122
 10 Luminance perception at absolute threshold		125
10.1	Model of rod noise	126
10.2	Our approach	128
10.2.1	Photon counts	129
10.2.2	Simulation	130
10.2.3	Temporal integration	131
10.3	Validation	132
10.3.1	Performance	132
10.3.2	Results and discussion	132
10.3.3	Perceptual experiment	133
10.4	Conclusions	137

11 Summary	139
11.1 Conclusion	139
11.2 Future work	141
Bibliography – Own Work	143
Bibliography	145

Chapter 1

Introduction

The main goal of this thesis is to describe performance of binocular vision in combination with other modern imaging features. Previously, a significant research focus has been given to the performance of the human visual system (HVS) in perceiving stereoscopic images and videos in isolation. In our work we investigate how does the perception change when other properties such as motion, real time interaction or high dynamic range reproduction are combined with stereoscopic 3D. Such understanding could potentially pave the way for a more natural and enjoyable viewing experience even on existing displays. We describe several directions of integrating a perceptual models into a computational optimization of displayed content. This often allows to increase the reproduced depth together with subjective realism and at the same time reduce discomfort caused by display limitations.

1.1 Motivation

The HVS is tuned to work optimally in real world conditions. Through the evolution we have learned how different pieces of visual information can be fused into a complete and meaningful picture of the surrounding world. The perception of depth is not an exception. Many visual cues help us to understand distances from objects around us, and each is useful for a different depth range. Some of the cues are purely monocular, such as texture gradients, perspective or accommodation. The binocular disparity on the other hand relies on a fusion of information from both eyes. All these cues should naturally work together. However, this may not hold if we try to reproduce them artificially on stereoscopic display.

Stereoscopic displays use various techniques to deliver different image into each eye and to simulate parallax that would be created by a 3D object viewed from two different viewpoints. However, in reality both images are displayed on a single image plane of the display which means that cues like accommodation are not reproduced correctly. This causes a mismatch called the *vergence-accommodation* conflict. The name stems from the fact that the eye vergence follows the object position, possibly in front or behind the screen plane, while the lens accommodation is always fixed to this screen plane. Such conflict causes a visual discomfort and fatigue and must be carefully tuned to achieve good stereoscopic quality [Lambooi et al. 2009]. This on the other hand usually leads to a compression of depth with respect to the real world and triggers a conflict with other pictorial cues. A well known consequence of this is the *cardboard illusion* where the mismatch between perspective foreshortening and disparity of an object causes its flat appearance reminding unrealistic cardboard cutouts.

We need to find a balance between a good depth reproduction, realism and viewing comfort as on current displays these factors have to be traded off. In this thesis we analyze how motion and high dynamic range of luminance influence our sensitivity to either of these qualities. This way we can find a better solution than if the disparity is considered in isolation. We discuss each of these two interactions in two separate parts in order to focus on each of them in more detail. Regardless of that, our research methodology and the algorithmic approaches are consistent across the entire thesis and unify the structure of this work.

We focus on the role of motion and high dynamic range (HDR) imaging as both represent important directions for near future displays. We see that motion parallax can be introduced by the user just by moving her head while wearing the head mounted display (HMD). Information about the gaze motion is getting more accessible with cheap sensors and advanced computer vision algorithms. The HDR reproduction is getting available in consumer electronics as the technology matures and prices get lower. Each of such new visual features is exciting on its own but on the way to achieve an ideal real world viewing experience it is necessary that all of them work together. It is even more important for currently available displays that cannot reproduce either of these sensations perfectly and can therefore introduce new problems. The motion tracking in HMD is subject to a delay, the gaze estimation tends to be imprecise, the dynamic range of current HDR screens is very limited compared to the range of the human vision. This is why it is interesting to ask questions about how such limitations will interact with each other and in our case what would they mean for limitations already known for stereoscopic displays.

1.2 Novel contributions

The techniques described in this thesis are based on perceptual research which we extend by our own measurements as required for our applications. Initial observations for usually simple stimuli are generalized to derive a perceptual model of the HVS behavior. We then convert this abstract model into a computational model suited to the format of our input data and the problem that we aim to solve. Using such model we can predict performance of human vision in more complex scenarios common in computer graphics and use it either as a metric or as a part of an optimization scheme that aims to improve perceptual qualities of the image or video sequence. Both the derivation of the original perceptual model from our initial experiment and its application as a computational model introduce some amount of assumptions about a selection of experiment properties that can be neglected and properties that have to be preserved as a parameter. Therefore, it is vital to close the loop by validating the final application using an additional user study testing that observations made for a typically very simple stimuli also hold for a complex image or video.

The results of our work have been previously published in renowned computer graphics and applied perception journals, as well as presented on international conferences. Here we provide a detailed list of contributions of this thesis.

Improving perception of binocular stereo motion on 3D display devices

[Kellnhofer et al. 2014a, SPIE]

We analyze technical issues connected with displaying a dynamic stereoscopic content using two common 3D display technologies – anaglyph and shutter glasses. We identify that although the temporal displacement in the left and the right eye presentation originates in different phenomena for each technology it still leads to a similar illusion of a fake depth and a potential degradation of the presentation quality. We propose a computational remedy that does not involve hardware changes or additional costs.

Additional materials are available on-line at <http://resources.mpi-inf.mpg.de/TemporalStereo#spie2014>.

Optimizing disparity for motion in depth

[Kellnhofer et al. 2013, EGSR]

We investigate motion in depth which is especially significant for stereoscopic 3D as it directly affects stereoscopic disparity. We identify how common frame-by-frame disparity processing may influence the perception of such motion and we show that this can cause drop in the user performance in tasks that require precise depth judgment. Benefiting from the knowledge of the HVS limitations we construct an optimization framework that corrects the distortion introduced during the previous processing and restore the original motion where the error would be visible to a human. We demonstrate contributions of our approach in a validation study.

Additional materials are available on-line at <http://resources.mpi-inf.mpg.de/TemporalStereo/>.

What makes 2D-to-3D stereo conversion perceptually plausible?

[Kellnhofer et al. 2015a, ACM SAP]

We conduct a series of experiments to investigate limits of the HVS ability to perceive various disparity degradation types including a removal of temporal frequencies thus a distortion of motion in depth. We formulate recommendations for disparity quality requirements when it is intended for human perception rather than machine vision. This is useful for disparity processing, stereo quality metrics and also 2D-to-3D conversion methods which we discuss in detail in another paper [Leimkühler et al. 2016, GI].

Additional materials are available on-line at <http://resources.mpi-inf.mpg.de/StereoCueFusion/WhatMakes3D/>.

GazeStereo3D: Seamless disparity manipulations

[Kellnhofer et al. 2016a, SIGGRAPH]

The fully personal motion introduced by the gaze of a viewer is considered as an input for an adaptive stereoscopic display. We use the additional information to refine the depth appearance around the attended region beyond what traditional disparity mappers can do. We derive a model of disparity change visibility that allows us to handle gaze changes after saccades seamlessly. By avoiding disturbing changes we make our solution immune to latency issues of low cost trackers. This also potentially allows a usage of zero cost machine learning based eye trackers such as the one proposed in our recent paper [Khosla et al. 2016, CVPR]. We also propose other applications that

do not require online eye tracker but use statistical data or saliency maps to enable preprocessing for an average viewer. We validate our approach in a user study.

Additional materials are available on-line at <http://resources.mpi-inf.mpg.de/GazeStereo3D/>.

Motion parallax in stereo 3D: Model and applications

[Kellnhofer et al. 2016b, SIGGRAPH Asia]

We focus on the motion perception on a higher level and describe how certain types of motion help humans to understand the world around them. Motion parallax conveys information about a scene structure and we have previously studied human ability to understand it [Kellnhofer et al. 2015c, SPIE, Kellnhofer et al. 2016c, JEI]. Here we observe that the motion parallax serves as a depth cue and we experimentally derive a joint disparity and motion parallax model for the apparent depth. This way we predict how each of those two cues contributes to the resulting depth when we scale them independently. We propose a disparity mapping approach that accounts for the depth from the parallax and redistributes the overall disparity budget into regions with less parallax where the additional cue is most needed. This allows for a better overall depth reproduction with a smaller depth budget and thus more comfort for a viewer. We also show how this reduces visual artifacts in autostereoscopic displays and we validate our method in a user study.

Additional materials are available on-line at <http://resources.mpi-inf.mpg.de/StereoParallax/>.

Manipulating refractive and reflective binocular disparity

[Dąbala et al. 2014, Eurographics]

We focus on photopic day-like conditions. We observe that reflective and refractive surfaces that are so attractive in bright images can cause significant problems in the context of stereoscopic 3D. They cause possibly excessive and uncomfortable disparity by extending the scene depth. They also easily produce unfusible images for each eye which leads to an undesired rivalry. Our computational framework predicts, evaluates, and restricts such artifacts to the level necessary to depict physically impossible but visually believable, comfortable, and generally preferred images. We demonstrate this in a user study.

Additional materials are available on-line at <http://resources.mpi-inf.mpg.de/StereoRefraction/>.

Stereo day-for-night: Retargeting disparity for scotopic vision

[Kellnhofer et al. 2014b, ACM TAP]

We describe the transition between day and night conditions where the photopic vision switches to mesopic and then scotopic conditions and the performance of the entire visual system dramatically degrades. We measure, model and computationally simulate how such changes affect our ability to see stereoscopic 3D. This leads to a general high dynamic range disparity retargeting that can simulate the vision performance in arbitrary luminance conditions on a traditional low dynamic range display.

Additional materials are available on-line at <http://resources.mpi-inf.mpg.de/DarkStereo/>.

Modeling luminance perception at absolute threshold

[Kellnhofer et al. 2015b, EGSR]

We move our attention to even more extremely low luminance close to the absolute threshold of the human vision. The low correlation between noisy signals of both eyes effectively disallows the binocular fusion and the stereoscopic vision. We then model properties of such noise that originates both from random events in our eyes and from quantum properties of light. We propose a computational model for a postprocess algorithm simulating the vision noise in a video content. In a user study our approach achieves a higher level of night depiction realism than previous alternatives.

Additional materials are available on-line at <http://resources.mpi-inf.mpg.de/DarkNoise/>.

1.3 Overview

Chapter 2 focuses on a discussion of the psychophysical background on which we base our models of HVS that are introduced later. Chapter 3 overviews related work in the field of computer science that tackles similar problems as this thesis. The rest of the thesis is split to two parts discussing two major aspects of advanced stereoscopic displays – motion and HDR.

The first part investigates different forms of motion that can be experienced during stereoscopic viewing. In Chapter 4, we tackle issues connected with proper presentation of motion on two different types of stereoscopic displays. In Chapter 5, we discuss how disparity distortions introduced in various phases of a processing pipeline influence a subjective quality of perceived depth. We focus on a proper reproduction of motion in depth as an important scene understanding cue and propose a perceptually motivated optimization that ensures its proper reproduction. In Chapter 6, we study how information about motion of the eye itself can help us to achieve a locally adapted disparity mapping for a better viewing comfort and enhanced depth reproduction without introduction of temporal artifacts. In Chapter 7, we analyze the motion on a higher level and show that motion parallax plays a significant role in conveying depth. After measuring the size of this effect we derive a disparity remapping method for a better overall depth reproduction on displays with a restricted depth budget by careful balancing of both depth and motion parallax depth cues.

The second part investigates stereoscopic vision in wide range of absolute luminance levels covering what we call the HDR. In photopic day-light conditions reflection and refraction are common as specular or transparent surfaces cause reflection and refraction of light rays. This leads to view dependent image distortions which consequently introduce additional disparity and binocular rivalry. In Chapter 8, we investigate how a perceptual modeling and an optimization driven image synthesis can be used to prevent negative consequences such as discomfort or inability to fuse both images while preserving 3D appearance and realism. In Chapter 9, we move to mesopic and scotopic conditions and we model how stereoacuity degrades at such difficult viewing environment. We propose a computational approach for simulating similar experience on a traditional LDR display incapable of the real scotopic reproduction. In Chapter 10, we push the luminance levels even further all the way to the simulation of individual photon quanta arriving to human eyes near the absolute threshold of human vision. At such levels stereopsis is no longer feasible due to the dominant presence of random events in the HVS. Instead we model statistical properties of the resulting visual noise

and propose a computation model for a simulation of its appearance.

Chapter 11 concludes the thesis and discusses possible future research directions.

Chapter 2

Background

Here we provide a theoretical background discussing basic terms used in the thesis as well as HVS mechanisms exploited in our applications. We present an overview of the previous psychophysical research upon which we build our own measurements and perceptual modeling.

2.1 Binocular vision

In order to discuss more specialized topics related to individual chapters of this thesis we first introduce essential principles and terms connected to stereoscopic 3D itself.

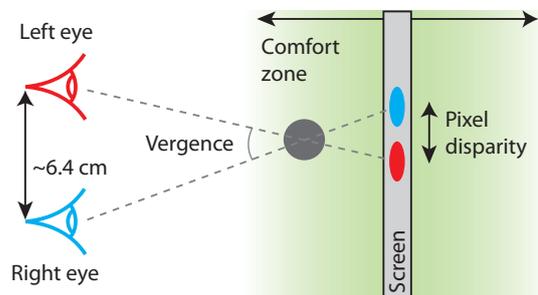


Figure 2.1. Basic scheme of stereoscopic presentation and parametrization of depth from binocular disparity using vergence angle and pixel disparity.

2.1.1 Stereopsis

Several visual depth cues are fused together by the HVS in order to obtain a robust estimate of surrounding depth [Landy et al. 1995]. Many of these cues, such as accommodation, occlusion, perspective, scale or texture gradient, are monocular but one of the strongest sources of depth information is stereopsis [Palmer 1999, Howard and Rogers 2012]. Utilizing binocular vision, stereopsis benefits from information in images perceived from two different viewpoints corresponding to the placement of human eyes approximately 6.4 cm apart (Fig. 2.1). This results in a parallax that can be estimated by fusion of both images and directly converted into an absolute depth image.

2.1.2 Modeling

We distinguish an absolute and a relative disparity. The absolute disparity corresponds to the *vergence* angle between view directions of both eyes and in psychophysics it is usually measured in arcmin units (Fig. 2.1). In computer graphics we typically refer to it as a *pixel disparity* and it is expressed as the distance between corresponding image locations in left and right eye images that is typically measured in pixel size units. Effectively it then describes the relative disparity to the screen. Note that the relation between both measures is non-linear.

The term *disparity* is then reserved for the difference of vergence angles and, therefore, stands for a relative disparity between two image locations expressed again in arcmin. An equivalent description can be made using screen pixel units. We refer to it as a *relative (pixel) disparity* to distinguish from the other terms. The terms *crossed* (or *negative*) and *uncrossed* (or *positive*) disparity are then used for disparity of objects in front and behind the screen plane respectively.

2.1.3 Visual discomfort in stereoscopic images

Stereopsis is a very immersive depth cue [Palmer 1999, Howard and Rogers 2012] but due to limitations of current display technologies it can be connected with negative perception effects that lead to visual discomfort and in long term fatigue. Overcoming such limitations is a major motivation of this thesis.

Most of current displays present the entire image on a single physical plane and cannot properly reproduce the focus cue that is needed for the eye *accommodation*. This leads to a conflict with the disparity cue as each predicts a different depth [Lambooi et al. 2009]. Such mismatch can be tolerated by the HVS in a small *comfort zone* (Fig. 2.1) around the screen depth plane [Shibata et al. 2011]. However, if the discrepancy is too large it can cause eye strain and headache especially for small viewing distances at which the accommodation cue plays a significant role. The solution is usually a compression of binocular depth which in turn may lead to an undesirable flat depiction. A significantly lower tolerance was observed for a vertical disparity [Tyler et al. 2012].

A similarly uncomfortable effect can also emerge from extensive relative disparity. If the disparity between the front and back of an depth edge cannot be fused inside so-called Panum's fusional area it results in a double vision effect called *diplopia* [Howard and Rogers 2012, Ch. 14].

A necessary condition for extraction of depth in binocular vision is fusion of both eye images. This process is similar to stereoscopic matching in computer vision and requires a local similarity between both views. Such assumption can be violated from various reasons and leads to an uncomfortable state called *rivalry* [Howard and Rogers 2012, Ch. 12]. Some display technologies suffer from ghosting where one view propagates to another one which triggers rivalry due to asymmetry. Improper adjustment of stereoscopic cameras may also lead to visual differences. Some mismatches between views are even physically correct as they emerge from view-dependent surface properties [Templin et al. 2012, Dąbala et al. 2014]. This is truth for reflections and refractions and we discuss in more details in Chapter 8.

Conflicts of depth cues are source of discomfort in general and can also break the immersion effect. Stereographic rules are followed in cinematography in order to prevent them [Lambooi et al. 2009]. A *border violation* occurs when objects with crossed disparity (in front of the screen) are occluded by a screen border and this way a conflict between these two strong depth cues is triggered [Lambooi et al. 2009]. We

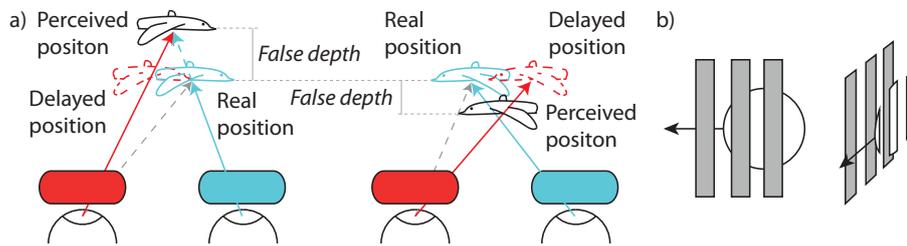


Figure 2.2. *a)* Pulfrich effect on a stereoscopically viewed horizontally moving bird. Red-cyan anaglyph colors are used to code individual eye images. The right eye image represents the real image-space position of the bird (cyan), the left eye image is perceived delayed (red). This results in the final percept being shifted in depth (black). *b)* Disparity-occlusion clue conflict arises in such conditions.

can reduce this problem by adaptive disparity remapping described in Chapter 6.

2.1.4 Role of display

Here we discuss limitations of specific stereoscopic display technologies and implications of such limitations on depth perception by the HVS. We utilize this in Chapter 4 to design a display algorithm improving the quality of depicted depth.

Anaglyph display

The anaglyph glasses are a popular and affordable technology suitable for occasional 3D content viewing and for quick content presentation. It uses simultaneous presentation of images to both eyes where the separation is provided by color filters. Frequency characteristics of these filters differ between individual technologies and it affects mainly the ability to reproduce original chromacity and the amount of crosstalk between eyes [Woods and Rourke 2004]. The most common configuration uses red filter for the left eye and cyan filter for the right eye. The big advantage over other technologies is that stereo images can be printed and thus is not limited to active, electronic display devices.

Left and right eye filters have not only different dominant colors but also level of transparency across the visible light spectrum [Woods and Rourke 2004]. This interacts with variable sensitivity of the HVS to different wavelengths [Thornton 1999] and makes each eye perceive a different level of illumination. In case of red-cyan glasses the left eye's red image is perceived darker than the right eye's cyan one. In summary, one image of the stereo pair appears brighter than the other.

The Pulfrich effect

A darker signal takes the HVS longer to process which causes the so-called “Pulfrich” effect. Consequently, a pair of stereo images with different brightness is therefore apparently shifted in time relatively to each other as described by Lit [1949]. If the user watches a moving object, the “bright eye” sees the object in position at time t while the “dark eye” sees it in a delayed position at time $t - \delta T$. This creates image disparity that is interpreted as an additional depth offset (Fig. 2.2a). If the motion speed is not constant the resulting disparity changes over time, introducing a false motion-in-depth cue. This principle has been used as a cheap and simple way of 2D-to-3D movie conversion [Beard 1991].

It, however, does not provide any control over the amount or sign of generated depth and therefore it is likely to conflict with other cues, such as occlusion (Fig. 2.2b). If there is an object moving from right to left behind a fence, it will appear to be shifted toward the viewer wearing darker filter on his left eye, e.g., when using cyan-red glasses. Consequently, the disparity cue will predict the object's depth to be in front of the fence violating the occlusion depth cue [Mendiburu 2009]. The effect is not symmetric as the opposite motion will generate opposite depth shift. Horizontal motion is, however, just one special case of general motion. Vertical motion will introduce vertical disparities which might reduce viewing comfort or even prevent fusion [Woods et al. 1993]. A typical example for this motion is falling rain.

Protocols

Capture or presentation protocols describe how image data for the left and right eye are captured or presented in time. The two basic types are *simultaneous* and *alternating* protocols. With simultaneous protocols both images either describe the world at the same time in case of capture protocol or are displayed at the same time in case of presentation protocol. With the alternating protocol images are either captured or displayed at alternating time sequences. We follow the notation on Hoffman et al. [2011] and denote the combination using abbreviations of capture protocol and presentation protocol in that order, e.g., SIM/ALT for “simultaneous capture with alternating presentation”.

The multiplexing technology usually determines the type of presentation protocol of choice. For time-sequential displays the alternating presentation is the only option. If simultaneous capture protocol was chosen it would produce conflict of time as the image presented in the second eye would be delayed with respect to its capture time. The resulting effect would introduce additional disparity, hence false motion-in-depth similarly as in the Pulfrich effect. Additionally, it can give raise to vertical disparity for vertical motion, reducing viewing comfort. Therefore matching the capture and presentation protocols, i.e., using SIM/SIM or ALT/ALT, is recommend.

However, as it was shown by Hoffman et al. [2011], the situation changes, when frame repetition is required. Such repetition is employed commonly in the cinema where multiple flashes of identical frames are presented to reduce flickering for movies with relatively low frame rates. It was pointed out that especially repeating frames with an alternating presentation protocol might introduce a false motion-in-depth sensation. The proposed model assumed that the time offset capture that matches to ALT/ALT for analogical single flash presentation should minimize depth perception issues. We will further use just ALT/ALT for such protocol with frame repetition. It was, however, observed that this choice is appropriate only for slow angular speeds of video content motion and SIM/ALT is a better choice for speeds above some threshold. This observation is explained by the *temporal disparity-gradient limit* $|\Delta\delta/\Delta t|$, a change of disparity over time, after which the HVS' disparity estimation fails. Hoffman et al. [2011] conclude with the recommendation to select between the two protocols discussed based on the probability of frame repetition.

Protocol choice might be further complicated if the capture frame-rate is changing over time which is a typical case in real time rendered content. Time-sequential presentation based stereo 3D technologies using active shutter glasses typically operates at 120 Hz to avoid flickering. Therefore rendering at 60 Hz for each eye is necessary in order to avoid frame repetition. That is too much for most of the current middle-range consumer HW and most up-to-date games. We can save some performance by generating every other frame using warping of the previous frame. Didyk et al. [2010a]

used blurring to hide the resulting warping artifacts. They argued that blurring of in one eye does not reduce the overall sharpness. However, warping might not always be sufficient cure for the performance problem, as one would start to see quality decrease if too many interpolated frames were inserted. In this case frame repetition is required and used. It means that the preferred capture protocol might change over time according to conclusions of Hoffman et al. [2011].

2.2 Motion in depth

Motion in depth (MID) is a very important type of motion for stereoscopic 3D as it can utilize a change of disparity over time in order to enhance content dynamics. In this section we describe the role of MID for judgment of the scene layout, our sensitivity to MID, and its interaction with other depth cues.

2.2.1 Role in depth judgment

Dynamic changes of binocular disparity naturally arise through any form of object motion in the surroundings. Even when the eyes perfectly converge on an object moving in depth, which results in the null absolute disparity for this object, the relative disparity with respect to other objects creates a strong cue for detecting motion-in-depth (MID), and estimating its direction and velocity [Erkelens and Collewijn 1985]. While monoscopic cues such as changing object size and its visibility/occlusion configurations, perspective deformations of inclined surfaces, and lens accommodation may contribute to the motion judgment as well, dynamic binocular disparity greatly improves the precision of motion perception [Gray and Regan 1998].

Such reliable motion judgment is required in many everyday tasks such as estimating the time when the approaching object will reach a specified position, called also the time-to-contact (TTC), determining the object impact direction, or performing the interception task of one moving object by another. Clearly, these tasks are of high relevance in many computer game and training simulator scenarios as well, where the participant performance may critically depend on the precision of perceived motion.

It is believed that two binocular mechanisms might contribute to the MID perception, but their precise role is still an open research problem [Harris et al. 2008]. A changing disparity over time (CDOT) mechanism (called also the stereo-first mechanism) determines relative disparities between scene elements and monitors their changes. An interocular velocity differences (IOVD) mechanism (called also the motion-first mechanism) relies on combining two monocular velocity signals that are derived based on temporally coherent motion patterns separate for each eye.

2.2.2 Human sensitivity

The sensitivity studies for the MID detection, which have been performed for various temporal frequencies of disparity modulation, revealed the peak sensitivity within the range 0.5–2 cycles-per-degree [Tyler 1971, Richards 1972], and the high-frequency cutoff at 10.5 Hz [Neinborg et al. 2005], which is significantly lower than 60 Hz as measured for temporal modulation of luminance contrast.

Harris and Watamaniuk [1995] and Portfors-Yeomans and Regan [1996] investigated the sensitivity of human visual system (HVS) to speed changes in MID, which arise due to the CDOT and IOVD mechanisms. They found the sensitivity to the motion in depth

(MID) speed expressed as vergence angle differentials is constant with distance to screen, and it follows Weber's Law, where the ratio (the Weber fraction k) of discriminated speed change to a reference speed typically varied between 0.08 up to 0.25. Interestingly, the Weber fraction does not significantly depend on the magnitude of disparity [Brooks and Stone 2004, Fig. 6], and whether the object moves away or approaches the observer. At the same time, this fraction gets higher for complex scenes due to the presence of monocular cues [Harris and Watamaniuk 1995]. Based on those findings, we derive a model of perceived disparity velocity changes in Sec. 5.2.2.

The sensitivity to the MID is poorly correlated with the sensitivity to frontoparallel motion, but it is well correlated with the static disparity sensitivity under different base disparity and defocus conditions [Cumming 1995]. The HVS sensitivity to temporal disparity changes seems to be relatively low [Kane et al. 2014]. We exploit this property in our seamless remapping model in Chapter 6 and also in our 2D-to-3D conversion [Leimkühler et al. 2016].

A unique target with distinctive frontoparallel motion is known to have a very strong "pop-out" effect. However, no such effect was observed when disparity, as the only cue, induced MID of stereoscopically observed targets [Harris et al. 1998].

In Chapter 6, we complement these findings by measuring the HVS sensitivity for the speed of a scene depth range change as well as the speed of smoothly shifting the fixation point towards the accommodation plane (the screen plane) as relevant for stereoscopic displays.

2.2.3 Interactions with other depth cues

Interaction of binocular MID and monocular cues (in particular the change of size) typically leads to the overall improvement of motion judgement. Gray and Regan [1998] found that for separately considered monocular and binocular cues consistently underestimated or overestimated values of absolute TTC are obtained, while the accuracy improved significantly when both cues are available. As the linear horizontal width of a moving object decreases the reliability of monocular information drops [Gray and Regan 1998], and then the precision of TTC task might fully rely on the quality of binocular information. Surprisingly, binocular vision seems to be important in the TTC task for distances relevant for highway driving up to 75 m [Cavallo and Laurent 1988]. As observed by Regan and Beverley [1979] with increasing motion speed or inspection times (lower framerates) the changing-disparity cue becomes more effective in conveying the MID sensation than the changing-size cue. This is also the case when MID is accompanied by more complex shape changes than simple isotropic rescaling, which may arise for deformable or rigid, but non-rotationally symmetric objects. Also, the detection thresholds for just noticeable MID are typically lower for the binocular cues than for their monocular counterparts. Regan and Beverley demonstrated that a change in size can be cancelled by an antagonistic change in relative disparity, and proposed a simple weighted-sum model to combine both cues.

Heuer [1987] reported that for contradictory cues, rivalry can be observed instead of summation, which may lead to the instability of dominating cue. Brenner et al. [1986] suggest that conflicting cues might be responsible for large differences between subjects in the motion judgment, and propose that most likely scene interpretation that is selected by subjects should minimize the cue conflicts. Gray and Regan [1998] observe that the human performance in the TTC task is decreasing for distorted stereo configurations.

All the above indicates that the high accuracy of dynamic disparity information is required to enable reliable MID judgement, which is instrumental in numerous practical

tasks. In many object motion scenarios dynamic disparity is the only reliable or the most effective MID cue to perform those tasks. Even in the presence of other strong MID cues, their effectiveness can be seriously degraded when combined with distorted binocular disparity. Our perceptual study in Sec. 5.3 is an example where distorted stereo has a significant effect on task performance. As we discuss in Sec. 3.1.1, such distorted disparity information is quite common in stereo 3D imaging and computer graphics applications.

2.3 Eye motion

In this section we briefly discuss the dynamic aspects of the stereovision during eye motion. In Chapter 6, we propose a dynamic mapping that adapts disparity to enhance depth and viewing comfort locally around the gaze location. Such temporal change of disparity will inevitably lead to motion in depth (MID, Sec. 2.2). We consider two cases that fully determine the HVS operation modes in the response to such depth changes:

(1) When a target slowly changes its position in depth and screen space, the eye vergence is combined with a smooth pursuit eye motion to fuse the left and right images and maintain the target in the foveal region.

(2) When fixation switches between two targets that significantly differ in the depth and screen location, vergence must be combined with saccadic eye motion to achieve the same goal.

In the former case, the speed of MID is the key factor that determines the visibility of resulting depth changes and activates different eye vergence mechanisms (Sec. 2.3.1). In the latter case, the saccadic suppression is the dominant factor in hiding depth changes (Sec. 2.3.2).

2.3.1 Eye vergence

Eye vergence is performed through a fast and possibly imprecise transient (trigger) mechanism that is activated for larger depth changes as well as a slower and precise sustained (fusion-lock) mechanism that compensates for the remaining fusion inaccuracies [Semmlow et al. 1986]. Slower depth changes with the ramp velocity below 1.4 deg/s can be fully processed by the sustained mechanism, while the motoric eye vergence (transient or sustained mechanisms) might not be even required for small depth changes that are within Panum's fusional area, in which case sensoric fusion in the brain might be sufficient. For stereoscopic displays the eye vergence is excessively dragged towards the screen plane and at the screen depth the vergence error is smallest [Duchowski et al. 2014a]. This may be caused by the accommodation-vergence cross-link when the incorrect focus cue at the screen plane shifts the vergence towards the screen [Kim et al. 2014].

In Chapter 6, we control the speed of depth manipulations, so that only the sustained mechanism and the sensoric fusion are activated, which minimizes intensified efforts of the oculomotor system. At the same time, the eye vergence is kept as close to the screen plane as possible, which reduces the vergence error, the vergence-accommodation conflict, the frame violation effect [Zilly et al. 2011], and crosstalk between the left and right eye images [Shibata et al. 2011], and improves the viewing comfort and quality [Peli et al. 2001, Hanhart and Ebrahimi 2014]. Utilizing the slow mode of vergence adaptation also relaxes requirements for the eye tracking performance

and cheap solutions can potentially be used for easier integration without specialized hardware as shown by Khosla et al. [2016].

2.3.2 Saccadic suppression

The HVS cuts off the sensory information during fast saccadic eye motion to avoid the perception of blurred retinal signal, which is referred as the saccadic suppression. The duration of actual saccadic eye motion depends on its angular extent and falls into the range of 20–200 ms. Each saccade is preceded by a preparatory stage with a latency of 200 ms, where new sensory information is cut off 80 ms prior to the eye motion initialization to the saccade completion [Becker and Juergens 1975]. McConkie and Loschky [2002] have shown that a switch from a significant blur to a sharp image can be detected by the observer even 5 ms after the end of a saccade. However, the tolerable delay can grow to up to 60 ms [Loschky and Wolverson 2007] for more subtle blur changes as in multi-resolution techniques [Guenther et al. 2012, Geisler and Perry 1998].

The eye vergence is a relatively slow process that for stereoscopic displays takes in total about 300–800 ms [Templin et al. 2014a]. The actual time depends on the initial disparity and the vergence motion direction. In general, the motion towards the screen plane is faster than in the opposite direction with the maximum velocity of about 20 deg/s. The latency prior to the eyeball vergence initialization amounts to 180–250 ms [Semmlow and Wetzel 1979, Krishnan et al. 1973]. This effectively means that the eye vergence motion is typically continued after the saccade completion and approx. 100 ms are needed before new sensory information can actively guide the vergence to the target depth. In this respect, our seamless disparity manipulation in Sec. 6.4.1 shows some similarities as it may induce eye vergence motion after the saccade completion.

2.4 Motion parallax

Motion parallax is a depth cue that results from observer movement. As we move, objects closer to us move in the visual field faster than the objects that are further away. This relative velocity is used by the HVS to recover the distance between different points in the scene. Motion parallax can also be triggered when two points in the scene undergo a rigid transformation, e.g., translation (Fig. 2.3). In this section, we provide background information about this mechanism, and discuss previous work on modeling it. In particular, we concentrate on the relation of motion parallax to binocular disparity, as well as possible interactions between these two cues.

2.4.1 Modeling

Binocular disparity is parameterized as a difference in vergence angles of the point of interest and the fixation point (refer to Fig. 2.3). Motion parallax can be parameterized in *equivalent disparity* units, which enables a direct comparison to binocular disparity [Rogers and Graham 1982, Ono et al. 1986, Bradshaw and Rogers 1996]. Equivalent disparity is defined as the maximum relative displacement between a peak and a trough in the stimulus as the head (or the stimulus) moves across the interocular distance.

Modeling motion parallax by the means of the equivalent disparity makes an assumption about the absolute scale of depth in the scene. As motion parallax is a purely relative depth cue, we can choose an arbitrary scaling factor and apply to both the scene geometry and motions which will still generate the same 2D image sequence, and hence,

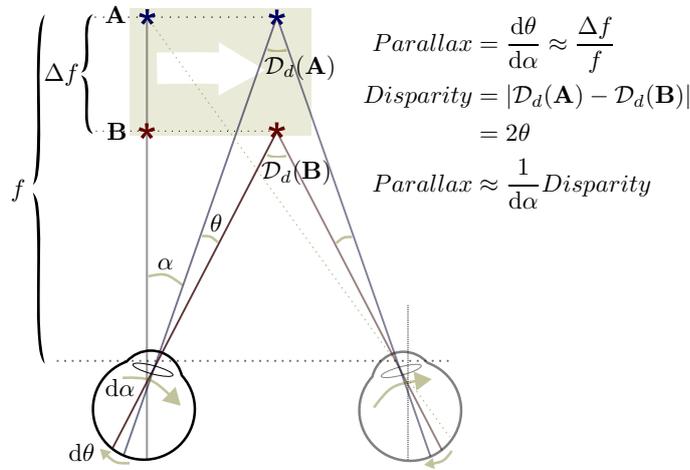


Figure 2.3. Motion parallax mechanics [Nawrot and Stroyan 2009]. Translation of a pair of points A and B to the right (or equivalently observer translation to the left) produces motion parallax. Assuming that A is the fixation point, its retinal position does not change, due to the pursuit eye rotation α at angular velocity $d\alpha/dt$, while at the same time the retinal position of B changes with respect to A by the angular distance θ producing retinal motion with velocity $d\theta/dt$. The relation $d\theta/d\alpha \approx \Delta f/f$ approximately holds, which, given the fixation distance f , enables to recover depth Δf from motion parallax. The right side of the figure shows the relation of motion parallax to the angular measure of binocular disparity $\mathcal{D}_d(\mathbf{A}) - \mathcal{D}_d(\mathbf{B})$ [Howard and Rogers 2012, Fig. 14.5] for asymptotic case of $f \rightarrow \infty$ as derived by Stroyan [2010].

the same motion parallax. However, assuming that human inter-ocular distance is fixed, each of these scenes would have different disparities. Similarly there will be multiple possible parallax values for each disparity distribution. This ambiguity can be replaced by measuring directly the relative depth as $m = \Delta f/f$ [Nawrot and Stroyan 2009] (Fig. 2.3). This particular form prevents division by zero and restricts possible values to the range from 0 to 1.

Note that as a static scene description the relative depth is not a definition of motion parallax itself. A necessary condition for existence of motion parallax is presence of rigid motion which requires a dynamic definition. Nawrot et al. [2009] show that the perception of motion parallax follows the ratio of the retinal motion velocity of distractor \mathbf{B} ($d\theta/dt$) and the eye motion velocity during pursuit of fixate \mathbf{A} ($d\alpha/dt$) (Fig. 2.3). These two signal allow the HVS to derive the relative depth as it can be shown that $d\theta/d\alpha \approx \Delta f/f$ [Nawrot and Stroyan 2009].

2.4.2 Perceived depth from motion parallax

While processing of depth from disparity and motion parallax seem to engage common neural mechanisms [Rogers and Graham 1982, Bradshaw et al. 2006], there are also notable differences between them. The HVS sensitivity to motion parallax as a depth cue has been measured in the equivalent disparity experiment where static sinusoidal depth corrugations have been used as the reference for perceived depth [Rogers and Graham 1982, Bradshaw and Rogers 1996, Bradshaw et al. 2006]. The respective sensitivity function has a shape similar to the disparity sensitivity function (DSF) [Bradshaw and

Rogers 1999] with a maximum around 0.4 cpd, however, the threshold magnitudes are 2 to 10 times higher than for disparity.

Unlike for binocular disparity, depth constancy was found rather poor for motion parallax, as the estimated depth increased with the square of a feature's distance [Ono et al. 1986]. When geometrically equivalent depth is presented through motion parallax or binocular disparity as the only depth cue, foreshortening between about 25%-125% has been observed for the motion parallax presentation [Durgin et al. 1995]. Nawrot et al. [2014] performed a matching experiment, where the perceived depth as induced by motion parallax has been compared to equivalent depth resulting from binocular disparity, and observed a near-tenfold depth foreshortening. Such depth foreshortening should be understood as a perceived distance reduction for a given object with respect to the fixated point. Nawrot et al. found that to model perceived depth due to motion parallax compressive nonlinearities (transducers) need to be applied in the ratio θ/α (Fig. 2.3). We extend the model of Nawrot et al. [2014], which was derived for monoscopic 2D moving stimuli, to stereoscopic stimuli with binocular disparity. This enables modeling perceived depth for different combinations of motion parallax and disparity, which is relevant for stereo 3D applications. Moreover, the model by Nawrot et al. does not account for the relative velocity on the display, which directly translates into retinal velocity. Such relative velocity is characterized by its own detection thresholds [Campbell and Maffei 1981], which affect the ability to detect motion parallax as well. Therefore, our model accounts for relative velocity in a presence of binocular disparity.

2.4.3 Disparity vs. motion parallax on 3D displays

Both the reduced threshold sensitivity and depth foreshortening for supra-threshold depth ranges indicate that motion parallax is a weaker depth cue compared to binocular disparity. However, since in stereoscopic displays disparity is usually strongly compressed for viewing comfort [Lang et al. 2010, Masia et al. 2013a, Chapiro et al. 2014], the apparent depth from motion parallax becomes comparable as it always refers to the original, uncompressed depth. In Sec. 7.2 we seek the opportunities of disparity range compression in regions, where depth perception is well supported by motion parallax. Note that disparity is constrained by the fixed interocular distance, while motion parallax can still be detected at far distances, given that the observer velocity is sufficiently high. An additional advantage of motion parallax is that in contrast to disparities which may cause diplopia on certain 3D displays, motion parallax always leads to physically plausible scene interpretations without additional effort.

Disparity and motion parallax cue fusion Different classes of tasks combining motion parallax and disparity depth cues lead to different models of cue fusion. In a surface detection task from random dots Turner et al. [1997] did not observe the performance to increase when the two cues were present, and noticed that disparity was dominant when contradicting cues were shown. Young et al. [1993] have looked into perceived depth from a combination of (inconsistent) motion and texture cues through perturbation analysis and argued for a weak fusion model, a weighted linear combination with weights decreasing if a cue is contaminated by noise. Such a strictly modular approach to both cues is not supported by experimental evidence that arises for example in surface recognition tasks involving motion parallax and disparity, where cooperative and facilitative interactions have been identified [Howard and Rogers 2012, Ch. 30.2.6]. Landy et al. [1995] presented a *modified weak fusion* model for cue combination, which can handle such interactions through cue promotion when they are

in an agreement with other cues. In Chapter 7, we propose a model, which is derived in a perceived-depth matching task, where cue interactions are explicitly measured for various combinations of geometrically correct motion parallax and compressed or expanded disparity values. The range of motion parallax and disparity magnitudes modeled fully covers the requirements of stereoscopic image manipulation (Sec. 7.2).

2.4.4 Kinetic depth effect

An object structure can be recovered from its rotating motion, which is referred to as the kinetic depth effect (KDE) [Wallach and O’Connell 1953]. Durgin et al. [1995] demonstrated that similar as for motion parallax the recovery of quantitative depth information from object rotation in monoscopic images is weaker than from binocular disparity. Recently, Bista et al. [2016] proposed an approach for the KDE triggering from a pair of photographs, where first a rough scene geometry (mesh) is reconstructed, then an optimal location for the scene rotation axis that is parallel to the screen is derived, and finally swinging rotation of the camera around this axis is introduced. In Chapter 7, we consider motion parallax, which is inherent for an animation rather than artificially generated rotations as in [Bista et al. 2016]. Also, we focus on the binocular vision and disparity manipulation rather than monocular images and mesh deformation to bring the relative velocity between rotating scene elements into desirable ranges as proposed by Bista et al. [2016].

2.5 Luminance perception

A luminance signal is essential for expression of disparity as it serves as its carrier. The influence of the high dynamic range perceivable to the human vision on binocular disparity perception is a topic of the second part of this thesis. Here we study previous theoretical research of this topic after covering basic properties of luminance perception as well vision performance under extreme lighting conditions.

2.5.1 Scotopic, mesopic, and photopic vision

Human vision operates in a wide range of luminance values (10^{-6} to 10^8 cd/m^2). We can distinguish *photopic* (10^8 to 3 cd/m^2), *mesopic*, (3 to 0.1 cd/m^2), *scotopic* (0.1 to 10^{-6} cd/m^2) vision, and scotopic vision *close to absolute threshold* (less than 10^{-3} cd/m^2).

In the human retina two types of photoreceptors: cones and rods, perform the visual tasks. In photopic conditions, *cones* are active. They have an uneven distribution with a strong peak in the fovea and contribute to color vision. They are inactive at night [Palmer 1999], and in Chapter 9 we assume they do not contribute to the modeled scotopic effects. We also do not consider the mesopic range, in which both rods and cones are active.

In scotopic night vision, only *rods* are active. They have a different response to light of different wavelengths [Wald 1945] and do not contribute to color vision. Their peak density is lower than for cones, but their distribution over the retina is more uniform and shows a slower falloff with the retinal eccentricity.

The rod and cone responses are modeled in many applications, such as image quality metrics [Mantiuk et al. 2011b], color appearance models [Kuang et al. 2007], and tone mapping operators [Ferberda et al. 1996, Durand and Dorsey 2000, Ward et al. 1997,

Pattanaik et al. 2000, Thompson et al. 2002, Khan and Pattanaik 2004, Wanat and Mantiuk 2014]. In mesopic conditions, a linear combination of rod and cone response is typically considered. A biologically-inspired model that predicts the offset in the L, M, and S cones due to rod responses has been used in tone mapping [Kirk and O'Brien 2011].

2.5.2 Photon quantum perception

Human vision is based on translating light into nerve signals. Light can be modeled as rays, waves, individual particles, or as their quantum statistics. In Chapter 10, different from a commonly taken viewpoint in computer graphics and vision, we choose the quantum-statistics point of view. Here, “light” for a space-time interval is not a single value anymore, but modeled as a distribution of probabilities to observe a certain number of quanta.

Light enters the human eye through the pupil, and is subject to different scattering and absorption events, before it reaches the retina, which is covered by receptors converting incoming light to electric signals.

Rods convert light into nerve signals using a *photo-chemical* cascade. Each rod contains rhodopsin, that is isomerized by exposure to light, resulting in a small change of potential to become a nerve signal [Alpern 1971]. In each following step of the cascade, non-linear functions amplify the signal, while at the same time suppressing noise. The temporal aspects of photo-transduction are the cause of afterimages [Ritschel and Eisemann 2012].

Not all photons hitting a receptor are actually transduced into an electrical signal (false negative) because it might happen that the photon does not hit a rhodopsin molecule. The ratio of transduction (ca. 0.06–0.2, [Hecht et al. 1942]) is called *quantum efficiency*. At the same time, it happens that rhodopsin is transduced in the absence of light (false positive) [Barlow 1956]. These aspects will be detailed in our treatment of near-absolute-threshold light levels in Sec. 10.1.

Finally, other entoptic phenomena, which are not directly caused by light in the common way, such as floaters, phosphenes, visual snow, the blue-field entoptic effect [Riva and Petrig 1980], or afterimages [Ritschel and Eisemann 2012], can occur under specific conditions but are not related to scotopic vision.

2.5.3 Perception of luminance and depth

Since luminance and depth edges often coincide, e.g., at object silhouettes, full-resolution RGB images have been used to guide depth map upsampling both in the spatial [Kopf et al. 2007] and the spatio-temporal [Richardt et al. 2012, Pajak et al. 2014] domain. Analysis of a database with range images for natural scenes reveals that depth maps mostly consist of piecewise smooth patches separated by edges at object boundaries [Yang and Purves 2003]. This property is used in depth compression, where depth edge positions are explicitly encoded, e.g., by using piecewise-constant or linearly-varying depth representations between edges [Merkle et al. 2009]. This in turn leads to significantly better depth-image-based rendering (DIBR) [Fehn 2004] quality compared to what is possible at the same bandwidth of MPEG-style compressed depth, which preserves more depth features at the expense of blurring depth edges.

The spatial disparity sensitivity function determines the minimum disparity magnitude required to detect sinusoidal depth corrugations of various spatial frequencies [Howard and Rogers 2012]. The highest resolvable spatial frequency is about 3–4 cpd

(cycles per degree), which is almost 20 times below the cut-off frequencies for luminance contrast [Wandell 1995]. Similar investigations in the temporal domain indicate that the highest sinusoidal disparity modulation that can be resolved is about 6–8 Hz [Howard and Rogers 2012], which is significantly lower than the 70 Hz measured for luminance [Wandell 1995]. As analyzed by Kane et al. [2014], the picture is different for disparity step-edges in space and time, which are important in real-world images. They found that, for step-edge depth discontinuities, observers might still notice blur due to removal of spatial frequencies up to 11 cpd, indicating that while overall disparity can be smoothed significantly, this is not the case for depth discontinuities. They could further show that filtering temporal frequencies higher than 3.6 Hz from a step signal remains mostly unnoticed. Their findings indicate that the temporal disparity signal might be sparsely sampled and even more aggressively low-pass filtered, without causing visible depth differences. In Sec. 5.1, we conduct similar experiments for natural scenes involving monocular cues.

Surprisingly, depth edges appear sharp, even though human ability to resolve them in space and time is low. One explanation for this is that the perceived depth edge location is determined mostly by the position of the corresponding luminance edge [Robinson and MacLeod 2013].

In previous work, perception was taken into account for stereography when disparity is given [Didyk et al. 2012], but it was routinely ignored when inferring disparity from monocular input for 2D-to-3D conversion. Leimkühler et al. [2016] describe how information present in a single monocular luminance image alone can be used to reconstruct depth in a way similar to how HVS is believed to work. Interestingly, depth discontinuities that are not accompanied by luminance edges of sufficient contrast poorly contribute to the depth perception and do not require precise reconstruction in stereo 3D rendering [Didyk et al. 2012].

Stereoacuity in scotopic conditions In a vast majority of graphics techniques dealing with stereo vision and binocular disparity processing, photopic vision is tacitly assumed [Lang et al. 2010, Didyk et al. 2011]. Lit [1959] reports an almost 20-fold disparity threshold increase in scotopic conditions with respect to photopic ones. While the threshold steadily increases across mid-photopic and mesopic conditions with decreasing adaptation luminance, at the transition between mesopic and scotopic vision (where cones become inactive) a clear discontinuity in stereoacuity can be observed [Lit 1959, Lit and Hamm 1966]. Disparity estimation in the human visual system (HVS) is similar to windowed cross-correlation [Cormack et al. 1991, Filippini and Banks 2009], which is commonly used in computer vision. In this respect the difficulties in finding window correspondence for night photographs [Subr et al. 2012], can serve as an indicator of possible problems that the HVS might experience in dark conditions. Such problems include the reduced quality of eye optics due to a larger pupil size, more blur due to loss of central cone vision in the fovea and reduced acuity of peripheral rod vision, as well as neural circuitry and photon noise [Hess et al. 1990].

In our further discussion we focus on scotopic stereoacuity dependence on spatial frequencies and luminance contrast in the image, as well as issues of the comfortable disparity range.

Spatial frequency content Banks et al. [2004] have investigated random dot stereograms of various density and they showed that spatial stereoresolution is limited by the Nyquist sampling limit at low dot densities and the spatial frequency of luminance

patterns at high densities. Moreover, stereoresolution deteriorates with increasing retinal eccentricity, which is not attributable to a binocular vision deficit in the periphery, but rather the increasing size of receptive fields and low-pass filtering in the eye optics. Livingstone and Hubel [1994] have investigated line stereograms and noticed that in scotopic conditions stereoacuity is more reduced than Vernier acuity for relative line position, which are both examples of hyperacuity. This may suggest that cortical averaging mechanisms associated with hyperacuity are more complicated in the case of stereo vision and more sensitive for the poorer quality of rod input, where higher spatial frequency patterns are strongly suppressed [Shlaer 1937].

Luminance contrast As discussed by Legge and Gu [1989] and Heckmann and Schor [1989], binocular disparity sensitivity strongly depends on the magnitude of luminance contrast. Didyk et al. [2012] considered this effect in the context of disparity manipulation using a stereoacuity model proposed by Cormack et al. [1991]. The model has been derived based on measurements for narrow-band-filtered random dot stereograms, and predicts an over tenfold disparity threshold increase for luminance contrast change from suprathreshold (over 10 just-noticeable-differences (JND)) to near threshold levels (2 JNDs). Since the model has been obtained through averaging the results for randomly chosen photopic and upper range mesopic luminance levels akin to a CRT display, the model validity for scotopic conditions is to be questioned. In Chapter 9, we investigate this issue. Since the sensitivity to luminance contrast is more than tenfold reduced for the change of adaptation luminance from photopic to scotopic conditions [Wandell 1995, Fig. 7.21], our goal is to model the resulting reduction of stereoacuity.

Disparity range Another interesting finding on scotopic stereoacuity is a fourfold enlargement of Panum's area of binocular fusion with respect to photopic conditions. This is well beyond fusion limits that can be predicted from the well-known spatial frequency and contrast effects [O'Shea et al. 1994]. This means that for night scenes reproduced on high dynamic range (HDR) displays, disparity compression, which is often performed to reduce visual discomfort, can be relaxed. This does not apply to low dynamic range (LDR) displays, which are considered in Chapter 9, as the viewer luminance adaptation is mostly at photopic levels.

Disparity of reflective and refractive surfaces Appearance of refractive and reflective surfaces changes with the viewpoint and, consequently, it is different for left and right eye. This creates a challenge for the HVS and its ability to extract depth information from disparity. It was suggested that a prior based on knowledge of the physical behavior is used to predict locations of reflected objects [Blake and Bülthoff 1990]. Murry et al. [2013] proposed that extensive vertical disparities as well as too large horizontal disparity gradients are used to estimate the reliability of a disparity signal. This allows HVS to discard disparity information where it would lead to a false perception and switch to an alternative source of estimate.

As a result of view dependent surface characteristic, multiple layers of disparity can be distinguished by a viewer. The performance in such task was observed to be lower for layer separation below 2 arcmin, too dense luminance patterns, or too large disparities [Weinshall 1989, Tsirlin et al. 2008].

Chapter 3

Previous work

In this section we provide overview of computer graphics application that are relevant to our work. We focus on different approaches for processing of disparity signal in stereoscopic images and videos. We also provide overview of methods for adaptive image processing based on gaze information. Additionally, we give brief summary of methods for HDR retargeting.

3.1 Disparity processing

3.1.1 Disparity mapping

Disparity range compression is one of the most common disparity manipulations [Shibata et al. 2011, Zilly et al. 2011, Hoffman et al. 2008, Lang et al. 2010, Didyk et al. 2011] employed to avoid the accommodation-vergence conflict. Since this task shares many similarities with luminance range compression, standard tone mapping operators [Reinhard et al. 2010] can easily be adapted for disparity processing [Lang et al. 2010]. However, special care should be taken to avoid depth reversals when using local operators which for luminance could be explicitly used to expand the local contrast. Didyk et al. [2011] proposed a perceptual model for disparity that mimics the disparity processing done by the HVS and applies this for disparity manipulations. In the context of automultiscopic displays, the problem of extreme depth compression was addressed [Didyk et al. 2012, Masia et al. 2013a, Chapiro et al. 2014]. These techniques aim at fitting disparities into a very shallow range taking care that the crucial disparity details are preserved. Such techniques can be also driven by additional saliency information. In the context of real-time solutions, simple but efficient methods include baseline and convergence manipulations [Jones et al. 2001, Oskam et al. 2011]. The common feature of all these techniques is maintaining good image quality in all regions regardless of the observer's current gaze direction. In Chapter 6, we go beyond that and try to improve perceived quality based on the available gaze information.

3.1.2 Temporal considerations

Typically disparity is manipulated in individual frames, and temporal processing is limited to “smoothing” between different disparity ranges at scene transitions [Lang et al. 2010, Yan et al. 2013]. Also, it is ensured that for moving scene elements local warping distortions are propagated along the motion flow to the successive frames [Lang et al. 2010]. A limited temporal extent of per-frame temporal smoothing and low-pass filtering characteristic of first order smoothing terms used do not allow to

maintain high-frequency temporal features of disparity dynamic for complex motions. In Chapter 5, we go beyond such local smoothing and enable explicit global control of disparity changes over time. This way we can preserve both spatially local disparity manipulations and temporarily global disparity dynamics.

In video retargeting applications rigidity in temporally salient image regions is often enforced [Krähenbühl et al. 2009, Wang et al. 2010]. This can be performed in disparity-driven image warping as well [Lang et al. 2010], but here the goal is to avoid geometric deformation of moving objects, which are strong gaze attractors, rather than to prevent deformations of perceived motion trajectory. Wang et al. [2010] used a second-order smoothing term to minimize creation of “virtual” camera motion. Hoffman et al. [2011] analyzed the impact of image refresh rate and stereo 3D display technology (precisely, the eye-view separation method) on the visibility of flicker, motion smoothness, and distortions in perceived depth. In Chapter 5, we use a second-order derivative as a descriptor of motion in depth dynamics and enforce its preservation during stereoscopic remapping which often involves compression of the disparity range. Our approach differs from the previous work as it does not try to produce smooth outputs but rather preserves the temporal details in the original signal which are important for the reproduction of motion in depth and tasks that depend on it (Sec. 5.3).

3.1.3 Viewing comfort enhancement

While many factors may impact visual comfort when looking at stereoscopic displays the conflict between the eye convergence and accommodation is usually identified as the most prominent one [Lambooj et al. 2009, Shibata et al. 2011]. Similar to other disparity manipulation techniques [Jones et al. 2001, Lang et al. 2010, Didyk et al. 2011, Koppal et al. 2011, Yan et al. 2013], our approach ensures that the comfortable range of binocular disparity is always maintained. Yano et al. [2002] and Speranza et al. [2006] investigated the impact of object motion on the visual comfort. They found that the rate of disparity changes over time, which is determined by the object velocity, may strongly affect the visual comfort. Also, frequent changes from crossed disparity with vergence point in front of the horopter to uncrossed disparity with vergence point behind the horopter may significantly reduce the comfort.

A relatively little attention has been given to issues connected with rivalry and the lack of fusion. Yang et al. [2012] utilize rivalry to achieve a vivid tone mapping of 2D images through a dichoptic presentation. Highlights and their view dependent appearance has been considered as a source of rivalry by Templin et al. [2012] and a method improving the fusion has been proposed. For a more general light transport including a combination of reflections and refractions no relevant previous work exist. Such images will contain multiple layers at various depths and their individual appearances as well as their separations have to be taken into account. In Chapter 8, we propose one possible solution to this problem.

Stereo 3D typically evokes higher positive emotions and stronger feeling of immersion in games compared to the 2D mode and often leads to a better accuracy of performed tasks, especially those that involve spatial 3D interaction [Kulshreshth et al. 2012]. For example, Hubona et al. [1999] have found that stereoscopic viewing improves both precision and speed in the object positioning and resizing tasks, while object shadows are far less effective cues. In VR applications, which involve self-motion based on optical flow, binocular 3D information facilitates the judgement of moving object direction through the flow parsing in the HVS into the self-motion and object motion components [Matsumiya and Ando 2009]. All these applications involve the

movement of objects in some form, and potentially can benefit from our motion in depth optimization described in Chapter 5.

3.2 Gaze-driven applications

In Chapter 6, we will introduce gaze-driven method for enhancement of stereoscopic content. Here we discuss similar previous work as well as other gaze-driven applications in computer graphics.

3.2.1 Gaze-driven disparity manipulation

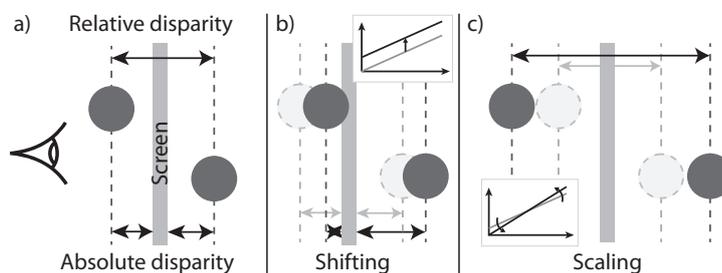


Figure 3.1. (a) Two objects as displayed in depth on a stereoscopic screen with their absolute disparity from the screen and the relative disparity between them. (b) Shifting of disparity moves both objects jointly, and thus changes absolute but preserves relative disparities. (c) Scaling of disparity changes mutual distances between objects and therefore both absolute and relative disparities.

While the existing disparity manipulation techniques summarized in Sec. 3.1.1 are successful on their own, their immediate adaptation for gaze-contingent setups is not an obvious task. Relatively few efforts addressing this problem almost exclusively focus on shifting the whole disparity range with respect to the display plane (Fig. 3.1b).

Fisker et al. [2013] investigate the gaze-driven disparity adjustment towards the screen plane in an informal experiment and report promising results in terms of visual comfort and perceived depth. Bernhard et al. [2014] perform a full-scale experiment where an abrupt disparity adjustment to the screen plane is compared to the static disparity case in terms of the vergence time. In this case, shorter timings are desirable for reducing the viewing fatigue. They find that the abrupt disparity adjustment often increases the vergence time, and suggest that stereo fusion might require some readjustments in such conditions. This might be because the vergence facilitation effects due to peripheral vision processing is invalidated [Cisarik and Harwerth 2005]. It is worth noting that in that experiment, the actual saccade is triggered by the color change of a test square, whose destination is precisely known, and the disparity adjustment can be performed with a minimal latency. In practical applications, the problems reported by Bernhard et al. can be aggravated, and the visibility of the disparity change is more difficult to hide. For this reason, all disparity manipulations that we propose are performed in a seamless manner so that the stereo fusion is never disrupted and all latency issues are irrelevant.

The work of Peli et al. [2001] is conceptually close to our idea presented in Chapter 6, and the authors measure the probability of detecting motion in depth for the fixated object

when its disparity is shifted to zero at various speeds. The experimental method, however, limits the free exploration of the scene by the observer who is directly instructed to look at particular target locations. A similar setup with a virtual hand as a target controlled by a hand-tracking device is explored for virtual reality applications with head mounted displays [Sherstyuk et al. 2012]. Chamaret et al. [2010] uses a visual attention model to predict the new region-of-interest (RoI) to gradually reduce its disparity to zero. They experimentally derive the maximum disparity change that remains unnoticeable as 1.5 pixel steps. Hanhart and Ebrahimi [2014] extend this work by employing an eye tracker for determining the RoI. They assume a disparity change of 1 pixel per frame without the frame-rate notion, and obtain favorable user judgments of such disparity manipulations when compared to the static disparity case.

In Chapter 6, we extend the work of Peli et al. [2001] by systematically measuring the just-noticeable speed of disparity changes for different initial disparity values. Our measurements are performed for continuous disparity shifts (Fig. 3.1b) rather than the discrete steps where the accumulated effect of disparity manipulation is not considered [Chamaret et al. 2010]. For the first time, we perform similar measurements for changes in the disparity range (Fig. 3.1c), and build a model that integrates both scaling and shifting of disparities. Based on the model we propose a novel gaze-contingent disparity mapping that enables seamless and continuous disparity manipulations, which significantly enhances the perceived depth.

3.2.2 Other gaze-driven applications

Gaze location tracking has also been used in other applications. In foveated rendering [Guenter et al. 2012], the efficiency of image generation can be improved by maintaining high image resolution only around the gaze location. The authors reported that seamless rendering can be achieved with a latency below 40 ms. In a similar way, chrominance complexity [Liu and Hua 2008] and level of detail [Murphy and Duchowski 2001] can also be gradually degraded with distance from the gaze location. Besides improving the rendering performance, the gaze location has been used to improve the image quality and the viewer experience. In the context of tone mapping, the luminance range can be used more effectively by reducing image contrast in regions that correspond to the peripheral vision [Jacobs et al. 2015]. Gaze-contingent depth-of-field effects have been modeled to improve the rendering realism [Mantiuk et al. 2011a], reduce the vergence-accommodation conflict [Duchowski et al. 2014b], or enhance the depth impression [Vinnikov and Allison 2014]. Although all these techniques lead to either reduced costs in rendering or a better image reproduction, they often require the frame update to be strictly within the saccadic suppression period. In all cases the users express dissatisfaction if there is a noticeable lag due to the insufficient performance of the eye tracker or the rendering.

All these practical results clearly indicate that the use of saccadic suppression to hide the content change for the new fixation is very sensitive to the type of performed changes, the eye tracking precision, and the overall system latency. In gaze-contingent disparity manipulations, abrupt depth changes must be completed within the saccadic suppression, as new sensory information acquired afterwards actually guides the eye vergence motion (Sec. 2.3.2). On the other hand, the saccade must be effectively completed for a precise determination of the target depth, which leaves little room for fully informed scene depth manipulation. While there exist methods for predicting the saccade landing position based on some initial saccade direction and velocity measurements, e.g., using a ballistic model [Komogortsev and Khan 2008], any inaccuracies in this respect could

be a serious hindrance for any gaze-driven disparity manipulation effort. For all those reasons, Chapter 6 advocates seamless depth manipulations, when the new fixation is established.

3.3 HDR processing

HDR tone mapping is a broad topic and its full extent is beyond scope of its thesis. We refer to an overview of existing methods by Eilertsen et al. [2013] or recent work in the field [Aydin et al. 2014, Eilertsen et al. 2015]. Here we focus tone-mapping solutions for night scenes as this is important for our methods proposed in Chapter 9 and Chapter 10. In particular, we describe the role of perceived noise in scotopic vision and, in this context, we overview other sources of noise in images such as sensor noise, film grain, and synthetically generated noise, which under certain conditions can improve perceived realism and quality. Finally, we discuss photon-accurate eye modeling as is required near absolute threshold, which is central for Chapter 10.

3.3.1 Tone mapping: Night scene depiction

A key goal of tone mapping operators (TMO) is to reproduce scotopic scene appearance on a photopic display [Reinhard et al. 2010] by simulating a blue-shift and the loss of color vision, visual acuity, contrast and brightness characteristic for night vision [Ferwerda et al. 1996, Pattanaik et al. 1998, Khan and Pattanaik 2004, Kirk and O'Brien 2011, Wanat and Mantiuk 2014]. Typically, such simulations cover higher levels of scotopic luminance (0.001–0.1 cd/m²) including the transition to mesopic conditions, while luminance levels near absolute thresholds are not specifically addressed. Furthermore, the time-course of adaptation [Durand and Dorsey 2000, Pattanaik et al. 2000], the bleaching phenomenon [Gutierrez et al. 2005], or stereovision in darkness [Kellnhofer et al. 2014b] were modeled in computer graphics.

Nightly impressions have been convincingly reproduced in painting [Livingstone 2002], digital arts, computer games, and feature films without referring to any rigorous simulation of scotopic vision. Empirical solutions inspired by “day-for-night” shooting have been proposed by Thompson et al. [2002]. The success of empirical techniques indicates that rigorous simulations of scotopic vision not always lead to a subjectively optimal night-like look, especially in photopic conditions. Consequently, our strategy is to apply psychophysical data when available, and otherwise refer to empirical techniques, including the case when such data does not generalize to images presented on photopic displays.

Most importantly, Thompson et al. [2002] observed that adding noise to day-for-night tone mapping can improve the scotopic impression. They add static, zero-mean, uncorrelated Gaussian noise with a fixed standard deviation to each pixel, to achieve subjectively optimal visual results. Still, it is not clear how to apply their approach for animated content where calibration in absolute luminance is crucial such as close to absolute thresholds. An example of a video showing a transition (Fig. 10.1, left to right) from photopic over mesopic conditions down to scotopic conditions near the absolute threshold illustrates the two remaining main challenges: First, a transition from a noise-free image over subtle noise to a state in which only grey noise is expected to remain. To this extent, we introduce a calibration by absolute luminance not available from previous work. Second, changing image content, e.g., a simple camera pan, will require the noise to change. A simple overlay would result in a “shower door effect”

[Kass and Pesare 2011]. In Chapter 10, we model accurate change dynamics, principled by physiological data to feature additive and multiplicative components.

3.3.2 Image noise

Noisy images are often undesirable in image synthesis and imaging applications, where denoising techniques are common. However, noise can be explicitly added to enhance perceived image sharpness [Johnson and Fairchild 2000]. Fairchild and Johnson [Fairchild and Johnson 2005] have hypothesized that noise as a repetitive pattern tends to be suppressed by the HVS, which might subjectively enhance image saliency. In general, procedural noise is often used in image synthesis to enhance the visual richness of rendered images [Lagae et al. 2010].

Sensors in digital cameras are prone to different temporal and spatial sources of noise [Janesick 2001]. In particular, temporal photon and dark current shot noise show similarities to the nature of noise in rods (refer to Sec. 10.1) and are also modeled via Poisson distributions. Readout noise could be considered as an analog of retinal circuitry processing beyond rods and cones, which we ignore in our visual noise simulation described in Chapter 10.

The exposure and development of silver-halide crystals dispersed in the emulsion of analog films results in forming tiny blobs of metallic silver or dye clouds, which creates the familiar film grain effect whose density tends to follow a Gaussian distribution [Altman 1977]. Film grain as a form of noise might be manipulated or even intentionally added for technical or artistic reasons in movie post-production [Seymour 2011] and can be acquired through film-stock scanning or synthesized following its noise-power spectrum [Stephenson and Saunders 2007, Gomila et al. 2013]. Stereoscopic processing of film grain was recently described by Templin et al. [2014b]. Similarly, in digital photography, the suppression of a synthetic look and masking of digital sensor noise are desirable [Kurihara et al. 2008]. Simulating the complete photographic process, including important characteristics of photographic materials, such as its response to radiant energy, spectral sensitivity, emulsion resolution, and graininess [Geigel and Musgrave 1997] can render results very realistic. While some analogies to our work in Chapter 10 are apparent, extremely low light levels are not supported in [Geigel and Musgrave 1997] as it requires photon-accurate simulation (Sec. 2.5.2).

Part I

Stereoscopic 3D and motion

Chapter 4

Optimizing disparity for screen plane motion

Stereoscopic 3D imaging is nowadays a wide-spread and affordable means to achieve a convincing game or movie experience. The human visual system (HVS) uses a combination of different perceptual cues to estimate spatial layout from 3D images. Different from common luminance imaging, stereo 3D display technology provides additional binocular disparity cues. An extensive body of work has investigated various static properties of binocular stereo content, such as manipulation and editing to achieve viewing comfort both in terms of technical requirements as well as in faithful perceptual modeling. In this chapter, we describe the interplay of binocular stereo motion with different display and rendering technologies. Temporal disparity changes can introduce conflicts with other cues that hamper scene understanding, e.g. occlusion if apparent depth order is altered. This increases the difficulty of depth-oriented tasks in simulations and games where the HVS combines various depth cues to estimate the spatial layout of objects and their motion. We find that motion can have a strong influence on perceived depth, especially in connection with limitations of display devices in everyday use. Such display devices vary in several key properties:

- **Spatial resolution.** Displays of high spatial resolution can present smaller changes of motion, be it in the screen plane or in depth, resulting in smoother motion.
- **Temporal resolution.** The display can repeat frames if its refresh frequency is higher than the one of the image data source.
- **Temporal protocol.** The left and right eye images can be presented simultaneously or sequentially as well as it can be presented continuously (hold-type LCD) or in flashes (cinema projector).
- **Multiplexing.** Polarization, color coding or parallax barriers are commonly used to separate images between the left and right eye if presented simultaneously.

In physical reality the viewing angle which is different for each eye guarantees perception of a pair of stereo images. Conventional displays show the same content to both eyes and therefore cannot reproduce binocular disparity. Therefore “multiplexing” has to be added between the display and the viewer. Such multiplexing is most commonly done using, either

- **Color.** Color of left and right eye image is modified before presentation and color filters in glasses then separate two images.

- **Polarization.** Polarizer layer on the display changes polarity of emitted light for the left and right eye, and polarization filter in glasses then isolates the proper signal for each eye.
- **Parallax barrier and lenslet arrays.** Opaque barriers or tiny lens on the display ensures that individual pixels are only visible to one or the other eye.
- **Time-sequential presentation.** Left and right eye images are presented sequentially and glasses with active shutter are used to block inactive eye. Passive polarized glasses can alternatively be used to bring the active element of polarizer to the displaying device itself.

We investigated how these properties interact with each other in real display devices and how the perception of stereo 3D is altered. We conclude with recommendations for stereo 3D content optimization for some specific display technologies. We suggest the following improvements:

- Compensation of false motion in depth for anaglyph display (Sec. 4.1).
- Compensation of false motion in depth for time-sequential displays (Sec. 4.2).

We believe that such problems are yet researched by the display and computer graphic community, and addressing them will improve the viewing experience for dynamic 3D content at only a small computational cost and implementation effort.

4.1 Correction for the Pulfrich effect

In this section, we describe an approach for compensation of false motion-in-depth for the anaglyph glasses.

Type	Luminance [cd/m^2]			Attenuation [log]		
	No filter	Left	Right	N/L	N/R	R/L
Red-cyan	218.10	21.82	76.64	1.00	0.45	0.54
Green-magenta	218.10	34.86	28.32	0.80	0.89	-0.09
Amber-blue	218.10	15.35	1.567	1.15	2.14	-1.0

Figure 4.1. Our measurements of anaglyph glasses filter attenuations in log units. Common red-cyan glasses, Trioscopics green-magenta glasses and ColorCode 3-D amber-blue glasses.

We measured attenuation for several types of anaglyph glasses (Fig. 2.2c) using luminance meter Minolta LS-100 (See Tbl. 4.1). For most widely spread red-cyan glasses we got relative attenuation of the left eye as 0.55 log units. This is enough to create a delay of 5 to 10 ms according to Howard, Section 23.1[Howard and Rogers 2012]. Even ColorCode amber-blue glasses which are well-known for good reproduction of chromacity show similar magnitude of left- versus right-eye attenuation, but with opposite sign. Therefore producing the opposite shift in depth. Some glasses do not show large differences between eyes such as the green-magenta combination in our tests.

We measured the impact of the Pulfrich effect on the stereo perception in a perceptual experiment. We asked 5 subjects with tested stereo vision to watch a textured

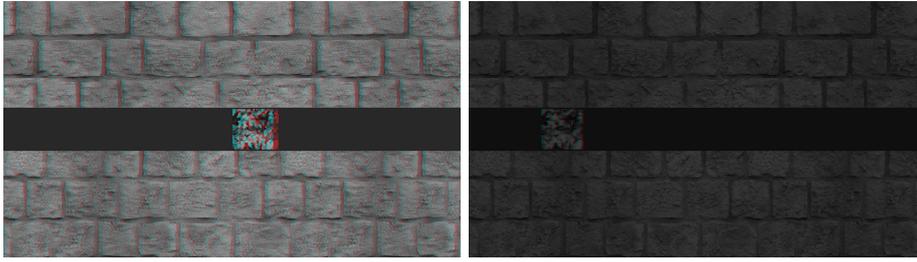


Figure 4.2. Experimental measurement of Pulfrich effect on red-cyan anaglyph glasses. User adjusts the delay of the right eye with lighter filter to negate introduced motion in depth. Two different luminance conditions shown.

square target stimulus moving horizontally in a sinusoidal motion (Fig. 4.2) on a Samsung SyncMaster 2233RZ 120 Hz LCD with the luminance range from 20 cd/m^2 to 200 cd/m^2 under normal office lighting conditions. The motion pattern covered viewing angle of 33 degrees. The stimulus depth was constant and in the plane of surrounding texture. Due to the Pulfrich effect, the target appeared to move in depth while moving left and right. Participants were instructed to adjusted the setup by pressing the “left” or “right” key, until the stimulus remains stable in depth. We experimented with 4 different stimulus motion frequencies varying from 0.3 to 1.2 Hz, 3 different disparity magnitudes (crossed, zero and uncrossed) and 2 different brightness levels within the range of the display. The theoretical model did not predict dependency on any of these attributes.

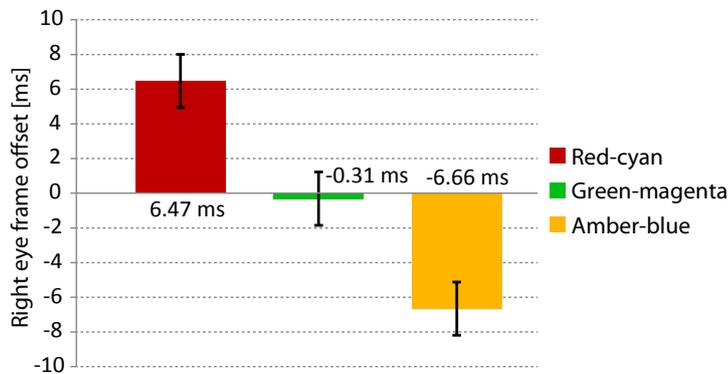


Figure 4.3. Time offsets in milliseconds introduced as delays for right eye frames to compensate for Pulfrich effect. Error bars denote confidence intervals for $p = 0.01$ according to Tukey HDS test.

The graph in Fig. 4.3 shows average time offsets for the right eye necessary to compensate each glass type. We have run one-way ANOVA with Tukey HDS post-hoc test and found that all pairs of mean values are different with $F(2, 102) = 81.06$, $p < 0.01$. The mean offset for red-cyan glasses 6.5 ms lies in the interval predicted by the physical measurement of filter densities.

As predicted, there was no significant difference detected among other experiment attributes. That can be due to insufficient extent of the study. Larger study would be required to prove that differences of means are negligible.

Our results confirm that the magnitude of the Pulfrich effect strongly depends on the type of glasses technology. Our website¹ provides a simple HTML5 applet to measure your own equipment and visual system in a web browser. We recommend applications using stereo to allow for adjustment of a delay of one eye. The delay can be either provided for most common anaglyph technologies or tuned by user using interface similar to our applet. Similar procedure is common for the contrast adjustment in graphical applications, e.g., computer games. Our measurements provide useful compensation values for some examples of a 3D equipment.

4.2 Correction for time-sequential presentation

In this section, we devise an approach to switch between alternative stereo capture and presentation protocols, based on a novel content-dependent prediction of eye pursuit motion probability.

A common way how to display 3D content on conventional displays is based on time-sequential presentation of the left and right eye image. Active shutter glasses are used to occlude the eye that is currently not required. The resulting ALT presentation protocol raises depth perception problems if frame repetition is used as discussed in the previous section.



Figure 4.4.  Representative examples of stimuli used in our experiment in Sec. 4.2. *a)* Rotating stimuli used to reconstruct experiment of Hoffman et al. [2011]. *b)* Periodic stimuli where eye pursuit motion fails at higher angular velocities. *c)* The same stimuli with an additional feature that improves eye pursuit motion. *d)* Stimuli with an easily trackable picture of a car.

We reproduced the fourth experiment of Hoffman et al. [2011] where depth stability was measured with either the ALT or SIM capture protocol for the ALT presentation protocol with frame repetition. We used green-magenta anaglyph glasses which were shown to have minimal Pulfrich effect to simulate the protocol on our 120 Hz LCD display. Our observations led to the same conclusion when applied to a rotating circle stimulus (Fig. 4.4a) as described. However, different conclusions have to be drawn from observations made when using more complex, 3D-rendered stimuli as found in interactive applications such as computer games. For example, we introduced periodical horizontal motion in the scene and used the ALT presentation of the left and right anaglyph image to simulate time-sequential display. We always compared the measured multi-flash protocol with a ground-truth SIM capture and presentation where no depth distortions are expected [Hoffman et al. 2011]. We then studied the relative motion-in-depth between the reference and multi-flash stimuli shown on the same display.

For slow motion our observations matched those of Hoffman et al. [2011] When the motion speed increased, the effect started to vary between images and some became unstable in depth for the ALT capture protocol. For other stimuli however, perception remained stable relative to a reference image and followed the model for slow speeds

¹<http://resources.mpi-inf.mpg.de/TemporalStereo#spie2014>

even at high speeds. We found, that the model for slow speeds is not valid for images with highly periodic texture patterns without significant features such as mosaics or rocks (Fig. 4.4b). For photographs of cars or people (Fig. 4.4d), the slow speed model was followed even at high speeds. We suspect that the reason for this difference is the inability to correctly pursue moving objects with smooth pursuit eye motion, when no visually significant and unique features can be distinguished on the object. Therefore, the measurement done for periodic textures were actually done without eye pursuit motion even though subjects were instructed to pursue the moving objects.

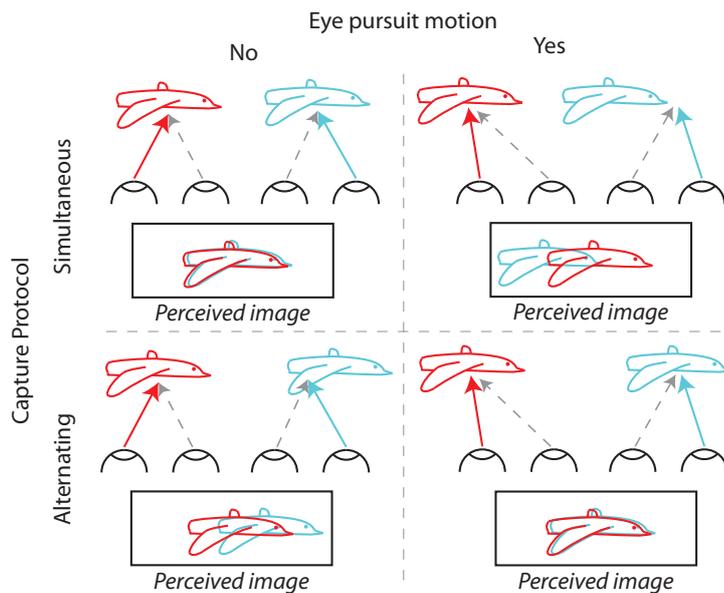


Figure 4.5. Perception of flying bird with zero disparity presented using ALT protocol without frame repetition. Red and cyan colors denote the eye active in the frame.

Fig. 4.5 explains the geometry of different presentation protocols and different eye pursuit motion conditions on perceived disparity. It shows that when assuming perfect estimation of smooth eye pursuit the temporarily accumulated disparity error is zero for alternating capture protocol with eye pursuit motion, or simultaneous capture protocol without eye pursuit motion. There is a false depth from disparity created with both other combinations.

To verify this assumption, another experiment was done using the same configuration and a periodic monochromatic image. We inserted a single unique feature into the pattern, a green colored circle, so that it became easier to pursue (Fig. 4.4c). We then found that the observation of apparent motion in depth became consistent with that for non-periodic images. We also did similar experiment with the original rotational setup. Here, we found difficulty to evaluate perceived motion in depth. We observed, that even though given point is properly pursued on its circular trajectory, it seems to move back and forth in depth with respect to the middle bar. However, the reason for this effect is found in the middle bar itself: When we instructed the participants to pursue the rotation motion, users lost their track with the static vertical bar. So it was the bar that exhibited screen space motion relative to the eye and therefore it moved in depth. As viewers were attracted to the moving circle, they did not observe its own motion and only saw different relative position at the transition phase of circle above the bar. This invoked

the illusion of motion in depth for the circle.

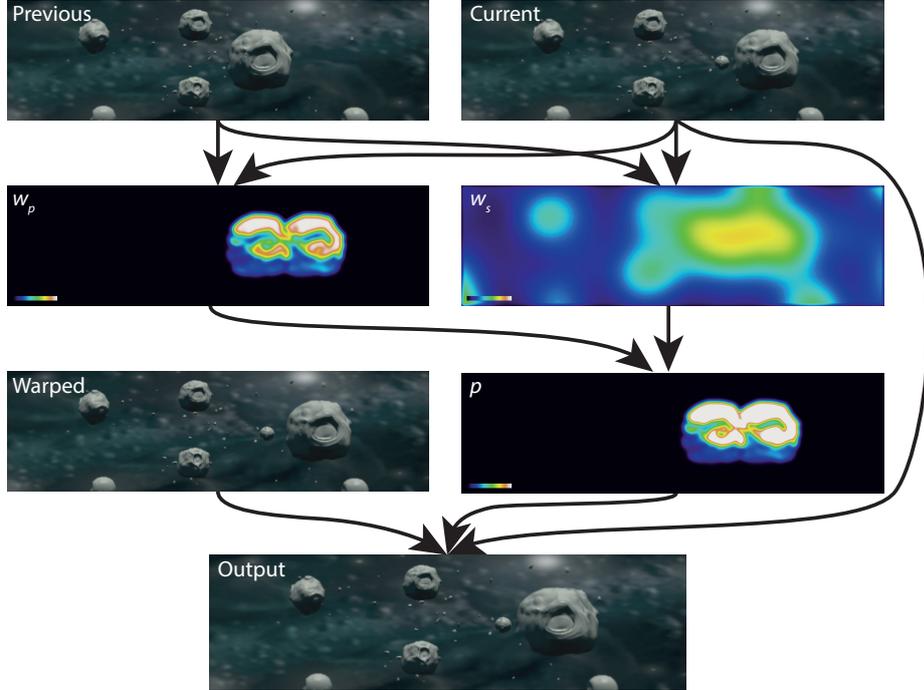


Figure 4.6. Diagram of our protocol for frame repeating ALT presentation. Previous and current left frames as inputs to our framework. Optimized left frame as output.

We therefore generalize the recommendation given by Hoffman et al. [2011] and conclude that the rendering approach should be chosen not only based on capture rate but also on the content and expected attention of the viewer. We suggest a spatially adaptive approach based on a novel eye pursuit motion probability map. The offset-compensated capture protocol is advised by default as it works well when no frame repetition is involved [Hoffman et al. 2011] and converges to the simultaneous capture protocol when motion speed goes to zero. Simultaneous rendering should only be chosen for moving regions that are not pursued by viewer. Simultaneous rendering is simulated using motion-flow based warping of the frame to allow for local transformations. The eye pursuit motion probability map is driven by local blending between both protocols maintaining spatio-temporal smoothness of the original sequence (Fig. 4.6). The proposed approach combines benefits of both protocols for different types of content. It is conservative as in the worst case, it produces results identical to the less suitable of them and not worse.

To justify the blending between two protocols we estimate the disparity means for each of two extreme cases. Fig. 4.8 shows how additional disparity in relative units denoted as dx is cumulated through six screen states of a triple-flash ALT/ALT protocol with eye pursuit motion in the way as was depicted for single flash protocols in Fig. 4.5. Each eye receives image composed from several images at different time. As a result of predicted motion and smooth eye pursuit, retinal projections of these images are shifted by dx_i . The final image can be considered as a low-pass filtered with mean value approximated by average of dx_i . Then the disparity is difference of means for left and

Cap. protocol	Eye purs.	Sequence		Mean		Disp.
		Left	Right	Left	Right	
Simultaneous	No	0, 2, 4	0, 2, 4	2	2	0
Simultaneous	Yes	0, 2, 4	1, 3, 5	2	3	-1
Alternating	No	0, 2, 4	1, 3, 5	2	2	0
Alternating	Yes	0, 2, 4	0, 2, 4	2	3	-1

Figure 4.7. The mean values of spatial offsets between the retina projection and expected motion trajectory and resulting false disparity for various capture protocols and eye pursuit motion assumptions. The relative units are multiples of product of spatial speed and frame duration.

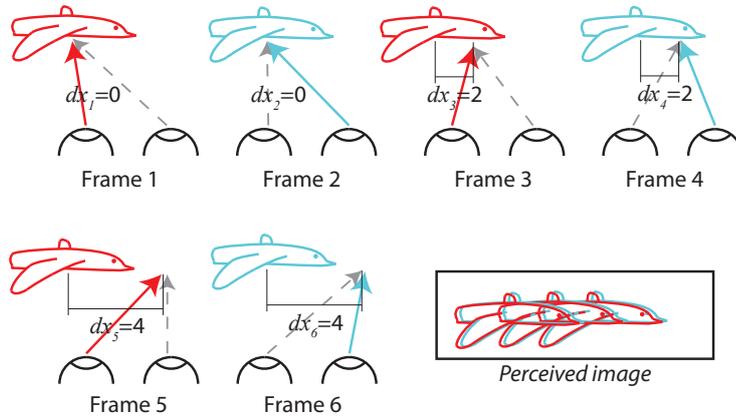


Figure 4.8. Cumulative disparity error during smooth pursuit of a flying target displayed using the ALT/ALT protocol with triple frame repetition. Red and cyan colors denote the left and right eye active in the respective frame.

right eye. In the case described in Fig. 4.8 we get a mean as 2 units for both eyes and therefore no additional disparity. Values for all combinations are summed in Tbl. 4.7. The difference sign denotes with the false additional disparity, which is observed in our experiments as a shift in depth.

For the method description to follow, we assume without loss of generality to display the left frame first. First the right frame is rendered at the simulation time and left frame is rendered at a time decreased by the delay of the right frame's first presentation (e.g., 8.3 ms for 120 Hz display). That matches the ALT/ALT protocol with an offset described by Hoffman et al. [2011]. The left frame is displayed without correction.

We compute an eye pursuit motion probability map for the right eye using a combination of dynamic saliency and similarity between frames. The saliency tells us if the user is motivated to pursuit given part of the image while the similarity tells us if user is likely to determine the motion flow and to perform the pursuit properly. We use the *Phase Spectrum of Quaternion Fourier Transformation* method [Guo et al. 2008] to find a spatio-temporal saliency map $w_s(\mathbf{x}_i)$. To detect the similarity between frames we compare matching samples of current and previous frame $f_t(\mathbf{x}_i)$ and $f_{t-1}(\mathbf{x}_i)$ and we get the periodicity weight $w_p(\mathbf{x}_i)$ as:

$$w_p(\mathbf{x}_i) = \min(\max(1 - 2|f_t(\mathbf{x}_i) - f_{t-1}(\mathbf{x}_i)|, 0), 1) \quad (4.1)$$

Finally, a Gaussian low-pass filter of radius 30 arcmin (typically 16 pixels) is applied to

achieve spatial smoothness. The final eye pursuit motion probability map then is

$$p(\mathbf{x}_i) = (1 - w_p(\mathbf{x}_i)) \cdot w_s(\mathbf{x}_i). \quad (4.2)$$

We produce the final image by local blending between the original frame and the warped frame. We may not want to introduce any compensation and possible artifacts into nearly static scenes where the default rendering protocol is sufficient. Therefore we take the motion speed into account. The largest speed where the slow-speed-model holds was measured by Hoffman et al. [2011] as

$$s = \frac{C}{2f} \quad (4.3)$$

where f is the number of frame repetitions and C is the temporal disparity-gradient limit of $|\Delta\delta/\Delta t|$ approximated as 950 arcmin/sec of change of viewing angle at the eye. Therefore we obtain limit speed s as 237.5 or 158.3 arcmin/sec for double or triple frame repetition. This way speed coefficient $w_c(\mathbf{x}_i)$ is derived:

$$w_c(\mathbf{x}_i) = \max\left(\frac{|\Delta\delta(\mathbf{x}_i)|}{s\Delta t}, 1\right) \quad (4.4)$$

where Δt is time difference between consequent frames and $\Delta\delta(\mathbf{x}_i)$ is the angular difference between the position of pixel \mathbf{x}_i in the previous and the current frame, approximated for a perpendicular viewed distant display as

$$\Delta\delta(\mathbf{x}_i) = \tan^{-1}\left(\frac{\|\mathbf{x}_i - A_f\mathbf{x}_i\| \cdot P}{d}\right) \quad (4.5)$$

where P is matrix size of pixel and d the screen distance. The permutation matrix A_f describes the local motion flow image f . Then we can derive the final blending weight $w(\mathbf{x}_i)$ as

$$w(\mathbf{x}_i) = \text{clamp}(2 \cdot w_c(\mathbf{x}_i) \cdot p(\mathbf{x}_i), 0, 1) \quad (4.6)$$

We warp the frame f using warping map B_f based on the motion flow A_f to approximate the frame that would be rendered by the SIM capture protocol:

$$B_f = \frac{1}{2f} \cdot A_f \quad (4.7)$$

The final frame is then generated by blending between the alternating and simulated simultaneous capture protocols:

$$\hat{f}_t(\mathbf{x}_i) = (1 - w(\mathbf{x}_i))f_t(\mathbf{x}_i) + w(\mathbf{x}_i)f_t(B_f\mathbf{x}_i) \quad (4.8)$$

We compared this proposed approach with a simple Gaussian blurring of the frame and with blending of in-between frame warping map B_f instead of blending of warped images. Fig. 4.9 shows application examples for simple periodic texture and rendered 3D scene. Simple blurring with symmetrical kernel does not change the mean value of disparity distribution and therefore was ineffective in improving depth stability. Blending of warping maps reproduced depth stability comparable to the proposed method but exhibited artifacts perceived as deformations. The proposed method also leads to visual artifacts which are perceived as double edges when observed statically, however such edges are blurred with previous frame repetitions in animation sequence which results in overall smother appearance than the ALT/ALT protocol.

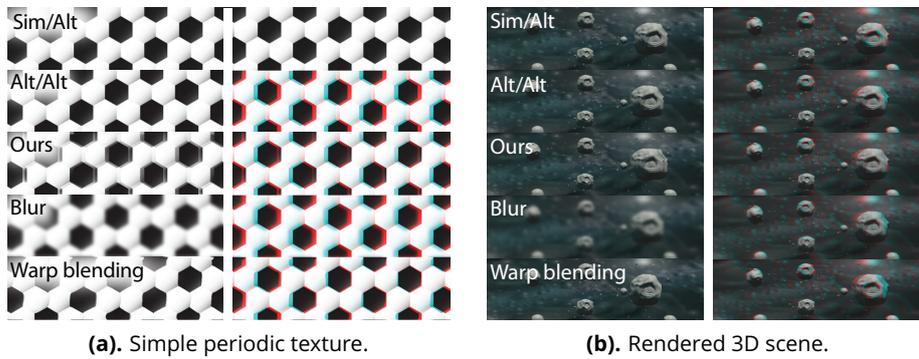


Figure 4.9. Left column of each example shows frames processed by our method. Right column shows overlaid left and one consequent right frame which simulates viewing without eye pursuit motion. The image was presented in the screen plane without additional disparity.

4.3 Conclusions

This chapter focused on two specific issues related to presenting a stereoscopic 3D content. It was shown that two of widely used stereoscopic display technologies are the subject of potential disparity distortion when a motion is introduced in the content.

First, the Pulfrich effect was analyzed for anaglyph glasses and then experimentally measured with several pieces of a consumer available 3D eye-wear. Our results lead to a suggestion of a temporal compensation for an anaglyph presented content.

Second, the left and right eye image capture and presentation protocols were discussed in the context of time sequential display technologies. Conditions leading to content dependent distortions of the disparity of moving objects were theoretically predicted and practically observed. A spatially adaptive saliency based approach was then proposed to combine advantages of two existing capture protocols and to minimize the disparity distortion for a dynamic 3D content on displays with a sequential presentation.

Both the topic of the motion in depth perception and the perception of depth in conditions of the motion have received a relatively less attention than the static 3D imaging in last years. A more comprehensive study and a wider choice of technologies should be included to create a complex recommendation for the 3D production and presentation. In future, we would like to focus on the interplay of disparity with other depth cues, such as texture, shading or scale. The stereo image capture processing is another challenging step.

Chapter 5

Optimizing disparity for motion in depth

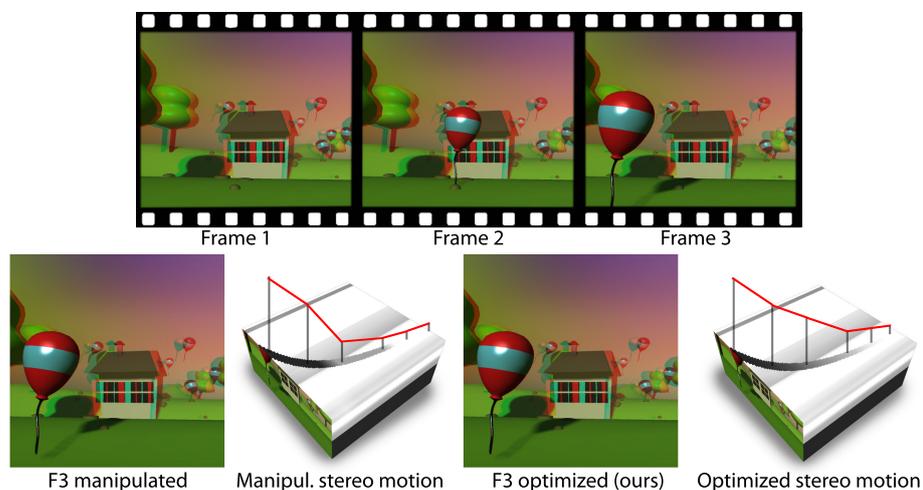


Figure 5.1.  Three frames of an animation showing a motion of a balloon in depth (*Left to right*). The outcome of disparity manipulation and our optimization, which reduces distortions in the balloon motion perception, are shown for frame *F3*. The cross-section through the space-time cube shows disparity as a function of time and space. When the balloon enters the zone between the house and the tree, the disparity manipulation results in abrupt disparity changes over time that are not present before the manipulation. Our optimization prevents the manipulation from distorting this originally smooth change.

Watching stereoscopic media such as movies and computer games can be an exciting experience, often described by statements such as “... and then this character really jumped *out* of the screen!” Remarkably many such statements refer to a *change* of perceived depth rather than a static condition. Until now, computational modeling of temporal changes of disparity (stereo motion) has only received little attention. Regrettably, manipulation of disparity, which is routinely performed to improve viewing comfort or to achieve artistic objectives, can impede perceived motion in depth. For example disparity compression in local scene regions may induce acceleration of motion-in-depth for objects traversing such regions. Such uncontrolled binocular cues can be perceived as annoying motion artifacts, which in navigation simulators or remote manipulators can affect the performance in precision-demanding tasks such as collision avoidance. In this chapter we show how to process binocular disparity to reproduce both faithful depth perception and stereo motion.

Let us consider the balloon approaching the viewer in Fig. 5.1 as an example. Here state-of-the-art solutions will improve viewing comfort by selective disparity compression in empty scene regions and preserve the balloon shape only at individual instants of time. If the manipulation causes compression or expansion of some depth regions with respect to others, the balloon will exhibit sudden change of speed i.e., acceleration, as it moves in depth from one such region to another. Temporal coherence of such disparity manipulation can be maintained by smoothing and propagating the resulting disparity changes along the motion flow, but the objective here is not to preserve the motion appearance fidelity per se. Moreover, since the disparity change is a strong cue for motion itself [Gray and Regan 1998], such uncontrolled disparity manipulation may introduce a cue conflict with respect to important pictorial cues such as the balloon size change due to perspective scaling. Clearly, a more holistic approach to the object stereo motion is required that accounts for local scene configurations as well. In this chapter we address this problem in the following five contributions:

- A perceptual analysis of stereo motion
- An experimental measurement of human sensitivity to both spatial and temporal disparity distortions
- An optimization to detect and preserve stereo motion cues
- A perceptual study of stereo motion task performance
- A 3D warping to create a stereo image pair with arbitrary, spatially-varying disparity from a polygonal 3D scene.

5.1 Experiment: Disparity distortion visibility

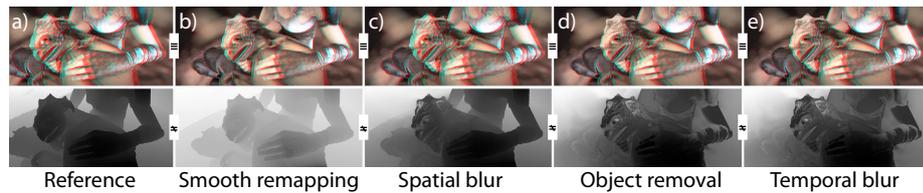


Figure 5.2. We intentionally introduce the depth distortions typically produced by 2D-to-3D conversion into close-to-natural computer generated images (*a*, *top*) such as the one from the MPI Sintel dataset [Butler et al. 2012] where ground truth depth is available (*a*, *bottom*). User response to stereo images (*b–e*, *top*) showing typical disparity distortions (*b–e*, *bottom*) gives an indication whether a certain amount of distortion results in functional equivalence for natural images or not. According to numerical measures such as PSNR or perceptual disparity metrics, the depth is considered very different (*inequality sign*, *bottom*), whereas it is functionally equivalent (*equivalence sign*, *top*).

Disparity that is input of our optimization can originate in many sources and can be affected by various distortions from acquisition, processing, compression or remapping. Before proposing a solution for removal of some of these distortions in an automatic way, we want to better understand how seriously do they affect human perception of depth. To this end, we conducted a series of perceptual experiments.

The majority of images and videos available is 2D, and automatic conversion to 3D is a long-standing challenge [Zhang et al. 2011]. A potential approach for such conversion that utilizes multiple monoscopic cues and fuses them together in a way similar to the HVS was also discussed in our recent paper [Leimkühler et al. 2016]. The requirements imposed on the precise meaning of “3D” might differ: For applications such as view synthesis, surveillance, autonomous driving, human body tracking, relighting or fabrication, accurate physical depth is mandatory. Obviously, binocular disparity can be computed from such accurate physical depth, allowing for the synthesis of a stereo image pair using image-based rendering. However, it is not clear what depth fidelity is required to produce plausible disparity in natural images, which include other monocular cues.

In this section we argue that physically accurate depth is not required to produce plausible disparity. Instead, we provide evidence that as long as four main properties of the disparity hold, it is perceived as plausible. First, the absolute scale of disparity is not relevant, and any reasonable smooth remapping [Jones et al. 2001, Lang et al. 2010, Didyk et al. 2012] is perceived equally plausible and may even be preferred in terms of viewing comfort and realism. Therefore, we can equally well use disparity that is the same as the physical one under a smooth remapping. Second, not every detail in the scene can be augmented with plausible depth information, resulting in isolated objects that remain 2D or lack disparity relative to their content. We will see that, unless those objects are large or salient, this defect often remains largely unnoticed. Third, the natural statistics of depth and luminance indicate that depth is typically spatially smooth, except at luminance discontinuities [Yang and Purves 2003, Merkle et al. 2009]. Therefore, not reproducing disparity details can be acceptable and is often not even perceived, except at luminance edges [Kane et al. 2014]. Fourth and finally, the temporal perception of disparity allows for a temporally coarse disparity map, as fine temporal variations of disparity are not perceivable [Howard and Rogers 2012, Kane et al. 2014]. Consequently, as long as the error is 2D-motion compensated [Shinya 1993], depth from one point in time can be used to replace depth at a different, nearby point in time.

We have conducted an experiment to find how typical disparity distortions that may originate in processing or in 2D-to-3D stereo conversion affect the plausibility of a stereo image or movie. To this end, we intentionally reduce physical disparity in one of four aspects and collect the users’ response.

5.1.1 Description

Stimuli Stimuli were distorted variants of a given stereo video content with known, undistorted disparity. We used four video sequences from the MPI Blender Sintel movie dataset [Butler et al. 2012], and the Big Buck Bunny movie by The Blender Foundation, which provide close-to-natural image statistics combined with ground-truth depth and optical flow. Additionally, we used four rendered stereo image sequences with particularly discontinuous motion that are especially susceptible to temporal filtering. Stimuli were presented as videos for temporal distortions and as static frames for spatial distortions to prevent threshold elevation by presence of motion that would underestimate the effect for a theoretical completely static scene. The scenes did not show any prominent specular areas that required special handling [Dąbala et al. 2014].

Distortions were performed in linear space with a normalized range of $(0, 1)$. For stereo display, this normalized depth was mapped to vergence angles corresponding to a depth range of $(57, 65)$ cm surrounding a display at 60 cm distance. This distribution around the display plane reduces the vergence-accommodation conflict, however, in

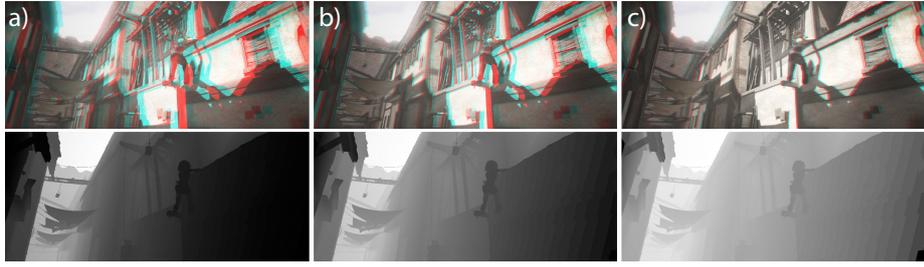


Figure 5.3.  Remapping. The top row shows the stereo image, the bottom row the disparity. The columns show different amounts of distortion due to remapping. *a)* Original stereo image. *b)* Remapping by a value of $\gamma = 0.8$, leading to equivalence. *c)* Remapping by a value of $\gamma = 0.6$, leading to non-equivalence.

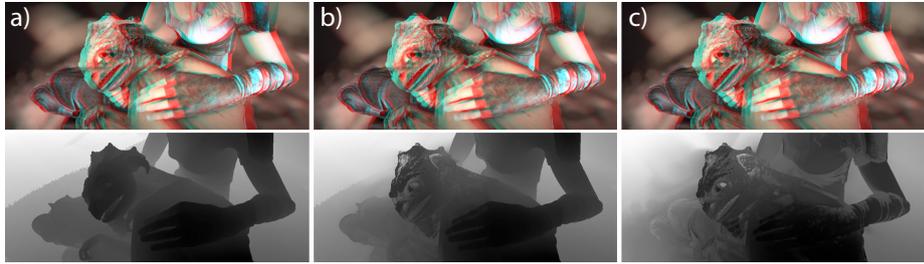


Figure 5.4.  Object removal. The top row shows the stereo image, the bottom row the disparity. The columns show different amounts of distortion due to object removal. *a)* Original stereo image. *b)* Removal of a circular region of radius $r_2 = 3$ vis. deg. around the character's head, leading to equivalence. *c)* Removal of a region of radius $r_2 = 6$ vis. deg. at the same location, leading to non-equivalence.

some cases a window violation occurred. As the salient content was located at the image center, the influence of this artifact likely was low. Finally depth-image-based rendering [Fehn 2004] was used to convert the monocular image with a distorted depth map into a stereo image pair. Stimuli were subject to exactly one out of five distortions: identity (placebo), remapping, object removal, spatial blur and temporal blur.

Placebo Firstly, the original content without any distortion is used as a control group. This is required to understand how often subjects report a difference when there is none, establishing an upper bound on what to expect if there really are distortions.

Remapping of the linear depth map D was performed for each location \mathbf{x} by means of a power curve $D'(\mathbf{x}) = D(\mathbf{x})^\gamma$ with a γ value of $r_1 \in \{0.9, 0.8, 0.6, 0.3, 0\}$. A value close to 1 indicates no change. Small values indicate a more compressive function. A value close to zero indicates a 2D stimulus with very little global disparity. We have chosen the power function as it is the most basic signal compression method that accounts for weaker abilities of depth discrimination with increasing depth by the human visual system. It is also inspired by the non-linear operator proposed by Lang et al. [2010]. Example stimuli are shown in Fig. 5.3.

Removal of entire regions was realized using luminance edge-aware inpainting from surrounding depth that restores structure but not disparity values. The regions removed were randomly positioned circles of radius $r_2 \in \{1, 3, 6, 12\}$ visual degrees. In practice, bilateral filtering with a strongly edge-preserving range radius parameter choice of 0.1 was used. The spatial radius of the filter was set to $0.5 \cdot r_2$ and values in the removed

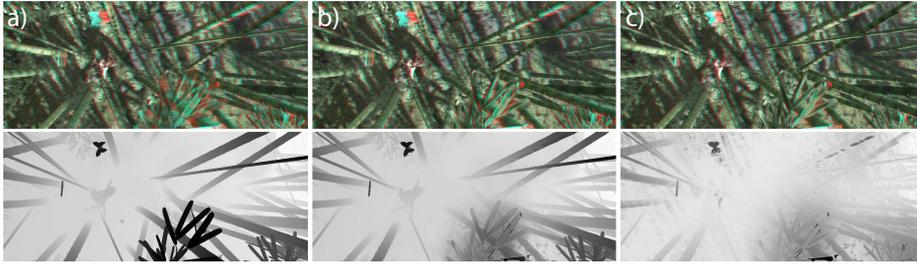


Figure 5.5.  Spatial blur. The top row shows the stereo image, the bottom row the disparity. The columns show different amounts of distortion due to spatial blur. *a)* Original stereo image. *b)* Blur with a spatial support of $r_{3,1} = 0.25$ vis. deg. and a range support of $r_{3,2} = 0.1$, leading to equivalence. *c)* Blur with a spatial support of $r_{3,1} = 2$ vis. deg. and a range support of $r_{3,2} = 0.1$, leading to non-equivalence.

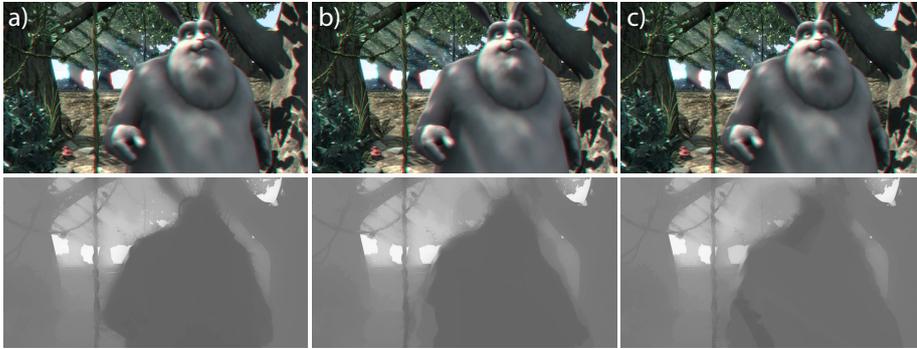


Figure 5.6.  Temporal blur. The top row shows the stereo image, the bottom row the disparity. The columns show different amounts of distortion due to temporal blur. *a)* Original stereo image. *b)* Blur with a temporal support of $r_4 = 0.25$ s, leading to equivalence. *c)* Blur with a temporal support of $r_4 = 1$ s, leading to non-equivalence.

region were weighted by zero in the center of the region and smoothly transitioned to 1 outside the region. This prevented visible discontinuity on the region boundary. The transition was generated by Gaussian blur of the binary mask of the region. Example stimuli are shown in Fig. 5.4.

Spatial blur was realized using bilateral filtering:

$$D'(\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \sum_{x_i \in \Omega} D(\mathbf{x}_i) G(\|\mathbf{x} - \mathbf{x}_i\|, r_{3,1}) G(D(\mathbf{x}) - D(\mathbf{x}_i), r_{3,2})$$

where $Z(x)$ is the normalizing partition function, Ω the spatial domain of D and $G(d, \sigma)$ is a zero-mean Gaussian distribution function with std. dev. σ chosen as $r_{3,1} \in \{0.25, 1, 2\}$ visual degrees for spatial range and $r_{3,2} \in \{0.1, 0.6, \infty\}$ for the intensity range from 0 to 1. The visual radius of 2 deg corresponds to ca. 80 px in our stimuli. Example stimuli are shown in Fig. 5.5.

Temporal blurring with a std. dev. of $r_4 \in \{0.025, 0.25, 1\}$ seconds was introduced. The blur was motion-compensated [Shinya 1993], that is, before combining pixels from a different frame, they were moved along their (known) optical flow. This assures, that temporal disparity details are removed for individual objects rather than blending the

disparity values of distinct moving objects in a dynamic scene. Example stimuli are shown in Fig. 5.6.

Subjects 17 participants took part in the experiment (23 ± 4 ys, 8M, 9F). Subjects were naïve with respect to the given task, 4 of them had a background in computer graphics or computer vision. All had corrected or corrected-to-normal vision. None reported any stereo vision deficiency and all were able to identify patterns and digits in test random dot stereograms.

Equipment Stimuli were shown using anaglyph on a DELL U2412M 60 Hz display with spatial resolution of 1920×1200 . As the videos are provided at 30 Hz, each frame was displayed twice. The magnitude of crosstalk was not measured. The combination of display, particular glasses and display settings were experimentally chosen so that ghosting was minimal and the same for every experimental condition. We argue that presence of minor ghosting is common in target consumer displays and therefore not violating the purpose of our study aiming to predict user experience in a practical scenario.

Procedure

In each trial, participants were shown the undistorted reference image and a distorted variant in a randomly shuffled vertical arrangement for 3 seconds and were asked to answer the question:

“Do both images provide equivalent stereo fidelity?”

by pressing one out of two keys on a keyboard. Asking for equivalence instead of preference removes the influence of a subjective bias for a particular disparity distribution which might not even favor the ground-truth in all cases. An example is edge-preserving filtering that typically results in an edge enhancement, which in turn might lead to overall preferred depth appearance by some subjects.

Each trial was followed by a screen with confirmation where the subject could take a rest. A blank screen was displayed for 500 ms immediately before the stimuli was shown. The experiment comprised of 2 repetitions for each of the 4 videos or images being presented with 1 placebo, 5 different remappings, 4 removals, 3×3 spatial blurs and 3 temporal blurs yielding the total of $2 \times 4 \times (1 + 5 + 4 + 3 \times 3 + 3) = 64$ trials, and lasted for approx. 40 minutes.

Analysis

We compute sample means and confidence intervals (binomial test, 95% confidence intervals (CIs), Clopper-Pearson method) for the percentage of trials in which a distorted and an original stimulus are considered equivalent (Fig. 5.7). The response is aggregated over all four scenes. Equivalence is rejected using two-sample *t*-testing (all $p < 0.01$). Additionally, the effect of reduction can be seen from comparing their CIs to the control group, in particular, its lower bound (Fig. 5.7, dotted line). CIs that do not intersect the placebo CI after the Clopper-Pearson correction indicate the presence of an effect.

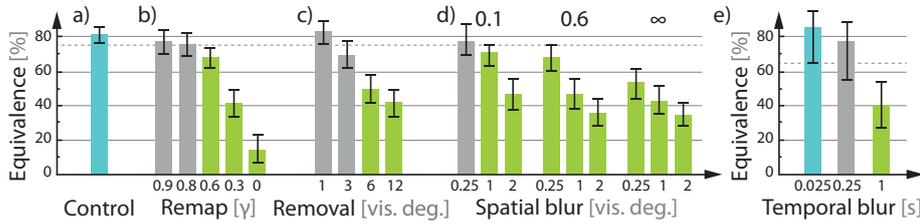


Figure 5.7. Perceptual experiment analysis (Sec. 5.1.1): The horizontal axis shows different bars for different distortions. The vertical axis is equivalence in percentage. A high value means that the distortion is more equivalent to a reference. A green bar has a significantly different equivalence compared to how equivalent the reference is to itself, which is only ca. 80 %, not 100 %. Bars are grouped by distortions. Inside each group the distortion is the same, just more or less strong in one (b,c and e) or two (d) respects. The outcome is discussed in Sec. 5.1.2.

5.1.2 Results

In this section, we discuss the outcome of the above experiment, compare this to observations made for artificial stimuli, compare our equivalence outcome to the prediction of established metrics, recall the scope and limitations of the experiment, and finally propose some recommendations for assessing disparity quality.

Observations

Placebo The control group, which is not distorted at all, is considered equivalent to the reference in $79.0\% \pm 4.0\%$ of the cases (Fig. 5.7, a). This indicates, that subjects roughly understand the task and do not give random answers. At the same time, it also shows the limits of what to expect from asking for equivalence: a fifth of the subjects reports seeing a difference when images are identical. Note, that this indicates, that a distortion that is equivalent will at best result in observing a score of ca. 80 %, not 100 %, which is not even achieved when no change at all is present.

Remapping Remapping values for $r_1 \leq 0.6$ (stronger deviation from identity, Fig. 5.3, c) are significantly nonequivalent (Fig. 5.7, b), indicating (but not proving) that more subtle remappings might be equivalent (Fig. 5.3, b). This is in agreement with the general practice of retargeting disparity values using smooth curves [Lang et al. 2010, Didyk et al. 2012] to better account for human perception on limited output devices.

Spatial blur For blurring (Fig. 5.7, d), not respecting edges ($r_{3,2} = \infty$), or edge-stopping blurring ($r_{3,2} = 0.6$ and $r_{3,2} = 0.1$) with a spatial Gaussian of std. dev. $r_{3,1} \geq 1$, resp. $r_{3,1} \geq 2$ vis. deg. is not equivalent (Fig. 5.5, c). This indicates that the slightly larger spatial extent and similar range support produce a functionally equivalent result (Fig. 5.5, b). It also highlights the importance of respecting luminance edges.

Object removal Not reproducing objects as large as $r_2 = 6$ vis. deg. (Fig. 5.4, c) or larger is significantly nonequivalent (Fig. 5.7, c), indicating that removal of smaller objects might not be objectionable (Fig. 5.4, b). As long as such objects are consistently embedded into the environment, which typically happens due to luminance-based edge-aware upsampling, the proper values of depth are not mandatory. This is in agreement

with common practice in 2D-to-3D stereo conversion, that does not manually label all objects with depth in a scene exhaustively.

Temporal blur For temporal blurring (Fig. 5.7, e) all reductions with a temporal Gaussian of std. dev. $r_4 \geq 1$ s have been found visually nonequivalent (Fig. 5.6, c). This indicates that temporal disparity sampling can be surprisingly sparse if it is motion-compensated [Shinya 1993], i.e., only disparity keyframes at ca. 3 Hz have to be fully recovered while the intermediate disparity frames can be temporarily interpolated (Fig. 5.6, b). Temporal upsampling (rotoscoping with keyframes) is a typical component of 2D-to-3D conversion and some other real time processing systems, producing imprecise but perceptually valid results.

Artificial and natural stimuli

Analysis of thresholds, i.e., what can be perceived, has helped to better understand how the human visual system perceives both luminance and stereo [Howard and Rogers 2012, Ch. 18.5]. For finding these, artificial stimuli such as sinusoidal gratings [Didyk et al. 2012] or step edges [Kane et al. 2014] in the absence of other cues are common. However, it is clear, that such thresholds are overly conservative and do not answer the question which two natural stimuli are functionally equivalent. For this reason, visual equivalence has been proposed for specialized natural luminance stimuli involving certain geometry, certain lighting and certain reflectance [Ramanarayanan et al. 2007].

For stereo, the outcome of the above experiment indicates that in natural images, even more edge-aware spatial blurring and temporal filtering is tolerated than what was reported for disparity-only stimuli by Kane et al. [2014]. While the reductions in our experiment (and application) might introduce conflicts between disparity and pictorial cues, the latter seem to play the dominant role in depth perception, and tolerance for disparity reduction is higher. Still, as can be seen in Fig. 5.7, d, edges at larger depth discontinuities must be preserved, and in the temporal domain (Fig. 5.7, e) disparity should follow the pixel flow, while the temporal update of specific disparity values can be sparse.

Comparison to other metrics

To see if common metrics could predict the equivalence found, we compute their prediction of the difference between the reference and all our distorted stimuli and perform both linear ($a + b \cdot x$) and log-linear ($a + b \cdot \log(x + c)$) fits to the equivalence value across all distortions and stimuli. As common metrics we have tested peak signal to noise-ratio on depth and on the image pair [Merkle et al. 2009], a perceptual disparity metric on depth [Didyk et al. 2011] and a structural image similarity metric [Wang et al. 2004] on image pairs.

Fig. 5.8, (e) shows a scatterplot relating mean equivalence ratings by subjects to the result of numerical metrics. If any linear fit to a metric would predict equivalence well, its response would need to form a line. Similarly, a log-linear fit would need to form a logarithmic curve. However, we see that all fits predict the actual perceived difference rather poorly. PSNR has a correlation of $R^2 = 0.44$ for the linear and $R^2 = 0.44$ for the log-linear fit (all correlation statements in this section are DOF-adjusted R^2 values with $p < .01$ regression significance). As expected, the above-mentioned notions lack perceptual foundation and cannot predict perceived differences. Of greater interest is the finding that the perceptual metrics also do not fully predict equivalence. Although a

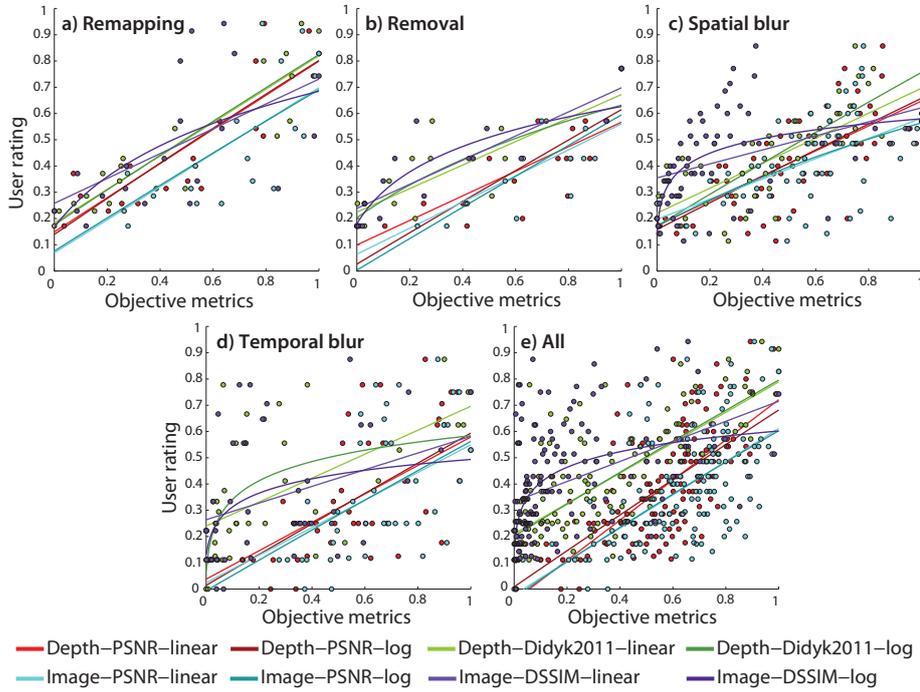


Figure 5.8. Linear and log-linear fits of numerical metrics to mean equivalence ratings by subjects. Different metrics are coded as colors. For each point, the response of the common metric defines the horizontal position and the mean equivalence rating of subjects defines the vertical position.

perceptual model of disparity results in the highest correlation of $R^2 = 0.57$ (linear) and $R^2 = 0.57$ (log-linear) and it is therefore a recommended option, it is still far from perfect. Further research in the area is needed. The idea to directly compare the image pair [Merkle et al. 2009] did not result in an improvement, except for the removal distortion.

We see that a linear fit to the perceptual model produces the best result, while providing only a weak correlation. In absence of any better model for equivalence, the fit with $a = 0.210$, $b = 0.579$ to Didyk et al. [2012] could serve as surrogate. We conclude, that even perceptual metrics cannot capture the task-specific challenge of visual equivalence for 2D-to-3D stereo conversion and that explicit user studies are required until a computational equivalence test is available.

Similarly poor performance of objective metrics was also observed when individual experiment conditions were analyzed separately for object removal (see Fig. 5.8, b), spatial filtering (see Fig. 5.8, c) and temporal filtering (see Fig. 5.8, d). The only exception was found in smooth remapping (see Fig. 5.8, a) where depth-based metrics performed well and achieved a correlation of up to $R^2 = 0.82$. It seems that existing metrics can deal much better with global monotonic manipulation as introduced by global remapping operators than with spatially varying artifacts that can arise from measurement imperfections or compression. See Table 5.1 for the complete list of measured correlations.

Table 5.1. Linear and log-linear correlation R^2 coefficients of study results with various metrics. Negated values used for PSNR. Measures that explain a certain distortion best are shown in bold face.

Experiment	Depth				Image pair			
	PSNR		Didyk2011		PSNR		DSSIM	
	Lin.	Log.	Lin.	Log.	Lin.	Log.	Lin.	Log.
Remap.	0.75	0.75	0.82	0.82	0.53	0.52	0.31	0.35
Removal	0.60	0.55	0.64	0.64	0.65	0.62	0.72	0.76
Spat. blur	0.49	0.48	0.60	0.56	0.30	0.30	0.15	0.32
Temp. blur	0.42	0.42	0.39	0.46	0.31	0.31	0.16	0.17
All	0.43	0.43	0.57	0.57	0.30	0.30	0.17	0.25

Scope and limitations

Working with natural images has inherent difficulties not found for artificial stimuli. Many other cues such as saliency [Borji and Itti 2013, Itti et al. 1998] due to luminance, motion, or stereo itself certainly affect the result. We have addressed this by performing the distortion either globally or in areas that are likely most salient (such as the moving character).

Another limitation of natural images is that we require both ground truth depth and natural images. While this data is easy to acquire for synthetic images, natural video content with ground truth depth maps is hard to come by. Most data sets available could be called to have a substantial “campus and LIDAR”-bias: they result from scanning an open street-level setting with houses and roads using a laser scanner. Practical stereo movie content however is drastically different, involving fractal natural objects, close-ups, human and non-human characters, careful scene arrangement, artistic field-of-view and cues from scene or observer motion.

Even the Sintel dataset does not have certain types of motion that are important in practice, in particular discontinuous motion that makes an important ingredient of vividly moving characters such as in sports broadcasting. This is why we add a selection of four movies with such motion to the set.

Finally, our experiment only provides evidence that equivalence is not covered well using common measures. Besides the recommendation for the closest fit with a moderate amount of correlation, this is a partially negative result: It indicates, that user studies are clearly superior over numerical comparison, but does not provide a way to measure functional equivalence computationally. Note, that this however is not yet possible for luminance images and remains future work in computational stereo perception.

Recommendations

The main conclusion to be drawn from the observations made is, that metrics, be it numerical or perceptual, are poor predictors of perceptual equivalence when assessing disparity quality. Instead, results should be compared by explicit user studies, which can reveal a picture entirely different from MSE, PSNR or even from perceptual disparity metrics.

The agreement with luminance edges is of particular importance. When looking at stereoscopic content, nothing is worse than a sharp disparity edge that does not align to a well-visible luminance edge. This leads to the recommendation to actually warp

the image and show it to subjects, as only the combination of depth and luminance will allow for any conclusion.

Finally, if the objective is stereo video, the comparison has to be made on video, where the tolerance for errors is even higher than it already is for images. In fact it is so high, that for typical natural image footage, blurs with a standard deviation of around an entire second did not show a significant non-equivalence, even after a large number of trials and subjects.

5.2 Our approach

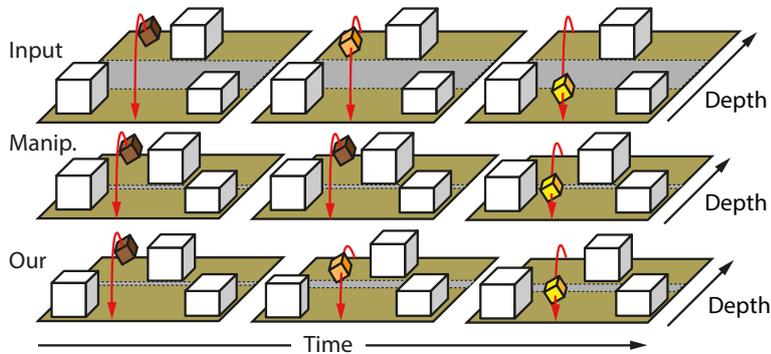


Figure 5.9. Starting from an input stereo content (*Top row*) with motion (*red arrow*), here shown in different frames (*Columns*) a typical manipulation is disparity compression to achieve viewing comfort (*Second row*). In this example, the flying cube will be slowed down in the proximity of the other cubes and will tend to jump over the empty space between the cubes (marked in grey), where the manipulated disparity is compressed. Our approach (*Third row*) finds a compromise that allows the manipulation where possible and restores motion in depth.

Our approach takes the temporal disparity field processed by any arbitrary disparity manipulation method and then restores the possibly altered motion-in-depth represented by the disparity change over time (Fig. 5.9). It matches it to the disparity change over time in the original temporal disparity field but it preserves the spatial disparity characteristic introduced by the manipulation. A post-process design of our method enables its application to a general disparity manipulation, e.g., disparity retargeting [Lang et al. 2010, Yan et al. 2013] or disparity compression [Didyk et al. 2011], without its modification or knowledge of implementation details.

To this end we devise a cost function for a potential mapping (Sec. 5.2.1) that is minimized, leading to a mapping that preserves disparity kinematics (Sec. 5.2.3). In our optimization we perform a perceptual scaling of disparity velocity changes to better account for the actually perceived changes of object velocity (Sec. 5.2.2). Spatio-temporal subsampling (Sec. 5.2.4) and GPU processing (Sec. 5.2.5) are required to achieve the real-time performance of our optimization solver. Finally, we use a novel 3D warping approach (Sec. 5.2.6) to synthesize a new stereo image pair that conforms to the optimized disparity map.

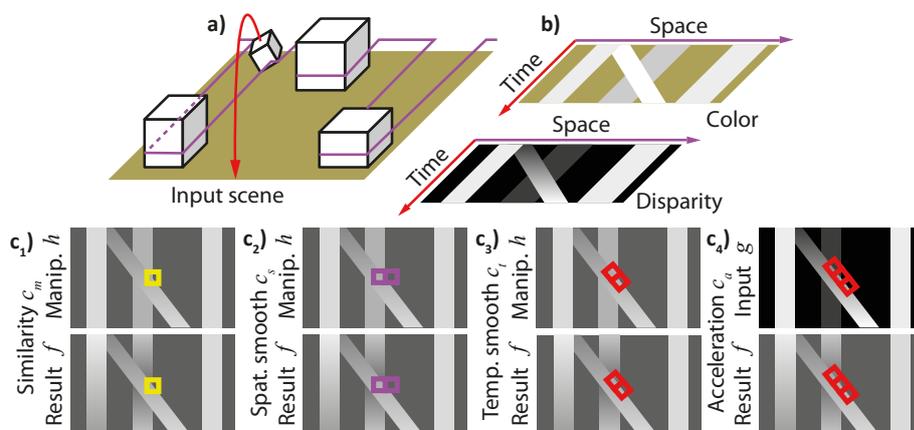


Figure 5.10. The cost function computation: (a) Input scene as in Fig. 5.9, where a small cube is moving along a red trajectory. Slicing the scene (violet line) and adding time as the second dimension results in a temporal light field-like RGB field for two eyes (b)-top and a temporal disparity field where brightness depicts depth (b)-bottom of the input scene. Our optimization (c) finds the time-varying disparity image f such that it preserves the manipulated stereo content in h and the input stereo motion in g . The similarity cost $c_m(f)$ matches disparity at the same space-time positions (c_1 : yellow box) in f and h . The smoothness costs $c_s(f)$ and $c_t(f)$ match the disparity difference at the same positions in space (c_2 : violet box) and time (c_3 : red box) in f and h . Note, how temporal smoothness is aligned with the motion flow. The acceleration cost $c_a(f)$ matches the second order central disparity difference at the same positions (c_4 : red box) in f and g .

5.2.1 Cost function

Let $\Omega = \mathbb{R}^2 \times \mathbb{R}^+$ be the space-time domain and $\mathcal{S} = \Omega \rightarrow \mathbb{R}$ be the set of all time-varying disparity images defined on it. Disparity in this work refers to vergence angles or pixel disparity measured in arc minutes, which requires a known observer-to-screen distance and screen size. Now, consider $g \in \mathcal{S}$ as well as $h \in \mathcal{S}$, which denote an original and a manipulated time-varying disparity image. A typical change from g to h could be disparity retargeting [Lang et al. 2010, Yan et al. 2013] or disparity compression [Didyk et al. 2011]. Our approach finds a third time-varying disparity image $f \in \mathcal{S}$ that optimally combines manipulation and stereo motion preservation with respect to certain costs as shown in Fig. 5.10.

Our cost function is designed to balance the following four factors. We want the optimized results f to remain similar to the manipulated stereo content h , and the disparity changes introduced, to be smooth in a local space-time neighborhood. At the same time we want to preserve the velocity of the original content g , and strongly penalize any acceleration changes. To this end we have to change the manipulated input once more. Doing this, we need to ensure that performed changes are spatially and temporally coherent. To this end we use four cost functions to be defined now.

First, the optimized time-varying disparity image f should be similar to the manipulated one h (Fig. 5.10, c_1)

$$c_m(f) = \int_{\Omega} (f(\mathbf{x}, t) - h(\mathbf{x}, t))^2.$$

where t denotes time and \mathbf{x} 2D position in the screen space. In what follows, the bold

notation is used for vectors.

Second, the change between the disparity images f and h should be spatially smooth (Fig. 5.10, c_2). In order to allow changes in compression manipulation independently for individual moving objects, the spatial smoothness term is weighted non-uniformly based on the inverted local spatial gradient magnitude of the manipulated disparity:

$$c_s(f) = \int_{\Omega} \left\| \begin{bmatrix} a_x(\mathbf{x}, t) & 0 \\ 0 & a_y(\mathbf{x}, t) \end{bmatrix} \cdot \left[\nabla_{\mathbf{x}}(f(\mathbf{x}, t) - h(\mathbf{x}, t)) \right]^T \right\|^2,$$

where $\nabla_{\mathbf{x}}$ is the gradient with respect to \mathbf{x} , and $a_x(\mathbf{x}, t)$, $a_y(\mathbf{x}, t)$ are power functions ensuring that the cost function is little affected on surfaces while smoothing at object silhouettes is strongly suppressed as in [Farbman et al. 2008]:

$$a_x(\mathbf{x}, t) = e^{-10 \left| \frac{\partial h(\mathbf{x}, t)}{\partial x} \right|} \quad \text{and} \quad a_y(\mathbf{x}, t) = e^{-10 \left| \frac{\partial h(\mathbf{x}, t)}{\partial y} \right|}.$$

We observed that by setting the exponent to -10 any visible degradation of sharp disparity transitions on boundaries of objects is prevented, while sufficient freedom to disparity changes is provided otherwise.

Third, an introduced additional modification of disparity should be smooth along the object motion (Fig. 5.10, c_3):

$$\begin{aligned} p_f(\mathbf{x}, t) &= p(\nabla f(\mathbf{x}, t) \cdot \mathbf{u}(\mathbf{x}, t)) \\ p_h(\mathbf{x}, t) &= p(\nabla h(\mathbf{x}, t) \cdot \mathbf{u}(\mathbf{x}, t)) \\ c_t(f) &= \int_{\Omega} \left(\underbrace{p_f(\mathbf{x}, t)}_{\text{Optim. Vel.}} - \underbrace{p_h(\mathbf{x}, t)}_{\text{Manip. Vel.}} \right)^2, \end{aligned}$$

where $\mathbf{u}(\mathbf{x}, t)$ is the 3D screen space-time normalized motion vector at position \mathbf{x} at time t , ∇ the gradient with respect to \mathbf{x} and t , \cdot the 3D dot product, and $p \in \mathbb{R} \rightarrow \mathbb{R}$ is a function that maps physical disparity velocity to perceptual units (Sec. 5.2.2).

Finally – and most different from previous work on energy-based image or disparity manipulation – the original acceleration in g should be preserved (Fig. 5.10, c_4). We therefore construct the acceleration term to match the second derivative of manipulated time-varying disparity image f to the second derivative of original disparity g along the motion path $\mathbf{u}(\mathbf{x}, t)$:

$$\begin{aligned} p_g(\mathbf{x}, t) &= p(\nabla g(\mathbf{x}, t) \cdot \mathbf{u}(\mathbf{x}, t)) \\ c_a(f) &= \int_{\Omega} \left\| \underbrace{\nabla(p_f(\mathbf{x}, t))}_{\text{Optim. Acc.}} - \underbrace{\nabla(p_g(\mathbf{x}, t))}_{\text{Orig. Acc.}} \right\|^2. \end{aligned}$$

Other methods for modification of image sequences often rely on per-frame temporal smoothness enforced by minimization of first temporal derivative [Lang et al. 2010, Didyk et al. 2011, Yan et al. 2013, Krähenbühl et al. 2009, Wang et al. 2010]. That reduces the change of optimized image property, disparity in our case, over time so it becomes as constant as possible. Such approaches have two key issues. First, they often operate on per-frame basis which might result in limited temporal extent of the optimization depending on the solver used. We instead use sparse samples in the temporal domain to capture the global temporal characteristic of the motion-in-depth in the range of several seconds. Second, first-order smoothing also removes high frequencies in the temporal disparity signal, turning every original motion into a smooth

motion. Instead we use the second temporal derivative (acceleration) of the disparity and match the manipulated time-varying disparity image f to the original acceleration in g . This reintroduces the original motion characteristic independent of its scale that could have been both locally and globally altered by changes in disparity range and distribution. Depth-map guided temporal upsampling later guarantees that we recover high temporal frequencies of disparity lost by sparse sampling (Sec. 5.2.4). Our method is therefore equivalent to the simple temporal smoothing only in the case of constant speed motion.

Our smoothness term $c_t(f)$ only states that the optimization to restore the motion should respect the manipulated disparity image h and should not alter it rapidly. Therefore it does not contradict the acceleration term $c_a(f)$.

5.2.2 Perceived disparity velocity changes

We want to model the HVS sensitivity to a change of disparity over time which, as we discussed in Sec. 2.2, contributes to the perceived velocity of MID through the CDOT mechanism. Such a model should ensure that perceptually important motion characteristics are well-reproduced and otherwise the optimization can safely ignore imperceptible motion distortions.

We assume that the sensitivity follows the Weber-Fechner law and based on the measurement in [Portfors-Yeomans and Regan 1996, 1997] we conservatively set the Weber fraction $k = 0.08$ irrespectively of the motion direction. Let $\dot{\alpha} = d\alpha/dt$ be the change of disparity over time. Then the perceived disparity velocity of $\dot{\alpha}$ is $p(\dot{\alpha}) = \frac{1}{k}(\ln(\dot{\alpha} + 1) - \ln(\epsilon + 1))$, where ϵ is the smallest disparity velocity that can be detected. To compare two disparity velocities $\dot{\alpha}$ and $\dot{\beta}$, using $\Delta p = p(\dot{\alpha}) - p(\dot{\beta}) = \frac{1}{k}(\ln(\dot{\alpha} + 1) - \ln(\dot{\beta} + 1))$, it is not required to know ϵ . Effectively, Δp is scaled in sensory just noticeable difference units (JND), which also means that the velocity differences $\dot{\alpha} - \dot{\beta}$ for which $\Delta p < 1$ JND are not perceivable. Since $k = 0.08$, one needs to change the disparity velocity $\dot{\alpha}$ by at least 8% to discriminate any difference.

The model can be directly applied to the temporal smoothness term $c_t(f)$, where the perceived disparity velocity change is computed for the time-varying disparity images f and h . This is also the case for the acceleration term $c_a(f)$, where the physical disparity velocity is converted into the sensory response units $p(\dot{\alpha})$, prior to the disparity acceleration computation (in the discrete formulation the acceleration is approximated by the second order central differences based on Δp).

5.2.3 Minimization

Finding the best disparity $f = \arg \min_{\hat{f}} c(\hat{f})$ with

$$c(f) = w_m c_m(f) + w_s c_s(f) + w_t c_t(f) + w_a c_a(f)$$

is a constrained (to the target disparity range) optimization problem. The values $w_m = 0.05$ for the data, $w_s = 1.0$ for the spatial, $w_t = 0.1$ for the temporal, and $w_a = 0.1$ for the acceleration weighting are used in all our results. In the following, we will discretize and linearize the problem, before solving it numerically.

Discretization The solution space is discretized into n_s spatial and n_t temporal elements, which altogether requires $n = n_s n_t$ new disparity values to be found. Thereby

the solution is a real vector $\mathbf{f} \in \mathbb{R}^n$. Let $\mathbf{g} \in \mathbb{R}^n$ and $\mathbf{h} \in \mathbb{R}^n$ be discrete versions of g , the original and h , the manipulated time-varying disparity images.

Optimization The first two costs can be written as a discrete differential. The data term cost is $\|\mathbf{f} - \mathbf{h}\|^2$. The spatial smoothness cost is $\|A_n(\mathbf{f} - \mathbf{h})\|^2$, where the matrix A_n in row i is 4 in column i , -1 at all four elements with index j that are a spatial neighbor to i and zero otherwise. The two other costs are non-linear due to the perceptual model. Let A_f and $A_b \in \mathbb{R}^{n \times n}$ denote discrete versions of the forward and backward motion flow u , respectively. This motion flow permutation matrix encodes in row i and column j , how much the i -th space-time pixel is a result of forward or backward motion flow of space-time pixel j . The temporal smoothness is

$$\|p(A_f \mathbf{f} - \mathbf{f}) - p(A_b \mathbf{h} - \mathbf{h})\|^2,$$

where $p(\hat{\alpha})$ is applied element-wise. Finally, the acceleration cost is

$$\| (p(A_f \mathbf{f} - \mathbf{f}) - p(\mathbf{f} - A_b \mathbf{f})) - (p(A_f \mathbf{g} - \mathbf{g}) - p(\mathbf{g} - A_b \mathbf{g})) \|^2.$$

5.2.4 Upsampling

Solving the above minimization problem at the full space-time resolution of common stereo content can consume an intractable amount of time and memory. However, we find that the *coarse-to-fine* optimization works well in a subsampled space-time domain, followed by on-the-fly upsampling to the original resolution (Fig. 5.11). To capture all properties of motion, temporal sampling frequency must be high enough, so that a majority of points on any surface are visible at least in three consecutive frames. In all video sequences considered in this chapter we used a uniform subsampling of 1 : 10 in time and 1 : 5 in space which well under the visibility threshold of 3 Hz and 1 visual degree measured in our experiment (Sec. 5.1). Only the spatio-temporal change of disparity is upsampled and applied to the current frame, to keep fine details. Two different strategies address the upsampling in time and space.

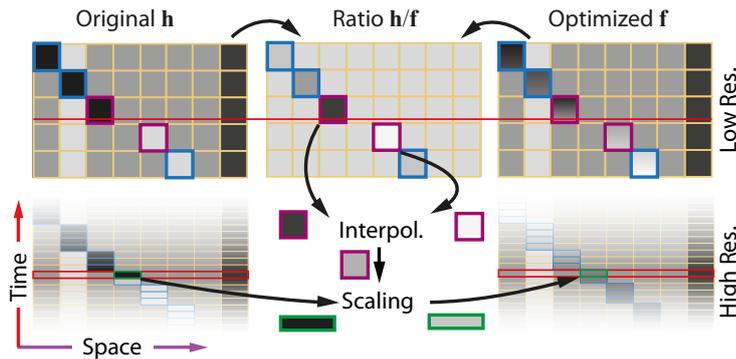


Figure 5.11. Temporal upsampling of a spatio-temporal disparity field. An object is moving in both image and depth space (*blue*). The ratio of the temporarily nearest original and optimized disparities \mathbf{h} and \mathbf{f} is interpolated with respect to the motion flow to get the disparity scaling at a given high-resolution frame time (*red*). Scaling is applied to the high-resolution disparity (*green*) to get the high-resolution output disparity.

Time In time, dense motion flow to the position in the previous keyframe and to the position in the subsequent keyframe are used. We detect occlusion in motion flow by comparing the depth value of the current pixel and the corresponding pixel along the motion flow. If those values differ significantly or the flow points outside the image, we consider the flow to be a disocclusion or occlusion and ignore the corresponding depth value in the interpolation. Instead of interpolating disparity in time, we interpolate disparity gradients and add them to the current high-resolution original frame. This preserves motion changes beyond the temporal sampling frequency of the optimization discretization.

Space In space, joint bilateral upsampling [Kopf et al. 2007] with the current high-resolution disparity image as the guidance is applied to the output of the temporal upsampling. Again, we interpolate gradients that are applied to the high-resolution frame. This allows for spatial details, finer than the spatial sampling frequency of the discretization of the solver.

5.2.5 Implementation

Optimization We use a gradient descent to find the best time-varying disparity image. The system cannot be optimized in closed form, due to the perceptual non-linearities and the boundary conditions of positive disparity. Starting from $\mathbf{f}^{(0)} = \mathbf{h}$, in every step i a correction vector $\mathbf{f}'^{(i)}$ is constructed from derivation of all costs and the solution is updated $\mathbf{f}^{(i+1)} = \mathbf{f}^{(i)} - \lambda \mathbf{f}'^{(i)}$. A $\lambda = 0.5$ and approximately 8 iterations were found to be sufficient for convergence to the solution for animations we tested. Adding more iterations and/or usage of smaller λ did not introduce any visible difference in final image sequence nor individual frames when compared visually.

GPU implementation The solver is implemented on a GPU and performs the updates of the mapping function in realtime. We maintain a window of previous frames combined with a prediction of future frames. The solver uses the last solution as the new initial guess. The deformation field is stored into a read-only 3D GPU buffer and each solver iteration is parallelized over all elements. If we consider the resolution of the subsampled space as a constant chosen based on content structure details rather than screen resolution the optimization runs in constant time independent of the target resolution. Please note, that a constant iteration count worked well in our experiments. Both spatial and temporal upsampling require processing linear in the number of output frame pixels. This is computationally equivalent to an application of a simple post-process filter, e.g., motion blur, which is a technique commonly used in real-time applications.

5.2.6 3D warping

Both, the manipulated and the optimized disparity maps do not necessarily correspond to any single pinhole camera projection. Therefore, such disparity patterns cannot be directly produced by conventional ray-tracing or rasterization. Instead, we have to modify the image locally. Typically the scene is rendered from a monocular center point of view and then image warping is performed [Lang et al. 2010, Didyk et al. 2010b]. Occlusions are resolved using a depth map, if available. However, disocclusions might still appear if some originally occluded region becomes visible, resulting in typical artifacts (Fig. 5.12).

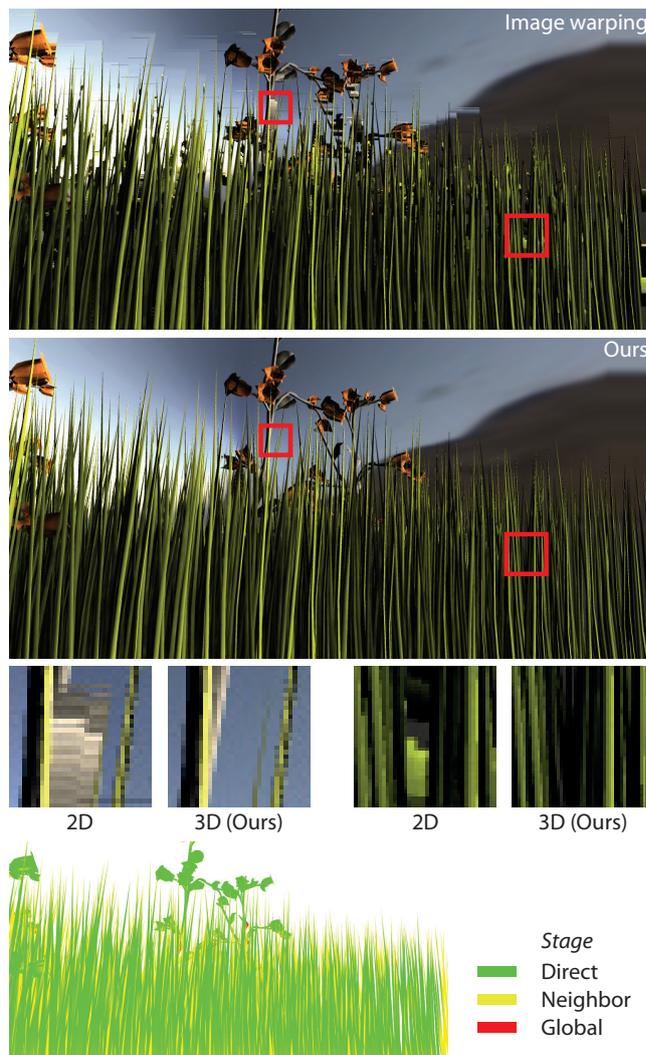


Figure 5.12. Common image warping and our approach applied to the left image of a stereoscopic image pair. The insets show close-up views of typical artifacts in common image warping. Color encoding in the bottom image illustrates, which of the three proposed vertex warping methods has been applied for a given image region.

To overcome this problem, we propose to render geometry twice. The first pass creates only a linear depth map that is manipulated and optimized in the pipeline described above. The second pass produces a color image pair, but moves vertices in the rasterization such that the resulting disparity conforms to the desired disparity. In order to allow for fine disparity mapping regardless of 3D geometry quality, we utilize tessellation on modern GPUs and adaptively subdivide triangles in object space up to the size of one pixel when projected on the screen. Larger thresholds can be used to favor performance. The rest of the rendering pipeline remains unchanged. We use a cascade of three increasingly approximate ways to reconstruct the appropriate vertex motion: *Direct fetching* followed by *neighbor fetching*, and finally *global disparity curves*.

Direct fetching In almost all cases, direct fetching is sufficient: Let \mathbf{v}_p be the pixel coordinates of the projection of a vertex \mathbf{v} . We read the depth map at \mathbf{v}_p and check if the difference between the depth of \mathbf{v} and the depth map value is smaller than a threshold ϵ . If this is the case, we read the disparity map and move the vertex to achieve the desired disparity between the left and right frames. This approach fails in the presence of occlusions or disocclusions, as the disparity map at \mathbf{v}_p does not contain the disparity that is to be assigned to \mathbf{v} .

Neighborhood stage If direct fetching fails, we first find the nearest depth in a 3×3 pixel-neighborhood of \mathbf{v}_p to increase robustness on object edges. If the minimal depth difference is smaller than ϵ we use the disparity value for the same position in the disparity map. Otherwise, we assume that the vertex was occluded in the original rendering. It might, however, not be occluded after disparity optimization or manipulation. Therefore we need to resolve what disparity it would have had if it was not occluded. We do that by searching the neighboring non-occluded depth map pixels for similar depth values. We assume that spatially close objects with similar depth are likely to have similar disparity.

We use 64 samples from the 2D Halton sequence to generate polar coordinates for sampling in a wider neighborhood of each vertex. The Halton sequence makes the sampling more robust to aliasing compared to regular sampling. The i -th sample position \mathbf{s}_i is

$$\mathbf{s}_i = \mathbf{v}_p + r_i^2(\cos(2\pi\alpha_i), \sin(2\pi\alpha_i)),$$

where $(r_i, \alpha_i) \in (0, 1)^2$ is the i -th element of the 2D Halton sequence. We use square of radius in order to sample the close neighborhood more densely. Once again, the depth difference smaller than ϵ indicates equality.

If there was no suitable value in the neighborhood either, it is still unclear where to move the vertex \mathbf{v} . As a last resort, we revert to a global disparity curve as explained next.

Global curve stage If the depth map sampling failed, there either is no visible object with similar depth in the disparity map or we failed to find it. In that case we cannot recover the correct disparity for the vertex but we try to minimize the error that would be observed as a rendering artifact. To this end, we reconstruct an approximation of the global curve mapping depth to disparity in a pre-process before the 3D warping. The mapping is constructed using radial basis functions with bandwidth prediction. In a first pass we predict bandwidth i.e., how many different disparity values map to a certain neighborhood of depth values. In a second pass, we reconstruct the mapping from depth to disparity with an adapted bandwidth.

In the first pass we build a standard histogram of the depth map. Populated bins will require a higher bandwidth, i.e., smaller kernel for reconstruction, less populated bins need a wider kernel. In particular, bins can be empty. In the second pass, we iterate over all depth-disparity value pairs. We use information about empty bins from the first pass to set the support of a hat reconstruction kernel of disparity values to the left and right neighborhood in the final histogram. The support matches to the distance to the next non-empty bin in a given direction. This way one depth-disparity pair may influence more than one histogram bin and therefore fill missing mapping intervals but does not cause blurring in other parts. Hat function filtering provides linear interpolation for empty intervals of mapping curve. We consider linear interpolation for these regions of depth range to be a conservative choice.

Discussion Conventional image warping [Lang et al. 2010, Didyk et al. 2010b] cannot deal with disocclusions. An alternative would be layered depth images (LDI) [Shade et al. 1998], that do contain all intersections of a viewing ray per pixel, not only the first. Similar to all image-based warping techniques, LDIs are prone to the undersampling problem, which might degrade the quality of synthesized images, in particular, for surfaces that originally have been seen under grazing angles. We avoid discretization altogether by first warping and discretizing later. Kim et al. [2011] suggest rendering from multiple perspectives to generate a 3D lightfield. Non-physical views can then be created as slices through this field. That, however, involves rendering of many data that will not be used. Our approach instead modifies the rasterization phase itself so that it only produces the final image with desired disparity. Our method only modifies the vertex projection phase of rendering and therefore is easily applicable wherever deferred shading [Deering et al. 1988] is used. We achieve that using vertex based warping directly on GPU. In our method, all disocclusions are resolved before the rasterization is performed, therefore, we do not lose any image information.

The resulting rendering might still produce artifacts if the global mapping curve does not match the local disparity mapping. We expect that the disparity manipulation roughly preserves some key properties such as depth ordering and therefore a global curve is a reasonable estimation of disparity at a given depth. While one could consider a more localized reconstruction of the mapping curve, in our scenes, we have not experienced any problems with the global curve approach and we observed a significant reduction of rendering artifacts due to the elimination of disocclusions (Fig. 5.12).

The rendering is done on per frame basis and the approach is therefore independent on temporal optimization. This makes it applicable to any other stereo content rendering problem.

Per-vertex warping makes the rendering performance highly dependent on the number of vertices in the scene. As the warping happens before culling, even vertices behind the camera will be processed. For real time applications an extension predicting what can be visible after the warping using simplified geometry could be implemented. This would prevent disparity sampling for occluded vertices which is the main performance issue. Another improvement would utilize temporal coherence to decrease the number of disparity map samples per vertex.

5.3 Validation

We performed two perceptual studies in order to evaluate the visual quality of object motion and to measure the performance in hit point-prediction for a ballistic target.

Please refer to Fig. 5.13 and the video on our website¹ for the stimuli.



Figure 5.13.  Three example trials from our performance study.

Setup In our experiments 10 and 8 different observers naïve in regard to the purpose of the experiment and with normal or corrected-to-normal (stereo) vision were observing a Zalman ZM-M240W polarized stereo display from a distance of 80 cm.

Preference study In the first experiment we displayed stereo video with compressed disparity (similar to the frame-by-frame application of global operator of Lang et al. [2010] without temporal smoothing) and the same video with our additional stereo-motion optimization side-by-side. Three computer generated scenes with motion in depth can be seen in Fig. 5.1 and Fig. 5.15. The saliency used to guide the compression was given either to the moving or the static object to simulate the artist’s intention or to emphasize an important object based on the scene’s semantics. The compression was set to be larger than what the display would require, in order to make the stimulus comparable to a reference video with non-manipulated disparity shown in the middle. Subjects were asked to indicate which of the two test sequences is more similar to the reference in terms of object motion. Our solution has been strongly preferred in 85.6% of the cases ($p < 0.01$, binomial test).

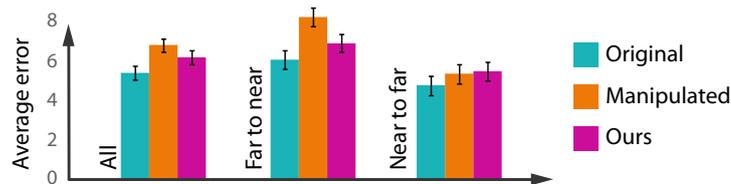


Figure 5.14. Motion direction errors with 95% confidence intervals.

Performance study In the second experiment the performance of hit point-prediction in terms of precision for a flying ball has been investigated (Fig. 5.13). We considered straight-line and ballistic motion trajectories of a ball moving in a random direction. The ball was shown only for a short initial interval after which observers used the mouse cursor to indicate its hit point on a ground plane. The closest world-space distance between the correct hit point and a ray through the clicked pixel was recorded. The stereo content was either unmodified, compressed, or compressed and processed using our approach. Fig. 5.14 summarizes the experiment outcome. An analysis of variance

¹<http://resources.mpi-inf.mpg.de/TemporalStereo#egsr2013>

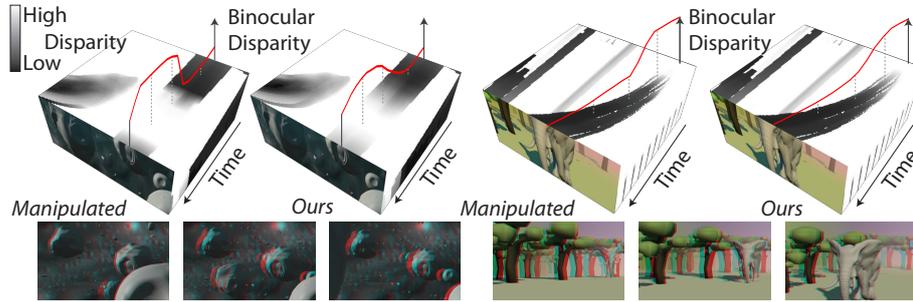


Figure 5.15. Two rendered scenes with complex deformation and camera motion used as stimuli in our study. *Top:* Slices through spatio-temporal disparity cubes after manipulation and our optimization. Time of animation proceeds from the back to front side of the cube. Plot shows values of disparity in time for single object in the scene as it moves in 3D space. Plot values were sampled from the cube, scaled and projected to match the cube orientation. *Bottom:*  Three frames from each scene.

(ANOVA) revealed a statistically significant effect ($F(2, 333) = 3.99, p < 0.02$) of reduced performance for the compressed disparity with respect to the unmodified one. When directions were analyzed independently there was the same effect for the far-to-near direction ($F(2, 165) = 5.57, p < 0.01$), but not for the near-to-far direction ($F(2, 165) = 0.64, p = 0.53$). We believe that impact of distortion is smaller in the near-to-far direction due to perspective scaling and smaller disparity change due to manipulation. This results in overall lower magnitude of error and smaller error differences between the original and manipulated disparity scenarios. As can be seen in Fig. 5.14, overall our method performed worse than unmodified disparity and better than the modified one, but in both cases we could not prove the significance of these effects.

5.4 Conclusion

This chapter introduced an approach to retarget disparity such that stereo motion can be reproduced faithfully. Our perceptual experiment showed that many spatial distortions of disparity can be tolerated by a human observer as long as they respect edges in the luminance signal. It has also shown what are the limits for detection of temporal artifacts in the disparity signal. Following these outcomes the stereo motion preservation problem was recast into a time-space-deformation problem that was solved using a numeric optimization procedure that allows for real-time performance. Our solution is independent of the particular stereo manipulation performed which makes it general. For the case of optimization-based manipulation, our perceptual disparity motion terms can be included in a combined optimization. Our validation study demonstrates that our disparity retargeting is strongly preferred over disparity manipulations that do not explicitly optimize for faithful motion in depth. The performance study clearly indicates that any disparity manipulation requires special attention in tasks that involve visual tracking of moving objects and precise judgment upon their possible collisions. Finally, we described a novel 3D warping approach to synthesize stereo image pairs that conform to a manipulated disparity map from polygonal 3D scenes. Application of this 3D warping is not limited to disparity maps produced by our system but is applicable to other manipulations as well.

Our approach is subject to several limitations. The perception of stereo motion might be affected by other factors, which have not been considered in this chapter, such as tracking and verging on the particular moving object, the object's luminance, texture, as well as its possible deformations. We relegate as future work a more in-depth investigation of those issues. Our current experiments were limited to computer-generated animations. Future work will need to show how approximations in the optical flow and scene depth reconstruction can affect our techniques.

Chapter 6

Gaze-driven disparity manipulations

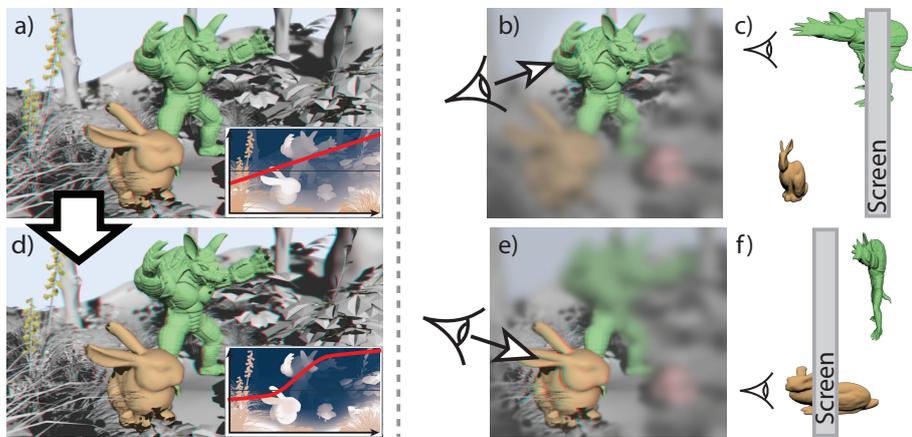


Figure 6.1. Beyond common disparity mapping (a) our approach adapts to attended regions such as the Armadillo (b) or Bunny (e) and modifies disparity in that region to enhance depth and reduce discomfort from the accommodation-vergence conflict (c,f). To this end, we build non-linear remapping curves and use our novel perceptual model to ensure a seamless transition between them (d).

In previous chapters we focused on motion that exists in the content independent on the will of a user. In this chapter we change this and examine consequences that relative motion of the image with respect to the retina that is caused by the motion of the eye itself has for reproduction of depth and possibilities for its local optimization.

A mental image of the surrounding world is built in the human visual system (HVS) by performing a sequence of saccadic eye movements and fixating at a sparse set of locations. This enables resolving fine spatial details in the fovea region of the retina, where the density of photoreceptors is highest, while it gradually reduces towards the retina periphery [Banks et al. 1991]. Gaze-contingent displays use eye tracking technology to monitor the fixation location and to conform the quality of depiction to the variable photoreceptor density. This way a more efficient control of the level of detail in geometric models [Murphy and Duchowski 2001], the image resolution in foveated rendering [Guenter et al. 2012], the state of luminance adaptation in tone mapping [Jacobs et al. 2015], the amount of blur in depth-of-field effects [Duchowski et al. 2014a], and the level of video compression [Geisler and Perry 1998] can be achieved, to name just a few key applications. When the fixation is shifted to another location the

image content must be adjusted accordingly, and saccadic suppression [McConkie and Loschky 2002, Loschky and Wolverson 2007], i.e., cutting off conscious registration of the blurred retinal signal due to fast saccadic eye motion (up to 1000 deg/s), is employed to hide the visibility of such adjustments. This imposes stringent requirements on the overall latency in the rendering system, as well as on the precision and sampling rate of eye tracking, which is used for the next fixation prediction.

The strategy of gaze-driven content manipulation is in particular interesting in the context of stereoscopic displays [Duchowski et al. 2014a]. Due to the well-known accommodation-vergence conflict [Hoffman et al. 2008, Zilly et al. 2011, Shibata et al. 2011] the range of disparities that can be shown on such screens is limited. To alleviate the problem, stereoscopic content has to be carefully prepared and manipulated [Lang et al. 2010, Didyk et al. 2011, Oskam et al. 2011]. This usually includes an aggressive compression of the depth range that can be presented on the screen, and, as a result, flattens the entire scene.

To address this problem, we propose gaze-contingent disparity processing that preserves depth information as much as possible around the fixation point and compresses it everywhere else. The key idea behind these manipulations is that they are performed at the eye fixation stage and remain imperceptible. We conduct a series of psychophysical experiments and observe that the HVS is insensitive to relatively fast, but smoothly performed depth manipulations, such as local depth range changes, or bringing the fixation point to the screen surface during the actual fixation. Based on the experimental outcome, we build a model that predicts the speed with which depth manipulations can be performed without introducing visible temporal instabilities. Such sub-threshold manipulations allow us to hide any latency issues of the eye tracker device, which makes our approach applicable even for low-cost devices. We investigate a number of applications for both real-time disparity optimization as well as offline content preprocessing. The contributions in this chapter are:

- a perceptual model that predicts the visibility of disparity manipulations during fixation,
- a metric of depth manipulation perceptibility,
- a real-time controller for adjusting stereoscopic content based on eye tracker data,
- a perceptual validation of the controller in terms of its seamless operation and local depth enhancement, and
- offline saliency-based solutions for disparity manipulation and scene cut optimization that improve viewing comfort.

6.1 Overview

In this chapter, we propose a new technique for manipulating stereoscopic content that accounts for the gaze information. To enhance perceived depth our method expands its range around the fixation location and reduces it in unattended regions that do not contribute significantly to depth perception. Additionally, objects around the fixation location are moved towards the screen to reduce artifacts such as visual discomfort (stereoscopic displays or virtual reality systems) or reduced spatial resolution (multi-view/lightfield displays). The main challenge here is to apply manipulations that adapt to rapid changes in fixations on the fly. We identify the following requirements guiding our design:

- depth manipulations should be performed with a speed nearly imperceptible to the observer so that the manipulations do not interfere with artistic designs,
- as the fixation point can change unexpectedly, it should always be possible to quickly recover to a neutral depth that provides acceptable quality across the entire image.

To address these requirements, we first study the sensitivity of the HVS to the temporal disparity changes (Sec. 6.2). As most disparity manipulations can be approximated by local scaling and shifting of depth (Fig. 3.1), we limit our study to these two manipulations. Based on the data obtained in the perceptual experiment, we next demonstrate how the visibility of temporal disparity manipulations can be predicted (Sec. 6.3.1). We use the resulting visible disparity change predictor to derive a sequence of disparity mapping curves, so that the target disparity can be achieved seamlessly in a minimal number of discrete steps (effectively frames) for any input disparity map (Sec. 6.3.2). This enables a number of applications for such formulated seamless disparity manipulation (Sec. 6.4). Besides the main real-time application, in which eye tracking data is available (Sec. 6.4.1), we demonstrate a few scenarios where gaze information can be either provided beforehand or predicted (Sec. 6.4.2–6.4.3). Furthermore, we propose a metric that predicts the visibility of any disparity manipulation for all possible gaze directions (Sec. 6.4.4).

6.2 Model for sensitivity to disparity manipulations

In order to determine how fast disparity shift and scaling can be applied before an observer notices changes, we conducted two separate threshold estimation experiments that were guided by the QUEST procedure [Watson and Pelli 1983].

6.2.1 Experiment 1: Disparity shifting

The goal of the first experiment was to determine the minimum speed at which a continuous shift of disparity becomes visible to an observer.

Stimuli Each stimulus consisted of a flat, circular patch that was textured using a high number of easily visible dots (random dot stereogram – RDS). The size of an individual patch spanned 18 deg. To investigate the impact of the initial disparity, we considered 7 different starting disparities $d_s \in \{20, 10, 5, 0, -5, -10, -20 \text{ arcmin}\}$ that were measured with respect to the screen depth. An example of stimuli used in our experiments is presented in Fig. 6.2a.

Task In order to measure the speed threshold, a two-alternative forced choice (2AFC) staircase procedure was used. At each trial a participant was shown two stimuli in randomized, time-sequential order. One of them was static while the other was moving in depth with constant velocity v_d . The direction of the motion was chosen to move the stimulus towards the screen as this is a likely scenario in a retargeting application. Each of the stimuli was shown for a period of 2.3 seconds, which was followed by 500 ms of a blank screen. The participant verged at the center of the stimulus and followed it as it moved in depth. The task was to decide which of the two stimuli contained motion or other temporal distortions and indicate the answer using arrow keys. The velocity of the moving stimuli was adjusted using the QUEST procedure. We chose to stop the staircase

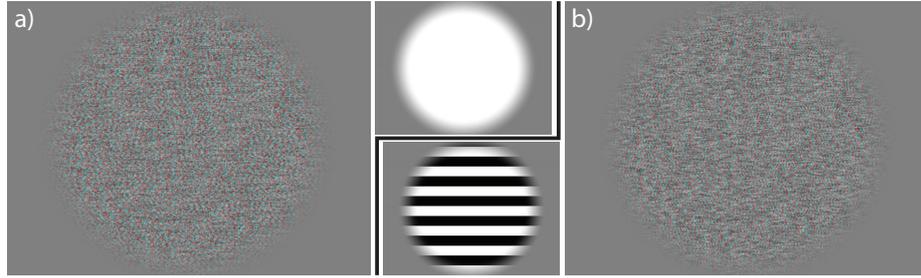


Figure 6.2. The random dot stereograms used in our experiments with disparity patterns in the middle. (a) Flat stimuli for Experiment 1. (b) Spatial corrugation for Experiment 2.

procedure when the standard deviation of the estimated threshold became smaller than 6.3 % of the initial estimate. The range of v_d considered by the procedure was set between 1 and 60 arcmin/s, which was determined in a pilot experiment conducted on five subjects.

Equipment In both experiments, the stimuli were presented using the NVIDIA 3D Vision active shutter glasses on a 27" Asus VG278HE display with a resolution of 1920×1080 pixels, at a viewing distance of 80 cm under normal, controlled office lighting. We avoided depth distortion due to the time-sequential presentation by excluding any frontoparallel motion [Hoffman et al. 2011].

Participants 14 participants (2 F, 12 M, 23 to 27 years old) took part in both our experiments. All of them had normal or corrected-to-normal vision and passed a stereo-blindness test by describing the content of several RDS images. Each of them completed threshold estimation procedures for all d_s in a random order. The subjects were naïve with respect to the purpose of the experiment. The average duration of the whole experiment was one hour. Participants were offered a break and they could resume the experiment on the next day.

Results The results of the experiments are presented in Fig. 6.3a. We observed a large variance of stereo sensitivity between subjects as expected for a general population [Coutant and Westheimer 1993]. We decided for a general model although personalization would be an option. While our initial hypothesis was that the speed threshold depends on the initial disparity, an additional analysis of variance did not show any effect ($F(6,72) = 0.42, p = 0.42$). This verified that the initial vergence does not influence the sensitivity significantly, and therefore, we model the threshold as a constant. Due to the significant variance in performance of individual users (ANOVA: $F(13,65) = 4.07, p < 0.001$), we used the median of all values as an estimate of the sensitivity threshold (the dashed line in Fig. 6.3a). Consequently, we model the disparity change thresholds as a constant:

$$v_b = c_0, \quad (6.1)$$

where $c_0 = 17.64$ arcmin/s.

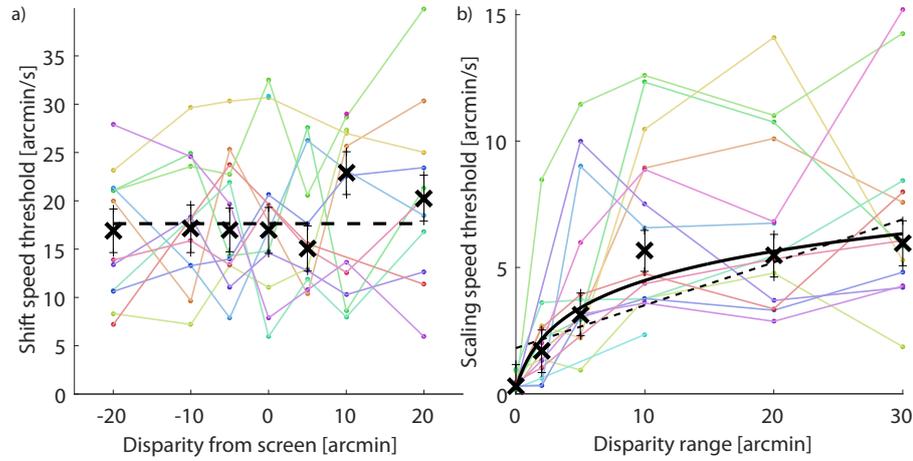


Figure 6.3. Results of Experiments 1 and 2 and our fitted model (a,b). Colors encode individual subjects; black crosses are median values across all subjects. (a) Thresholds as a function of the disparity from the screen (Experiment 1) with global median values shown as dashed line. (b) Thresholds as a function of the disparity range (Experiment 2) with both linear and logarithmic fit to median values.

6.2.2 Experiment 2: Disparity scaling

The goal of the second experiment was to measure how quickly the scene disparity range can be scaled before the temporal changes become visible.

Stimuli Similarly to the previous experiment, here we used a patch textured with a high contrast dot pattern. As we seek a speed threshold for disparity scaling, we considered a patch with a square wave disparity corrugation (Fig. 6.2b). To make our model conservative, we chose corrugation frequency to be 0.3 cpd as the HVS reaches its peak sensitivity for such a signal [Bradshaw and Rogers 1999]. We also used a square wave instead of sinusoidal one as it generalizes better for step functions [Kane et al. 2014] which successfully capture our manipulations that mostly occur between object edges. Because the sensitivity to disparity greatly depends on the amplitude of the disparity corrugation [Didyk et al. 2011], we consider different initial disparity ranges/amplitudes $d_a \in \{0, 2, 5, 10, 20, 30 \text{ arcmin}\}$. The values were chosen so that they do not result in a diplopia [Tyler 1975, Didyk et al. 2011]. The disparity corrugation was always centered around the screen plane, i.e., the average disparity of the patch is zero.

Task The procedure was similar to the previous experiment, with the exception that instead of the motion introduced to the entire patch, we introduced scaling to the disparity of the patch as a change of peak-to-trough amplitude over time. The maximum velocity that was considered by the 2AFC staircase procedure was set to 20 arcmin/s, and it was determined in a pilot experiment to be clearly visible. At such a speed diplopia could be reached during the exposure time of 2.3 seconds, but in practice, participants usually reported temporal change before this happened. Each participant performed one staircase procedure for each value of d_a in a randomized order.

Results The results of the experiments are presented in Fig. 6.3b. We observed a significant effect of the initial disparity range on the scaling speed threshold ($F(5,72) = 10.88$, $p < 0.001$) with a growing yet saturating tendency. The thresholds for disparity scaling are generally lower than for shifting. This is expected as disparity perception is driven mostly by the relative, not absolute, changes of depth. As a result, the sensitivity of the HVS to the relative disparity changes is much higher [Brookes and Stevens 1989]. The variance between users is again significant ($F(13,64) = 2.14$, $p < 0.05$). Similarly as in the previous experiment, we used the median as an estimate of the thresholds (black crosses in Fig. 6.3b) to which we fit an analytic function. Because a linear function yields low DoF-adjusted $R^2 = 0.50$ and does not adequately describe the saturating shape visible in the data (dashed line in Fig. 6.3b), we use a logarithmic function which is known to be adequate for describing many mechanisms of the HVS. As a result, we model the disparity range change thresholds as a function of the disparity magnitude:

$$v_g(s) = c_1 + c_2 \cdot \log(s + 1), \quad (6.2)$$

where s is the disparity range size in arcmin and $c_1 = 0.1992$ and $c_2 = 1.787$ are the fitting parameters with DoF-adjusted $R^2 = 0.89$.

6.3 Our approach

6.3.1 Visible disparity change predictor

Our disparity manipulation sensitivity model from the previous section predicts visibility of disparity changes for simple stimuli. To predict visibility of disparity manipulations for complex images, we define a predictor \mathcal{V} that for a given original disparity map $D_o : \mathbb{R}^2 \rightarrow \mathbb{R}$, two disparity mapping curves $d, d' : \mathbb{R} \rightarrow \mathbb{R}$, and a time $t : \mathbb{R}^+$ predicts whether the transition between the two curves in time t leads to disparity changes that are faster than the thresholds in Eq. 6.1 and Eq. 6.2. Formally, we define the predictor as:

$$\mathcal{V}(D_o, d, d', t) = \begin{cases} 1 & \text{if the transition is visible,} \\ 0 & \text{otherwise.} \end{cases}$$

In order to compute $\mathcal{V}(D_o, d, d', t)$, we have to check whether there is a location where either absolute or relative disparity (see Fig. 3.1a) changes become visible. The first case occurs if there exists a location \mathbf{x} for which the absolute disparity change is faster than the allowed speed in Eq. 6.1, i.e.,

$$\exists_{\mathbf{x} \in \mathbb{R}^2} \frac{|D'(\mathbf{x}) - D(\mathbf{x})|}{t} > v_b, \quad (6.3)$$

where $D'(\mathbf{x}) = d'(D_o(\mathbf{x}))$ and $D(\mathbf{x}) = d(D_o(\mathbf{x}))$. The second case occurs if there exist two locations \mathbf{x}, \mathbf{y} such that the relative disparity between them changes too fast (see Eq. 6.2), i.e.,

$$\exists_{\mathbf{x}, \mathbf{y} \in \mathbb{R}^2} \frac{|\Delta D'(\mathbf{x}, \mathbf{y}) - \Delta D(\mathbf{x}, \mathbf{y})|}{t} > v_g(\Delta D(\mathbf{x}, \mathbf{y})), \quad (6.4)$$

where $\Delta D'(\mathbf{x}, \mathbf{y}) = D'(\mathbf{x}) - D'(\mathbf{y})$ and $\Delta D(\mathbf{x}, \mathbf{y}) = D(\mathbf{x}) - D(\mathbf{y})$. With these two criteria, we can formulate our predictor as:

$$\mathcal{V}(D_o, d, d', t) = \begin{cases} 1 & \text{neither Eq. 6.3 nor Eq. 6.4 holds} \\ 0 & \text{otherwise.} \end{cases}$$

This definition holds for small values of t , as the relative disparity thresholds are a function of disparity magnitude (Eq. 6.2), which changes when different disparity mappings are applied. In our work, we assume that it is sufficient if t is equal to the period of one frame.

6.3.2 Seamless transition to target disparity

Our visibility prediction can be used to design a seamless transition between two disparity mapping curves d and d' . If the two disparity mappings are similar enough and $\mathcal{V}(D_o, d, d', t) = 0$ for t equal to the period of one frame, the transition can be done in one frame. However, this might not be the case if more aggressive disparity manipulations are desired. In such cases, it is necessary to spread the transition over a longer period of time to maintain the speed of changing disparities below the threshold values. To this end, we have to construct a sequence of new disparity mapping curves that will be applied sequentially in consecutive frames. At the same time, we want to keep the transition time as short as possible. More formally, for a given original disparity map D_o , and two disparity mapping curves d and d' , we want to find a shortest sequence of disparity mapping curves $d_i : 0 \leq i \leq n$, one for each frame i , such that $d_0 = d$, $d_n = d'$, and $\forall_{1 \leq i \leq n} \mathcal{V}(D_o, d_{i-1}, d_i, t_i) = 0$. To make the construction possible, we assume that each curve d_i is an interpolation between d and d' . Consequently, we define each curve d_i using corresponding interpolation weights w_i as:

$$d_0 \equiv d, \quad d_n \equiv d', \quad (6.5)$$

$$d_i(x) = (1 - w_i) \cdot d(x) + w_i \cdot d'(x) \quad 0 \leq i \leq n \quad (6.6)$$

$$w_0 = 0, \quad w_{i-1} \leq w_i, \quad w_n = 1, \quad (6.7)$$

where w_i is a sequence of the interpolation weights. This definition is equivalent to a formulation where disparity mappings are replaced with depth values from each stage of the transition:

$$D_0 \equiv D, \quad D_n \equiv D',$$

$$D_i(\mathbf{x}) = (1 - w_i) \cdot D(\mathbf{x}) + w_i \cdot D'(\mathbf{x}), \quad 0 \leq i \leq n,$$

$$w_0 = 0, \quad w_{i-1} \leq w_i, \quad w_n = 1,$$

for $D_i(\mathbf{x}) = d_i(D_o(\mathbf{x}))$. In order to make the transition seamless, we follow our visibility prediction described in Section Sec. 6.3.1 and obtain the following constraints that restrict absolute and relative disparity changes:

$$\forall_i \forall_{\mathbf{x} \in \mathbf{R}^2} \frac{|D_i(\mathbf{x}) - D_{i-1}(\mathbf{x})|}{t} \leq v_b \quad (6.8)$$

$$\forall_i \forall_{\mathbf{x}, \mathbf{y} \in \mathbf{R}^2} \frac{|\Delta D_i(\mathbf{x}, \mathbf{y}) - \Delta D_{i-1}(\mathbf{x}, \mathbf{y})|}{t} \leq v_g(\Delta D_{i-1}(\mathbf{x}, \mathbf{y})), \quad (6.9)$$

where t is the period of one frame and $\Delta D_i(\mathbf{x}, \mathbf{y}) = D_i(\mathbf{x}) - D_i(\mathbf{y})$. Now let us consider the term $D_i(\mathbf{x}) - D_{i-1}(\mathbf{x})$. It can be shown that:

$$\begin{aligned} D_i(\mathbf{x}) - D_{i-1}(\mathbf{x}) &= \\ &= (1 - w_i) \cdot D(\mathbf{x}) + w_i \cdot D'(\mathbf{x}) - (1 - w_{i-1}) \cdot D(\mathbf{x}) - w_{i-1} \cdot D'(\mathbf{x}) \\ &= (w_i - w_{i-1}) \cdot (D'(\mathbf{x}) - D(\mathbf{x})) \end{aligned} \quad (6.10)$$

By substituting this into Eq. 6.8, we can show that the constraint on the absolute disparity changes is equivalent to:

$$\forall_i \forall_{\mathbf{x} \in \mathbf{R}^2} \quad w_i - w_{i-1} \leq \frac{v_b \cdot t}{|D'(\mathbf{x}) - D(\mathbf{x})|} \quad (6.11)$$

Furthermore, using Eq. 6.10 we can also obtain:

$$\begin{aligned} \Delta D_i(\mathbf{x}, \mathbf{y}) - \Delta D_{i-1}(\mathbf{x}, \mathbf{y}) &= \\ &= (D_i(\mathbf{x}) - D_{i-1}(\mathbf{x})) - (D_i(\mathbf{y}) - D_{i-1}(\mathbf{y})) \\ &= (w_i - w_{i-1}) \cdot (D'(\mathbf{x}) - D(\mathbf{x})) - (w_i - w_{i-1}) \cdot (D'(\mathbf{y}) - D(\mathbf{y})) \\ &= (w_i - w_{i-1}) \cdot ((D'(\mathbf{x}) - D'(\mathbf{y})) - (D(\mathbf{x}) - D(\mathbf{y}))) \\ &= (w_i - w_{i-1}) \cdot (\Delta D'(\mathbf{x}, \mathbf{y}) - \Delta D(\mathbf{x}, \mathbf{y})) \end{aligned}$$

By substituting this into Eq. 6.9, we obtain a new form for the constraint on relative disparity changes:

$$\forall_i \forall_{\mathbf{x}, \mathbf{y} \in \mathbf{R}^2} \quad w_i - w_{i-1} \leq \frac{v_g(\Delta D_{i-1}(\mathbf{x}, \mathbf{y})) \cdot t}{|(\Delta D'(\mathbf{x}, \mathbf{y}) - \Delta D(\mathbf{x}, \mathbf{y}))|} \quad (6.12)$$

By combining Eq. 6.11 and Eq. 6.12, we can obtain the weights w_i that define the shortest transition between d and d' , such that it does not violate the constraints in Eq. 6.1 and Eq. 6.2:

$$w_0 = 0, \quad w_n = 1, \quad w_i = w_{i-1} + \Delta w_i, \quad (6.13)$$

$$\Delta w_i = \min_{\mathbf{x}, \mathbf{y} \in \mathbf{R}^2} \left(\frac{v_b \cdot t}{|D'(\mathbf{x}) - D(\mathbf{x})|}, \frac{v_g(\Delta D_{i-1}(\mathbf{x}, \mathbf{y})) \cdot t}{|\Delta D'(\mathbf{x}, \mathbf{y}) - \Delta D(\mathbf{x}, \mathbf{y})|} \right), \quad (6.14)$$

where t is the time of one frame. While different parametrizations of the transitions curves are possible, ours leads to a simple yet effective solution.

In order to construct the whole transition, we need to iterate Eq. 6.14 starting with $w_0 = 0$ until we reach $w_n = 1$. This is, however, computationally expensive as the evaluation of Eq. 6.14 requires iterating over all pixel pairs \mathbf{x} and \mathbf{y} , which leads to a quadratic complexity with respect to the number of pixels in the image. Instead, we propose a more efficient way of evaluating this equation by discretizing disparity maps into M values, so that there are only M^2 possible disparity pairs that we have to consider. If M is sufficiently large this will not create any accuracy issues. Assuming that the disparity range does not exceed -100 to 100 pixels, $M = 512$ results in errors not greater than $1/5$ of a pixel size. Consequently, we define an array H of size M such that $H[i] = 1$ if the disparity map D contains values between $\min(D) + i \cdot |\max(D) - \min(D)|/M$ and $\min(D) + (i+1) \cdot |\max(D) - \min(D)|/M$, and $H[i] = 0$ otherwise. Later, to evaluate Eq. 6.14, we consider all indices $i, j < M$ such that $H[i] = H[j] = 1$, and we refer to the corresponding values of disparities p_i and p_j .

6.4 Applications

Disparity manipulations are often performed by stereographers who use them as a storytelling tool. At the same time, additional disparity manipulations are applied to reduce the visual discomfort or to find the best trade-off between the image quality and

depth reproduction. We argue that the second type of manipulation should be performed in a seamless and invisible way, so it does not interfere with artists' intentions. In this section, we present applications of our model in different scenarios where such manipulations are crucial.

6.4.1 Real-time gaze-driven retargeting

In this section, we propose a real-time disparity manipulation technique that adjusts disparity information in the stereoscopic content taking into account gaze information. Our key insight is that depth information has to be accurate only around the fixation location and it can be significantly compressed in the periphery where depth perception is limited [Rawlings and Shipley 1969]. An additional improvement can be achieved by bringing the attended part of the image close to the screen [Peli et al. 2001, Hanhart and Ebrahimi 2014]. We make use of our technique for creating seamless transitions between different disparity mappings to assure that our manipulations do not introduce objectionable temporal artifacts and are robust to sudden gaze changes. An additional feature of our solution is that because the temporal changes to disparities are seamless, the technique is immune to latency issues of the eye trackers.

At every frame, our technique takes as an input the original disparity map $D_o(\mathbf{x})$ together with the current disparity mapping function d_p , and the gaze location \mathbf{g} provided by the eye tracking system. Then it proceeds in three steps (Fig. 6.4). First, it constructs a candidate mapping curve $d_c : \mathbb{R} \rightarrow \mathbb{R}$ which is optimal for the current frame. Next, it restricts d_c to $d_t : \mathbb{R} \rightarrow \mathbb{R}$ such that a quick recovery to a neutral linear mapping d_I in case of saccade is possible. As the last step, the current disparity mapping $d_p : \mathbb{R} \rightarrow \mathbb{R}$ is updated to $d : \mathbb{R} \rightarrow \mathbb{R}$ which is a single step of the seamless transition from Sec. 6.3.2. The mapping d is then applied to the image.

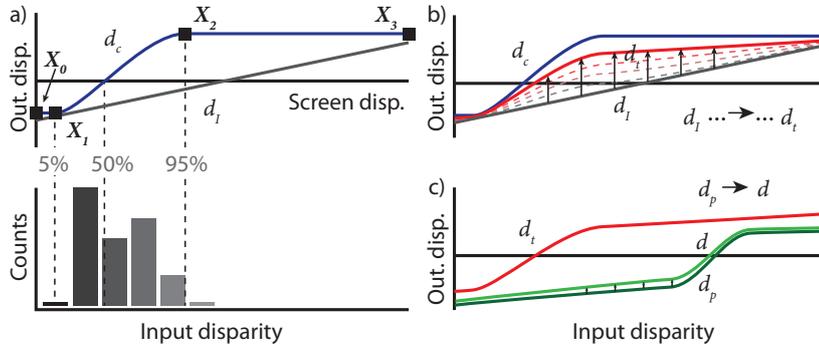


Figure 6.4. Construction of the disparity mapping for real-time retargeting. (a) A gaze-weighted disparity histogram is used to determine horizontal locations of the inner control points of the candidate curve d_c and the vertical offset necessary to minimize disparity from the screen in the gaze region. (b) An iterative transition algorithm finds the maximum transition from a neutral linear mapping d_I towards the candidate d_c in the defined time limit T_t as a target curve d_t . (c) The same algorithm is employed to update the previous curve d_p towards the target curve d_t and obtain a new mapping d to be used for the current frame.

Target curve construction To get the target curve d_t , we first build a candidate curve d_c parametrized by four control points $\mathbf{X}_{i \in \{0, \dots, 3\}} = [x_i, y_i]$ as presented in Fig. 6.4a. The

two outer points \mathbf{X}_0 and \mathbf{X}_3 restrict the entire scene to the comfort range $[r_{c,0}, r_{c,1}]$ of the displayable disparity:

$$\begin{aligned}\mathbf{X}_0 &= [\min(D_o), r_{c,0}] \\ \mathbf{X}_3 &= [\max(D_o), r_{c,1}].\end{aligned}$$

The two inner points \mathbf{X}_1 and \mathbf{X}_2 are responsible for the depth expansion around the gaze location. Therefore, their positions should span the range of disparities present around the fixation point. We define x -coordinates of \mathbf{X}_1 and \mathbf{X}_2 as the 5th (p_{05}) and 95th (p_{95}) percentile of the disparities around the gaze location. The percentiles are computed based on a histogram of D_o . To restrict its computation to the attended region and avoid temporal instabilities, we compute it as a weighted histogram, i.e., each disparity $D_o(\mathbf{x})$ contributes to the histogram according to the Gaussian $G_g(\mathbf{x}) = G(\|\mathbf{x} - \mathbf{g}\|, \sigma)$. Formally, we define the histogram H_G as:

$$H_G[i] = \sum_{\mathbf{x} \in \text{flR}(i)} G_g(\mathbf{x}), \quad i \in 0, 1 \dots M_G, \quad (6.15)$$

such that:

$$\begin{aligned}\text{flR}(i) &= \{\mathbf{x} : \min(D_o) + i \cdot z \leq D_o(\mathbf{x}) < \min(D_o) + (i + 1) \cdot z\}. \\ z &= |\max(D_o) - \min(D_o)| / M_G.\end{aligned}$$

The process of choosing the control points is presented in Fig. 6.4b. For the results in this chapter we chose σ to be 2.5 deg, as the stereoacuity significantly declines with the retinal eccentricity beyond this point [Rawlings and Shipley 1969], and the histogram size $M_G = 512$.

Initially, the inner segment of the curve d_t is constructed to map the disparities of the attended region to the entire available disparity range, i.e., $\mathbf{X}_1 = [p_{05}, r_{c,0}]$ and $\mathbf{X}_2 = [p_{95}, r_{c,1}]$. This forces the rest of the curve to be flat, but can also lead to scaling relative disparities beyond their original values. To prevent this, we limit the expansion between X_1 and X_2 by restricting the slope of the curve to 1. We achieve this by shifting the two control points toward each other with respect to the midpoint between them. Consequently, we define the control points X_1 and X_2 as:

$$\begin{aligned}\mathbf{X}_1 &= \left[p_{05}, \max(r_{c,0}, r_{c,1} + \frac{(r_{c,1} - r_{c,0}) - (p_{95} - p_{05})}{2}) \right] \\ \mathbf{X}_2 &= \left[p_{95}, \min(r_{c,1}, r_{c,0} - \frac{(r_{c,1} - r_{c,0}) - (p_{95} - p_{05})}{2}) \right].\end{aligned}$$

To bring the attended region close to the screen depth, we force the 50th (p_{50}) percentile of the disparities around the gaze location to map to 0. We achieve this by shifting all control points by p_{50} . The final control points are defined as:

$$\mathbf{X}'_i = \mathbf{X}_i - [0, p_{50}], \quad \text{for } i = 0 \dots 3. \quad (6.16)$$

To compute a smooth curve by the control points, we interpolate values between them using piecewise cubic Hermite interpolation and store the outcome in a discretized version using 256 bins.

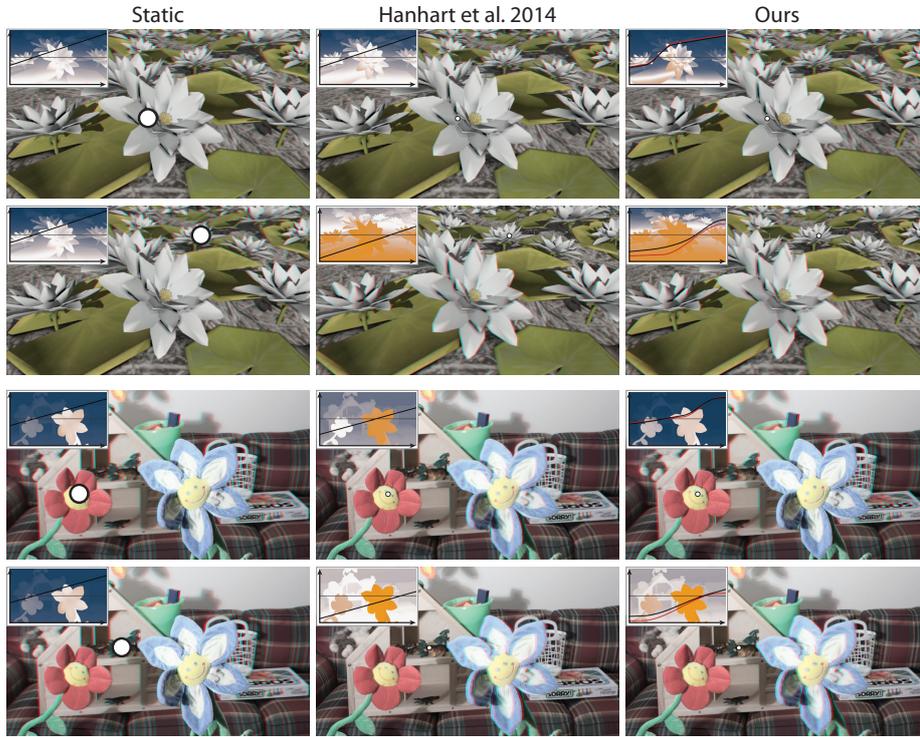


Figure 6.5. Comparison of our mapping (3rd column) with a static mapping (1st column) and the method of Hanhart and Ebrahimi [2014] (2nd column) as applied to our rendered image (top) and the image *Flowers* from the Middlebury dataset [Scharstein et al. 2014] for two different gaze locations (white dots). A disparity image with a mapping curve is shown in the insets. Crossed disparity is coded orange, uncrossed blue and screen disparity white. For our method the black curve is the rendered mapping d and the red curve is the target mapping d_t .

Quick recovery guarantee Depending on the depth variation in the scene and the gaze location, the disparity mapping curve d_c may correspond to very drastic changes in depth. This is undesired because we want to maintain a good depth quality even after a sudden gaze change. We solve this problem by refining d_c in such a way that using our seamless transition strategy we can recover from it within a predefined time period T_l . To guarantee this quick recovery, we derive the final target curve d_t by constructing a seamless transition from an identity disparity mapping d_I (Fig. 6.4a) to the candidate mapping d_c according to Eq. 6.14, and defining d_t as the mapping that is achieved at time T_l (Fig. 6.4b).

Seamless transition Although d_t is built in every frame, in order to prevent sudden disparity mapping changes, it cannot be directly used. Instead, in each frame we execute a single step towards this curve. To this end, we use Eq. 6.14 to compute a single step of a transition between previous disparity mapping d_p and d_t (Fig. 6.4c). Finally, we use the resulting curve d to generate a stereo image presented to the user (see Fig. 6.5 and Fig. 6.8) using the image warping technique of Didyk et al. [2010b].

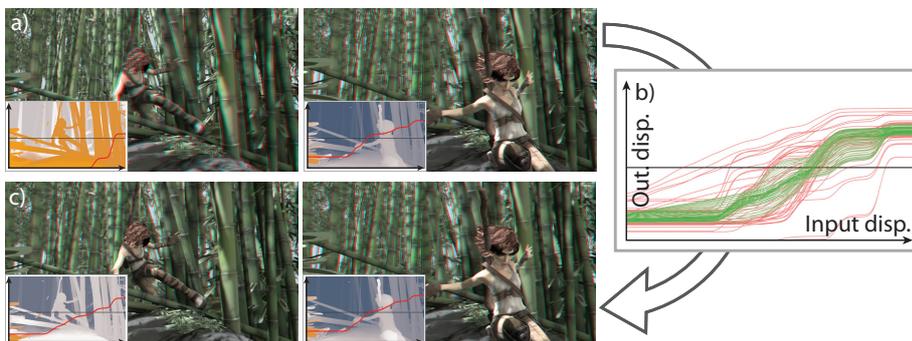


Figure 6.6. Our seamless transition applied to a per-frame saliency-based remapping [Lang et al. 2010]. (a) Two frames from the per-frame remapped sequence. (b) All per-frame (red) and our seamless transition (green) curves. (c) Our results. The Sintel video and disparity information are courtesy of the Blender Foundation and Butler et al. [2012], respectively.

6.4.2 Seamless disparity mapping in preprocessing

When the gaze location is not available, e.g., during post-production, our strategies can benefit from additional information about regions that are likely to be attended. For example, in movie production, it is common that attended image regions are known and purposely steered by a director. In other cases, a pre-viewing may be used to gather such data. In this chapter, we define this information as the probability distributions of gaze locations $S_k : \mathbb{R}^2 \rightarrow \mathbb{R}$ for each key frame $k \in [1, N]$. For the purpose of this chapter, we estimate S_k using an image based saliency estimator proposed by Zhang et al. [2013]. We also assumed that the entire video sequence is available so we can optimize the disparity mapping curves across the whole content. The key idea of this method is to compute per-frame optimal disparity mapping curves (Fig. 6.4a), and then optimize them so the transitions between them (Fig. 6.4c) are seamless according to our model.

For every key frame k we build a desired mapping curve \hat{d}_k (Fig. 6.6a). To this end, we follow the suggestion of Lang et al. [2010] and first build a histogram of disparity $H_w(D_k)$ similarly to Eq. 6.15 but with $G_g(\mathbf{x})$ replaced by the saliency map S_k . We also compute a standard histogram $H(D_k)$ using the same formula but with a constant weight $1/N_D$, where N_D is the number of pixels in D_k . To account for different sizes of salient regions across S_k , we normalize $H_w(D_k)$ by $H(D_k)$ prior to deriving the mapping curve u_k as a cumulative sum:

$$u_k[i] = \sum_{j=0..i} \frac{H_w[j]}{H[j]}.$$

This mapping attributes a larger disparity range to salient regions. We then derive \hat{d}_k by scaling u_k to the displayable range $[r_{c,0}, r_{c,1}]$ and shifting it to the screen to minimize the expected disparity from the screen estimated as the 50th (p_{50}) percentile of $H_w(D_k)$:

$$\hat{d}_k = u_k \cdot (r_{c,1} - r_{c,0}) + r_{c,0} - p_{50}.$$

There is no guarantee that the series of \hat{d}_k results in seamless manipulations, as drastic changes between neighboring frames can occur. We address this problem by finding a sequence of curves such that it provides seamless disparity changes. To this

end, we jointly optimize all curves d_k (Fig. 6.6b) according to the following strategy:

$$\begin{aligned} & \underset{d_k}{\text{minimize}} && E = |d_k - \hat{d}_k| \\ & \text{subject to} && \forall_{k \in [2, N]} \mathcal{V}(D_k, d_{k-1}, d_k, t_k - t_{k-1}) = 0 \\ & && \forall_{k \in [1, N-1]} \mathcal{V}(D_k, d_k, d_{k+1}, t_{k+1} - t_k) = 0, \end{aligned}$$

where t_k is the time stamp of the k -th key frame.

We solve this problem iteratively. We initialize $d_{k,0} = \hat{d}_k$. In the i -th step we compute the new candidate curve d'_k as:

$$\begin{aligned} d'_k &= (1 - \alpha)d_z + \alpha \cdot \hat{d}_k \\ d_z &= \frac{d_{k-1,i-1} + d_{k+1,i-1}}{2}, \end{aligned}$$

where $\alpha \in [0, 1]$ is obtained by solving a convex 1D problem:

$$\begin{aligned} & \text{maximize} && \alpha \\ & \text{subject to} && \mathcal{V}(D_k, d_{k-1,i}, d'_k, t_k - t_{k-1}) = 0 \\ & && \mathcal{V}(D_k, d'_k, d_{k+1,i}, t_{k+1} - t_k) = 0 \end{aligned}$$

using bisection. The mapping curve $d_{k,i}$ is then updated as:

$$d_{k,i} = (1 - \beta)d_{k,i-1} + \beta \cdot d'_k,$$

where $\beta = 0.1$ is a step size parameter. We stop the solver when $\max_k |d_{k,i} - d_{k,i-1}| < \epsilon$. We use $\epsilon = 0.1$ arcmin, which is achieved in less than 2 seconds for 50 key frames after ~ 100 iterations of our fast GPU solver running on a Quadro K2000M laptop GPU. If sparse key frames are used, then Eq. 6.6 is used to compute transitions between the intermediate frames. Samples from our results are presented in Fig. 6.6, and we refer readers to the video available on our website¹ for a full demonstration.

6.4.3 Scene cut optimization

Templin et al. [2014a] proposed an optimization for disparity transitions introduced by video cuts. They argued that minimizing the disparity difference at a cut reduces the eye vergence times and thereby improves the perceived image quality and scene understanding. To achieve this goal, disparity has to be retargeted on one or both sides of the cut and smooth transitions are required to blend to the original disparity. However, no precise model for such transitions was provided.

The seamless transition model for disparity mapping in Sec. 6.3.2 is well suited for this task. We optimize the cut by shifting disparities on both sides of the cut, which can be represented using a linear curve with a bias (see Fig. 3.1b). For simplicity, we assume that the time between subsequent cuts is always long enough to fit the entire mapping transition. Then, we can optimize each cut independently.

We first use the model of Templin et al. [2014a] to find the optimal bias h_o of the pixel disparity maps D_c and D_{c+1} on both sides of the cut at frame c . We follow their suggestion and solve the problem by minimizing:

$$h_o = \arg \min_h \sum_{\mathbf{x}} S(\mathbf{x}) V \left(D_c(\mathbf{x}) - \frac{h}{2}, D_{c+1}(\mathbf{x}) + \frac{h}{2} \right),$$

¹<http://resources.mpi-inf.mpg.de/GazeStereo3D/>

where $S : \mathbb{R}^2 \rightarrow \mathbb{R}$ is equivalent to the attention probability map S_k from Sec. 6.4.2 for the frame c . We use a uniform estimate in our examples. Function $V(a_0, a_1)$ stands for the vergence time model at the cut, where a_0 and a_1 denote the initial and target disparities [Templin et al. 2014a]:

$$V(a_0, a_1) = \begin{cases} 0.04a_0 - 2.3a_1 + 405.02 & \text{if } a_0 < a_1 \\ -1.96a_0 + 3.49a_1 + 308.72 & \text{if } a_0 \geq a_1 \end{cases}.$$

Linear mappings d_c and d_{c+1} are then built for each of the two cut frames with respective disparity shifts $h_c = -\frac{h}{2}$ and $h_{c+1} = \frac{h}{2}$ (Fig. 3.1b). For every other frame i with time stamp t_i , we use our transition model (Eq. 6.14) to derive the corresponding mappings d_i as a transition to the original mapping d_0 :

$$\begin{cases} \text{from } d_c \text{ to } d_0 & \text{if } i \leq c \\ \text{from } d_{c+1} \text{ to } d_0 & \text{if } i > c \end{cases}$$

for the duration $T_i = |t_i - (t_c + t_{c+1})/2|$. We again refer readers to the video on our website for an example of the resulting mapping.

6.4.4 Visibility visualization

In stereo content production when no assumptions can be made about the attended image region, our predictor of disparity change visibility (Sec. 6.3.1) can be used directly as a metric for the evaluation of a disparity mapping.

As an input we assume either two disparity mapping curves d and d' from two different frames, or the same disparity frame mapped by two different unknown curves as D and D' . The condition of the same frame can be relaxed if the distribution of physical depth in the scene does not change significantly over time. Additionally, we know the time span T between both inputs.

If only the mapped disparities D and D' are given, we construct the best approximation of the mapping curves between them rather than the mappings from the potentially unavailable original D_o (Sec. 6.3.1). The first curve d describes an identity mapping D to itself. The second curve d' describes a transition from D to D' and is constructed using a cumulative histogram, where each value from D' is accumulated to the bin corresponding to the value of D , and finally normalized by the number of accumulated values. The variance of values accumulated in each bin increases with a deviation from the global mapping assumption. The bins without any samples are filled by linear interpolation.

Now we can use the predictor $\mathcal{V}(D, d, d', T)$ to determine the visibility of a transition from d to d' in a binary way. To get the prediction in a continuous form, we can use our transition formula in Eq. 6.14 to compute the time T_c needed for a transition from d to d' as $T_c = n \cdot t$, where n is the number of discrete steps required. This allows us to formulate the metric score Q as the time needed relative to the time available:

$$Q = \frac{T_c}{T}.$$

The value units can be interpreted as just-noticeable differences (JNDs) and values lower than one can be considered imperceptible by the user, while values significantly larger can cause visible temporal artifacts as the depth is being transformed from one mapping to another.

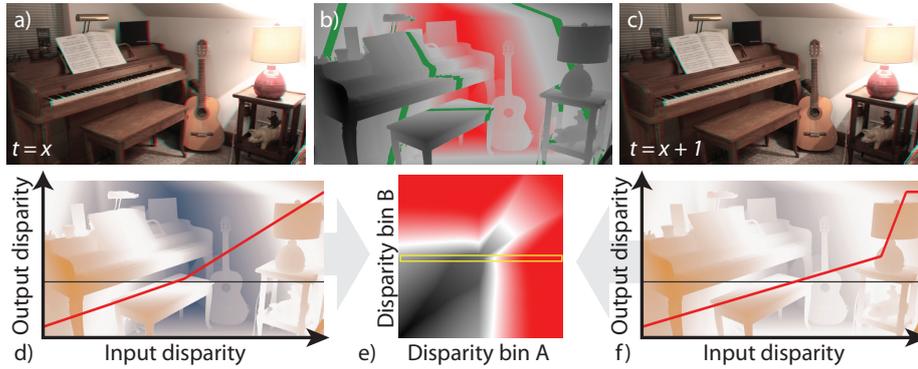


Figure 6.7. Output of our metric for 2 mappings as if transitioned over an interval of 1 second. (a, c) Boundary images. (d, f) Boundary disparity maps with their mapping curves. (e) The visibility matrix $Q(\mathbf{x})$ for every absolute and relative disparity. Red values ($Q > 1$) represent visible distortions. (b) Visualization of a single row of the matrix (yellow rectangle) as a distortion from each disparity pixel (red) with respect to the reference disparity (green) corresponding to the given row.

We also have an option to evaluate the metric for every absolute and relative disparity separately. This way, each pair of disparity values mapped by d and d' defines two linear mapping functions for which Q can be computed the same way. Enumerating all such pairs leads to a matrix representation $Q(\mathbf{x})$ and allows for a detailed inspection of the mapping properties and guiding the user towards the source of distortions. See Fig. 6.7 for an example.

6.5 Validation

6.5.1 Perceptual study

We evaluated the performance of our perceptual model and the disparity manipulation technique in a user experiment. To this end, we compared the depth impression and the temporal stability of our gaze-contingent disparity retargeting (Sec. 6.4.1) to three potential alternatives: first, a traditional static mapping which does not adapt to the gaze location in any way; second, an immediate shift of depth which brings the depth in the gaze location to the screen without temporal considerations (similar to [Bernhard et al. 2014]); and finally, the method of Hanhart and Ebrahimi [2014] as discussed in Sec. 3.2.1. Our model was derived for simple stimuli. To test its performance on complex images that contain more complex depth cues, we tested three variants of our method with different multipliers for the speed thresholds in Eqs. Eq. 6.1 and Eq. 6.2. We chose multipliers 1, 2 and 4.

Stimuli The techniques were compared using both captured and CG content. 4 stereoscopic images from the Middlebury dataset 2014 [Scharstein et al. 2014] and 2 from our own rendering were used as stimuli (Fig. 6.5 and Fig. 6.8).

Task We compared the three variants of our method with each other as well as with all alternative methods in a 2AFC experiment. At each trial a participant was shown the

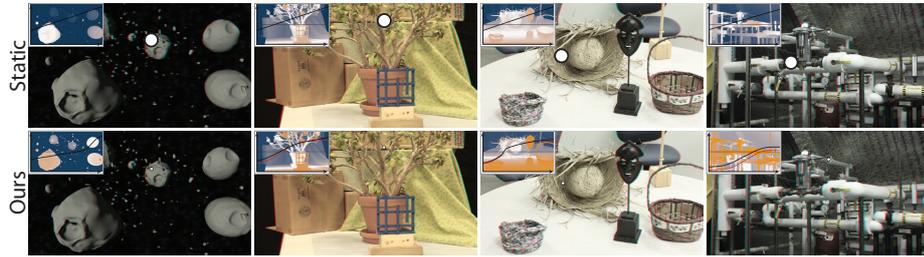


Figure 6.8. The stimuli used in our validation experiment (see also Fig. 6.5). From left: Our CG rendering and 3 images from the Middlebury dataset [Scharstein et al. 2014].

question and then the two stimuli in randomized, time-sequential order. Both contained the same content but with the disparity mapped in two different ways. Each of the stimuli was shown for a period of 10 seconds, which was followed by 800 ms of a blank screen. The participant answered one of the following questions:

- Which demo has more depth?
- Which demo is more stable?

The user could choose to repeat the sequence at will.

Equipment The stimuli were presented using the polarized glasses technology on a 24" Zalman ZM-M240W display with a resolution of 1920×1080 pixels, at a viewing distance of 80 cm under normal, controlled office lighting. The display technology was chosen not to interfere with the eye tracker Tobii EyeX that was used for the gaze-adaptive mapping. A chin rest was employed to improve the tracking performance and to prevent the participant from moving away from the optimal viewing angle.

Participants 14 participants (2 F, 12 M, 23 to 27 years old) took part in the study. All of them had normal or corrected-to-normal vision and passed a stereo-blindness test by describing content of several RDS images. The subjects were naïve to the purpose of the experiment.

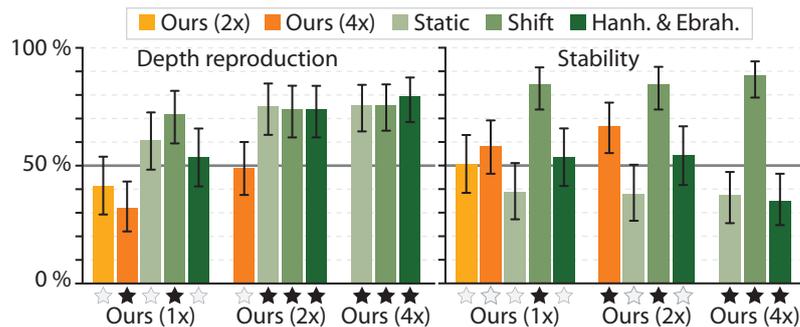


Figure 6.9. Results of our validation study for both the depth reproduction and stability questions. Each group of bars compares a variant of our method (multipliers 1, 2 and 4) against the other variants (warm colors) and competitor methods (green colors). 50 % is a chance level. A value above 50 % encodes participants' preference of the bottom label variant over the color-coded method. The error bars are confidence intervals. A significance in a binomial test is marked by a full star.

Results We have evaluated relative preference for each compared pair and each question (Fig. 6.9). Our method achieves a significantly stronger depth reproduction than a simple disparity shift (71.4%, binomial test $p < 0.01$) with the threshold multiplier 1, and than both a static mapping (75.0%, $p < 0.01$) and the method of Hanhart and Ebrahimi (73.9%, $p < 0.01$) for the multiplier 2. There was no significant difference between the depth reproduction of our method with the multiplier 1 and 2. This shows that this comparison of depth was a difficult task for users and required substantial disparity differences to be accomplished above the chance levels. There was significantly less depth reported for the multiplier 1 than for 4 (31.7%, $p < 0.01$); therefore, using a larger multiplier generally results in greater perceived depth, as expected.

Our method is significantly more stable than an immediate shift to the screen for the multipliers 1 (84.3%, $p < 0.01$), 2 (84.3%, $p < 0.01$) and even 4 (87.7%, $p < 0.01$). This illustrates that the latency of current eye tracking systems make performing modifications during the saccadic suppression difficult. This further supports our choice of relying on seamless disparity processing at the fixation. There was no significant difference in stability with respect to a static mapping and the method of Hanhart and Ebrahimi except for the highest multiplier 4 (35.8%, $p < 0.05$ and 35.0%, $p < 0.05$ respectively). The trend towards lower stability reports in a comparison to the static mapping visible for the lower multipliers is expected, as a presence of any visible difference between two stimuli will likely lead to a statistically significant difference in answers after a sufficient amount of trials. The discrepancy between close-to-chance results for the comparison of our multipliers 1 and 2 and the method of Hanhart and Ebrahimi, and on the other hand significant difference for the multiplier 4, suggests that the actual stability for the two lower multipliers is good.

The results show that our method can deliver more depth without sacrificing stability. The statistically higher stability of the multiplier 2 compared to 4 (66.7%, $p < 0.01$) and at the same time insignificantly but consistently higher depth reproduction than the multiplier 1, confirms that the multiplier 2 is a better choice for a complex stereo content. This is in agreement with previous observations about thresholds measured on artificial stimuli and their validity for realistic images, e.g., when measuring the perceivable color differences in CIELAB and CIELUV [Reinhard et al. 2010] or disparity differences [Didyk et al. 2011]. Further, our experiments show that the choice of the stimuli for the model construction (Fig. 6.2) generalizes for complex images, as the manipulations stay seamless when multipliers 1 and 2 are used, but become quickly visible when multiplier 4 is considered.

6.5.2 Limitations

Our perceptual model accounts only for disparity changes around fixation location; it does not account for peripheral sensitivity to motion. Although in our experiments we did not observe any problems, it might be interesting to investigate peripheral vision in the future, especially for wide-angle VR systems.

The “pop-out” effect, which brings scene objects in front of the screen, is often used as a storytelling tool. Our technique preserves it for quick temporal disparity changes, but the effect may diminish after the re-adaptation. This might only be a concern for standard stereoscopic displays. In autostereoscopic displays a significant “pop-out” effect is usually avoided as it leads to aliasing problems [Zwicker et al. 2006]. In VR displays, the “pop-out” does not exist as there is no notion of “in front of the screen”.

Our techniques rely on several methods that may introduce additional artifacts. In

particular, a poor estimation of visual saliency may lead to suboptimal results in our preprocessing application (Sec. 6.4.2). This is a general limitation of saliency-based manipulations, which can be improved by a director’s supervision or a pre-screening. The image warping technique used for generating our results can create monocular artifacts in disoccluded areas, if the disparity scaling is too large [Didyk et al. 2010b]. This together with cross-talk or aliasing during large shifts can potentially introduce artifacts perceived as additional 2D cues which can further affect the visibility of our disparity manipulations.

6.6 Conclusions

Gaze-contingent displays are gaining in popularity in various applications. Since such displays rely on the saccadic suppression to hide any required image changes from the user, their success strongly depends on the overall latency of the rendering system. In this chapter, we are interested in stereoscopic content authoring, which involves disparity manipulation, where the tolerability for the latency issues is very low. Our key insight is that near-threshold disparity changes can be efficiently performed at the eye fixation without being noticed by the user. This effectively makes the latency issues irrelevant and allows for use of such low-cost eye tracking solutions as deep learning based RGB eye tracker presented also in our recent paper [Khosla et al. 2016]. To this end, we measured the HVS sensitivity to disparity changes and formalize it as a metric. We employed the metric to guide the derivation of seamless transitions between frames in our gaze-contingent disparity retargeting. In this way, we improved the perceived depth significantly, while greatly reducing the requirements imposed on the eye tracker accuracy and latency. We also presented other applications of our metric in saliency-based disparity manipulations and scene cut optimization.

The benefits of our method extend beyond standard stereoscopic displays. New glasses-free 3D displays such as parallax-barrier or lightfield displays support only a relatively shallow depth range [Masia et al. 2013a]. As a result, the visual quality quickly degrades for objects that are further away from the screen plane. Head-mounted displays have also recently gained a lot of attention and including eye tracking in these devices is a natural next step. We believe that our method can provide a substantial quality improvement in all these cases. Gaze-driven techniques targeting specific display devices that use our model are an exciting avenue for future work.

Chapter 7

Motion parallax as a disparity supporting depth cue

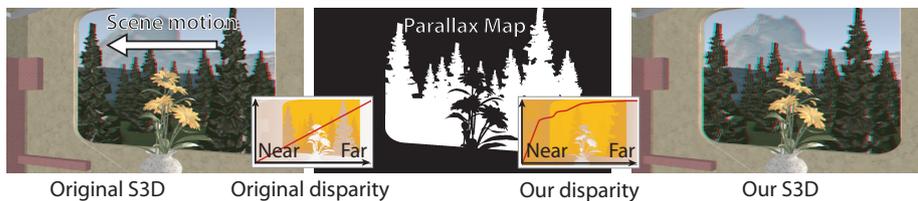


Figure 7.1. Starting from a stereoscopic video content with a static observer in a moving train (*Left*), our method detects regions where motion parallax acts as an additional depth cue (*Center, white color*) and uses our model to redistribute the disparity depth budget from such regions (*the countryside*) to regions where it is more needed (*the train interior*) (*Right*).

Previous chapters of this thesis discussed motion purely as a change of disparity over time, be it because of an object changing location and/or depth or because of a change of relative gaze location due to an eye saccade. We treated motion as a complication for a display, as a masking element for distortion perception or simply as an objective quality to preserve. In this chapter we look deeper in the structure of the motion itself and analyze the information that one particular kind of motion conveys about the depth configuration in a video sequence.

Perceiving layout of a scene is one of the main tasks performed by the human visual system. To this end, different depth cues [Cutting 1995] are analyzed and combined into a common understanding of the scene. Not all cues, however, provide reliable information. Pictorial cues such as shadows, aerial perspective or defocus, can often be misleading. Other cues, such as occlusion, provide only a depth ordering. There are also very strong cues such as binocular disparity or ocular convergence which provide a true stereoscopic impression. These, however, can only be reproduced in a limited fashion due to significant limitations of the current display technology. The first problem is the lack of correct accommodation cue in most of the current stereoscopic displays. This leads to a visual discomfort [Hoffman et al. 2008, Lambooj et al. 2009] when a large disparity range is shown to observers. While current research tries to address this problem with new, light-field displays [Masia et al. 2013b], these solutions have even stronger requirements regarding the depth range. This is due to the apparent blur for objects that are located at a significant distance from the screen plane [Zwicker et al. 2006, Wetzstein et al. 2012]. Because of the aforementioned reasons, it is crucial to take

any opportunity that allows us to improve depth reproduction without using additional disparity range.

One of the strongest pictorial depth cues is motion parallax. It arises when features at different depths result in different retinal velocities. The strength of this cue is relatively high when compared to other monocular cues. The relation remains also significant for binocular disparity [Cutting 1995]. This fact has been exploited in several applications, such as *wiggle stereoscopy* [Wikipedia 2015b] where motion parallax is used as a metaphor for stereoscopic images, or *parallax scrolling* [Wikipedia 2015a] used in games where by moving foreground and background with different speeds depth sensation is evoked. A striking example of motion parallax efficiency are species that introduce subtle head movements to enable motion parallax [Kral 2003]. This mechanism has been incorporated into cameras where apparent depth is enhanced by subtle motion [Proffitt and Banton 1999, v3© Imaging 2015]. Interestingly, motion parallax is not limited to observer motion, but also provides depth information whenever local motion in the scene follows a predictable transformation [Ullman 1983, Luca et al. 2007] (Fig. 2.3). These facts suggest that motion parallax is a very strong source of depth information for the human visual system, but it has never been explored in the context of stereoscopic image manipulations.

In this chapter, we address this opportunity and propose a computational model for detecting motion parallax and quantifying the amount of apparent depth it induces together with binocular disparity. To this end, we conduct a series of psychovisual experiments that measure apparent depth in stereoscopic stimuli in a presence of motion parallax. This is done for simple, sinusoidal corrugation stimuli. Based on the measurements, we propose a computational model that predicts apparent depth induced by these cues in complex images. Furthermore, we demonstrate how the model can be used to improve depth reproduction on current display devices. To this end, we develop a new, motion-aware disparity manipulation technique. The key idea behind it is to re-allocate the disparity range from regions that exhibit motion parallax to static parts of the scene so that the overall depth perceived by observers is maximized. To evaluate effectiveness of our manipulations, we perform additional validation experiments which confirm that by taking motion parallax depth cue into account, the overall depth impression can be enhanced without extending disparity budget (Fig. 7.1). More precisely, we make the following contributions:

- we design and perform psychovisual experiments quantifying the joint contribution of binocular disparity and motion parallax to depth perception,
- propose a computational model that predicts apparent depth induced by these cues for complex scenes,
- develop a new motion-driven disparity manipulation for stereoscopic and multi-scope content.

7.1 Joint motion parallax and disparity model

Motion parallax and binocular disparity have been studied in separation. In contrast and in spirit of this thesis, we assess the joint influence of motion parallax and binocular disparity on depth perception. To this end, we propose a model which describes a relation between the strength of these cues and the depth impression they induce. A direct advantage of such an approach is that the cue fusion is an integral part of the model, and it does not need to be modelled separately. The model will allow us to

identify combination of motion parallax and disparity values that result in a similar depth impression. We will use this equivalence in our applications to reduce binocular disparity in regions where depth perception is well supported by motion parallax.

7.1.1 Methods

To acquire data for building our model, we conducted a perceptual experiment where subjects were asked to match two stimuli according to the depth they provided. One of the stimuli contained both motion parallax and binocular disparity, while the other had only a disparity signal. The matches provided by users allowed us to find the disparity signal that is equivalent to a given parallax-disparity combination.

Stimuli The stimuli were random-dot stereograms with sinusoidal corrugations in depth. Both disparities and luminance-contrast values were modulated by a Gaussian function (Fig. 7.2, left). The spatial depth corrugation frequency was 0.4 cpd – the peak sensitivity of the HVS to both binocular disparity and motion parallax [Rogers and Graham 1982]. The luminance pattern was a random pattern with a uniform intensity distribution, which was additionally filtered using a low-pass filter with cutoff of frequencies at 10 cpd. To avoid texture depth cues that could provide additional information about the presented shape, the pattern was flatly projected on the surface. Stimuli include also a central black spot of 0.01 degrees to aid fixation. In our experiment, we used two types of stimuli: *dynamic* stimuli with both motion parallax and disparity, and *static* stimuli with disparity but without motion parallax. The dynamic stimuli were translated horizontally across the screen (Fig. 7.2, right) with different speeds.

We parameterize the stimuli using their relative depth from motion parallax $m = \Delta f/f$, binocular disparity d , and retinal velocity V (Fig. 2.3), i.e., (d_i, m_i, V_i) . A static stimulus is a special case of the dynamic one, i.e., $(d_i, 0, 0)$. Two possible parameterizations can be considered for the velocity: the absolute ocular velocity of the stimuli pursuit on the screen $\nu = d\alpha/dt$, or the relative retinal velocity of peaks and troughs in the stimuli $V = d\theta/dt$ (both expressed in arcmin/s). Although ν provides more straightforward control during the stimuli generation and does not depend on m , we argue that the latter is better at describing the HVS behavior. For large disparity amplitudes, small absolute velocity ν may result in clearly visible motion parallax. At the same time, the same absolute velocity may be insufficient to produce visible parallax for small disparities. In contrast, the relative velocity V does not have this problem. Consequently, we use V in the stimuli parameterization.

Note that stimuli can express different combinations of motion parallax and binocular disparity through the experiments, which do not need to be consistent. We specifically seek to measure the effect of non-consistent combinations, namely compressed and expanded disparity which is important for applications such as those shown in Sec. 7.2.

Equipment Stimuli were presented on a Zalman ZM-M240W polarized-glasses 3D display, with a screen size of 24" and a spatial resolution of 1920×1080 pixels, at a viewing distance of 60 cm.

Participants Twenty-seven subjects (5 F, 22 M, 24 to 27 years of age) that had normal or corrected-to-normal vision and passed a stereo-blindness test, took part in the experiment.

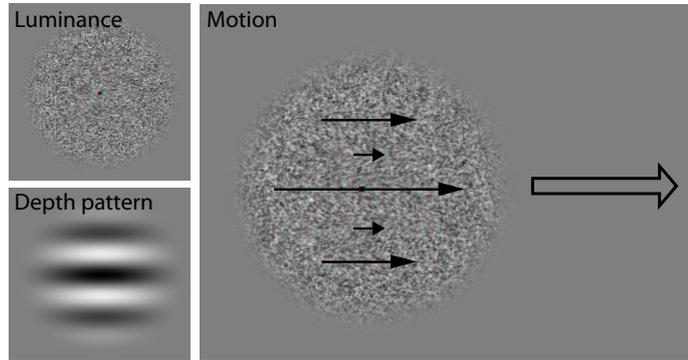


Figure 7.2. Stimulus used in our experiment. *Left:* Luminance and depth of our stimuli. *Right:* Anaglyph version of the same. For the dynamic stimulus variant, translating horizontally as indicated by the large arrow, motion parallax results in the flow depicted by the small arrows.

Procedure During each trial, subjects were given a pair of static and dynamic stimuli. They were instructed to fixate onto the marked central spot and adjust the disparity of the static stimulus until the perceived depth matched the one of the dynamic stimulus. This was done by pressing left and right keyboard keys. The stimuli were displayed sequentially and subjects could switch between them at will by pressing a key. A 500 ms-blank screen followed each switch. When switching to a dynamic stimulus, visibility was enforced for at least 1000 ms before switching back to a static stimulus to guarantee that the subject had time to observe the motion and did not judge depth based on binocular disparity only. For each dynamic stimulus each subject performed two trials. Performing the complete set of measurements took around 45 minutes per subject.

The three-dimensional stimulus space was covered by a set of samples combined from two subsets: In the first one, five regular steps of disparity and motion parallax were chosen from a typical range of values: $d \in [0..20]$ arcmin and $m \in [0..0.1]$, resulting in 25 samples. Stimulus translation velocity of 500 arcmin/s was chosen as it was observed to be safely suprathreshold in a pilot experiment.

Since the existence of detectable motion is a necessary condition for the perception of depth from motion parallax — five additional samples with stimulus translation velocity that varied from 0 to 300 arcmin/s were added for three disparity and motion parallax combinations (0, 0.025, .), (10, 0.075, .), and (5, 0.100, .). These points were chosen to lie in the part of the sampling space where the disparity is smaller than it would be in the real world for an object defined by the given motion parallax, and therefore the contribution of parallax to the estimated depth can potentially be significant. This helps us to derive the threshold velocity required for motion parallax detection. As a result additional 15 samples are measured and for two repetitions the experiments totals 80 samples per subject.

7.1.2 Data analysis and model fitting

The data from the experiment was used to derive a model of perceived depth due to a combination of motion parallax and disparity. The model maps from a joint stimulus involving both the motion and disparity cues to the matching depth that is perceived

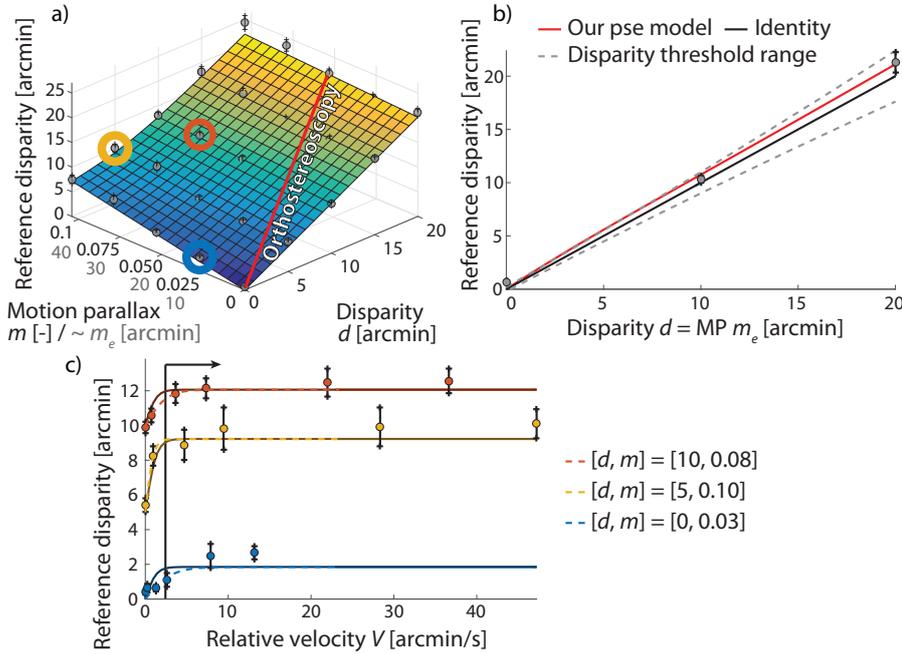


Figure 7.3. *a)* Data acquired from our experiment (gray dots) and our model showing the matched static disparity (vertical axis) as a function of binocular disparity and motion parallax (x and y axis). The red line shows the points corresponding to real-world observation conditions ($d = m$), plotted in *b)*. *b)* Comparison of measured points (gray dots), our model (red line) and the theoretical identity (blue line) for real-world observation conditions ($d = m_e$). Dashed lines mark deviation from identity within the disparity discrimination thresholds for a given disparity [Didyk et al. 2011]. Difference under the threshold range confirms that our model does not predict more depth for a moving stimulus than for a static stimulus if conflict of cues is avoided. *c)* Perceived depth as a function of relative velocity of motion V for three selected points from the two-dimensional space in *a)*. Dashed lines are fitted separately for each sequence while full lines were fitted jointly to build our model in Eq. 7.4. Corresponding points for the maximum velocity marked in *a)*. Vertical line is the 95th percentile threshold value.

using only disparity. We call this the matching static disparity.

We seek a function $\Phi : \mathbb{R}^4 \rightarrow \mathbb{R}$ of an equivalent static disparity for a combination of motion parallax m , binocular disparity d , velocity V , and angular distance s as a separation between neighboring peaks and troughs of our 0.4 cpd depth corrugation stimulus. We obtain the function by fitting an analytic expression to our experimental data. This is done in two steps: First, we fit a supra-threshold velocity model for d , m and the angular distance $s = S = 1.25$ arcmin that is fixed for our stimulus (Fig. 7.3a). Second, we model the effect of near-threshold velocities V using an additional, separable, multiplicative function which depends only on V and s (Fig. 7.3c). Effectively, we assume that the two steps are separable, i.e., the velocity influence is independent of the rest of parameters.

Supra-threshold velocity For supra-threshold velocity \hat{V} , stimulus angular distance S and each constant d , we express perceived depth as a function of m by fitting a linear

function to our data:

$$\Phi_{d,\hat{v},S}(m, \cdot) = am + b. \quad (7.1)$$

To obtain a model for varying d we interpolate the scale a and bias b . For b we use linear interpolation as it follows the identity prediction for depth perception as a function of disparity. However, a linear interpolation of the slope a would result in unreliable predictions in extrapolated regions. Instead, we use a normalized sigmoid function λ . Additionally, we enforce that in the absence of both motion parallax and binocular disparity, the perceived depth remains zero, i.e., $\Phi_{\hat{v},S}(0, 0) = 0$. Resulting function describing the perceived depth at supra-threshold level is given by:

$$\Phi_{\hat{v},S}(m, d) = (c_1 \cdot \lambda(d) + c_2) \cdot m + c_3 \cdot d, \quad (7.2)$$

$$\lambda(d) = \frac{2}{1 + e^{-d\beta}} - 1, \quad (7.3)$$

where $c_1 = -50.88$, $c_2 = 68.56$, and $c_3 = 1.006$ (DOF-adjusted $R^2 = 0.98$) (Fig. 7.3a). We set $\beta = 0.25$ to guarantee that the sigmoid saturates for the maximum disparity considered in the experiment ($d = 20$ arcmin) so that no extrapolation beyond our measurements occurs. The slope in the m -axis direction for $d \geq 20$ arcmin is then constant and equal to the sum $c_1 + c_2$. This determines the contribution of parallax to the total perceived depth for high disparity magnitudes. The value c_2 determines the slope of $\Phi_{\hat{v},S}(m, d)$ as a function of m for $d = 0$, that is, the effect of motion parallax as the only available depth cue in the absence of binocular disparity. The value of c_3 determines the slope in the d -axis direction for $d \geq 20$ arcmin. As its value is close to one our model provides safe extrapolation for larger disparities by matching each d to itself. However, the extrapolation for motion parallax magnitudes outside of the range we measured ($m > 0.1$) would be dangerous as no conclusions about further behavior of the function can be made based on our measurements. Therefore, for our applications, we clamp m to the range 0 to 0.1, which prevents a possible exaggeration of the contribution that larger parallax has to the depth perception.

Given the knowledge of absolute depth we can convert m to the equivalent disparity m_e in arcmin and then find orthostereoscopic points where $m_e = d$. Such conditions remove depth cue conflicts and therefore should create similar perception as the static stimulus. Fig. 7.3b shows that the deviation between our measured model and identity to a static reference is within the range defined by disparity discrimination thresholds at different disparity levels for depth corrugation stimuli of 0.4 cpd [Didyk et al. 2011].

Near-threshold velocity Fig. 7.3c shows the resulting matching static disparity for different velocities of motion for the three points sampled in the disparity and motion parallax space. Measurements show, how, for very low values of relative motion V , the HVS will not be able to perceive it and thus binocular disparity is the only depth cue present. Once a certain value of V is reached, motion is perceived, and motion parallax depth cue is triggered. We model this effect using a coefficient function $\Psi(\cdot)$ that describes visibility of motion parallax. It scales its effect between zero (perceived depth is determined by binocular disparity only) and our supra-threshold model:

$$\Phi(m, d, \Psi(\cdot)) = d + \Psi(\cdot) \cdot (\Phi_{\hat{v},S}(m, d) - d) \quad (7.4)$$

We first approximate $\Psi(\cdot)$ by fitting a sigmoid function to our measurements (Fig. 7.3c) as:

$$\Psi_0(V) = \frac{2}{1 + e^{-V \cdot c_4}} - 1 \quad (7.5)$$

with $c_4 = 1.553$ (DOF-adjusted $R^2 = 0.98$). Our velocity model saturates for $V > 3.0$ arcmin/s which is similar to differential motion thresholds of 3.2 arcmin/s as measured for moving bars in the fovea [McKee and Nakayama 1984, Fig. 2]. As speed measurements close to zero are unreliable the sigmoid function could lead to an overestimation of perceived depth for a near-static stimulus. To address this we threshold the sigmoid at 95 percentile, and obtain:

$$\Psi_1(V) = \begin{cases} 1 & \text{when } V \geq v_c, \\ 0 & \text{otherwise} \end{cases} \quad (7.6)$$

where $v_c \approx 2.36$ (Fig. 7.3c).

The retinal velocity was measured for sinusoidal depth corrugations with spatial frequency 0.4 cpd. To extend this for other corrugation frequencies, we rely on a study by Kee and Nakayama[1984, Fig. 5]. They have shown that thresholds of differential motion grow approximately linearly as a function of eccentricity. Given our experimental setup, function Ψ_1 is valid for eccentricity $S = 1.25$ deg. We extend its definition to arbitrary values of s by including the linear relation between the threshold and eccentricity. Consequently, we obtain:

$$\Psi(V, s) = \Psi_1\left(\frac{1.25}{s}V\right), \quad (7.7)$$

and state $\Psi(\cdot) = \Psi(V, s)$.

7.1.3 Discussion

We compare our measurements to those in the existing literature that have analyzed perceived depth induced by motion parallax in the absence of binocular disparity ($d = 0$), and find that they are consistent. In particular, Nawrot et al. [2014] report a foreshortening of perceived depth in the presence of motion parallax for similar viewing conditions. Similarly, Durgin et al. [1995] describe foreshortening of depth from motion parallax in the absence of binocular disparity, although using a substantially different procedure.

The part of the space between $d = 0$ and $d = m$ (the left side with respect to the red line in Fig. 7.3a) corresponds to the region where binocular disparity is compressed with respect to the geometrically correct motion parallax; in this region, depth induced by motion parallax increases the perceived depth. This perceived depth enhancement can be used to reallocate the binocular disparity budget, which is typically limited in practical applications, as shown in Sec. 7.2. When we express the motion parallax m in means of equivalent disparity m_e (the gray labels in Fig. 7.3a) we see that in an extreme case with zero binocular disparity ($d = 0$), motion parallax $m_e = 40$ arcmin alone induces depth corresponding to a static disparity of 7.5 arcmin or $40/7.5 \approx 20\%$ of veridical (geometric) depth.

The rest of the space, that between $d = m$ and $m = 0$ (the right side with respect to the red line in Fig. 7.3a), depicts the case when disparity is expanded with respect to its

original magnitude. In this case, motion parallax reduces perceived depth. Note that since motion parallax corresponds to the actual geometric configuration, it always drives the perceived depth towards its veridical value, irrespective of how binocular disparity is distorted. This is in accordance with modified weak fusion model [Landy et al. 1995], in which relative weights of interacting cues are dependent on their reliability.

7.1.4 Definition for points and pixels

The model presented in the previous section measures the contribution of motion parallax to the perceived depth of a scene in the presence of binocular disparity. In this section, we describe how this model that was derived from simple stimuli can be applied to obtain perceived depth between image pixels. The input is a single animation frame $I : \mathbb{N}^2 \rightarrow \mathbb{R}^3$, and as the output we seek a measure of perceived depth between pixels. We demonstrate the method for computer generated content, for which all scene information is available. The generalization of our technique to natural video streams is discussed in Sec. 7.3.2.

Motion parallax is well-defined for a pair of points (Fig. 2.3). However, its definition is usually limited to cases where the viewing position changes, yielding a globally-consistent motion of the scene. We extend this definition by observing that motion parallax introduces a depth impression in all regions of the scene that exhibit rigid motion. Therefore, to compute depth from both motion parallax and binocular disparity in our image I , we first need to determine whether two pixels, and thus two points in the scene, follow the same rigid transformation.

To detect whether two image pixels, \mathbf{x}_A and \mathbf{x}_B , undergo the same rigid transformation, we first obtain for each pixel \mathbf{x} its 3D world position $\mathcal{P}(\mathbf{x})$, and a transformation matrix $\mathcal{M}(\mathbf{x})$ describing its positions change over time. For the computer generated content, this can be easily done by storing a position buffer and transformation matrices for each pixels. For natural content, one needs to perform 3D scene reconstruction. In the following text, we assume that for each pixel in the image, we have given the two functions $\mathcal{P} : \mathbb{N}^2 \rightarrow \mathbb{R}^4$ expressed in homogeneous coordinates, and $\mathcal{M} : \mathbb{N}^2 \rightarrow \mathbb{R}^{3 \times 4}$ providing a transformation matrix for each pixel.

Rigidity of the transformation of the two points \mathbf{x}_A and \mathbf{x}_B can be checked by comparing transformation matrices $\mathcal{M}(\mathbf{x}_A)$ and $\mathcal{M}(\mathbf{x}_B)$. However, a direct comparison of the two matrices is usually unstable and not robust due to small inaccuracies. Inspired by work of Braunstein et al. [1990], we propose a different measure of rigidity. Given two pixels \mathbf{x}_A and \mathbf{x}_B , we compare the results of applying their transformation matrices $\mathcal{M}(\mathbf{x}_A)$ and $\mathcal{M}(\mathbf{x}_B)$ to a set of n different pixels $\mathcal{N} = \{\mathbf{x}_i \in \mathbb{R}^2 : 1 \leq i \leq n\}$, and compute:

$$\gamma(\mathbf{x}_A, \mathbf{x}_B) = \frac{1}{|\mathcal{N}|} \sum_{\mathbf{x}_i \in \mathcal{N}} \|\mathcal{M}(\mathbf{x}_A) \mathcal{P}(\mathbf{x}_i) - \mathcal{M}(\mathbf{x}_B) \mathcal{P}(\mathbf{x}_i)\|_2. \quad (7.8)$$

To reliably measure the difference between two transformations, \mathcal{N} should contain at least three points that are not co-linear. In our case, we use a 3×3 pixel neighborhood of \mathbf{x}_A as \mathcal{N} . We assume that the two matrices correspond to the same rigid motion if $\gamma(\mathbf{x}_A, \mathbf{x}_B)$ is small, thus defining the final rigidity measure for the pixels as:

$$\Gamma(\mathbf{x}_A, \mathbf{x}_B) = \begin{cases} 1 & \text{when } \gamma(\mathbf{x}_A, \mathbf{x}_B) < \epsilon, \\ 0 & \text{otherwise.} \end{cases} \quad (7.9)$$

This measure allows for a small amount of non-rigid motion, but in general, ϵ should be small to tolerate only small differences between both transformations. In our applications we use ϵ as 0.001 times the diameter of the \mathcal{P} value domain bounding sphere.

Once we have determined rigidity Γ , in order to apply our model to the pixels, we need to know their velocities and spatial distances as well as their depth and pixel disparity. For the pixel disparity we are interested in its per-pixel angular measure (i.e., vergence angles, Fig. 2.3) $\mathcal{D}_d : \mathbb{N}^2 \rightarrow \mathbb{R}$. This is computed taking into account the viewing distance, and the interocular distance of a standard observer (6.4 cm). For velocities, we compute absolute velocities, $\mathcal{D}_v : \mathbb{N}^2 \rightarrow \mathbb{R}^2$, measured in visual angles, and similarly we express the angle between screen-space locations of both points $\mathcal{D}_s : \mathbb{N}^4 \rightarrow \mathbb{R}$. The linear depth $\mathcal{D}_m : \mathbb{N}^2 \rightarrow \mathbb{R}$ is only used to express the motion parallax m and thus needs only to be known up to a factor.

Given these, we can apply our parallax visibility model to a pair of pixels to obtain the contribution of this cue between them as:

$$\zeta(\mathbf{x}_A, \mathbf{x}_B) = \Psi(\Gamma(\mathbf{x}_A, \mathbf{x}_B) \cdot \Delta\mathcal{D}_v, \Delta\mathcal{D}_s), \quad (7.10)$$

and our full model to obtain the perceived depth between them as:

$$\Theta(\mathbf{x}_A, \mathbf{x}_B) = \Phi(\Delta\mathcal{D}_m, \Delta\mathcal{D}_d, \zeta(\mathbf{x}_A, \mathbf{x}_B)), \quad (7.11)$$

where:

$$\begin{aligned} \Delta\mathcal{D}_m &= \frac{|\mathcal{D}_m(\mathbf{x}_A) - \mathcal{D}_m(\mathbf{x}_B)|}{\max(\mathcal{D}_m(\mathbf{x}_A), \mathcal{D}_m(\mathbf{x}_B))}, \\ \Delta\mathcal{D}_d &= |\mathcal{D}_d(\mathbf{x}_A) - \mathcal{D}_d(\mathbf{x}_B)|, \\ \Delta\mathcal{D}_v &= \|\mathcal{D}_v(\mathbf{x}_A) - \mathcal{D}_v(\mathbf{x}_B)\|, \\ \Delta\mathcal{D}_s &= \mathcal{D}_s(\mathbf{x}_A, \mathbf{x}_B). \end{aligned} \quad (7.12)$$

Note that, according to the definitions above, $\Delta\mathcal{D}_d$ is disparity between \mathbf{x}_A and \mathbf{x}_B (d in the model as given by Eq. 7.4), $\Delta\mathcal{D}_m$ is our measure for motion parallax between both pixels (m in the model), the relative velocity $\Delta\mathcal{D}_v$ between them (V in the model) is a difference of per-pixel optical flow vectors and $\Delta\mathcal{D}_s$ between them (s in the model) is directly their angular distance.

With the definition in Eq. 7.11 we can compute the relative depth between any two pixels of the input image sequence. The application of this to complex images for disparity manipulation is presented in the next section.

7.2 Our approach

While many techniques exist for disparity manipulations [Lang et al. 2010, Didyk et al. 2011, 2012, Masia et al. 2013a], none of them takes motion parallax information into account. As shown in our experiments, motion parallax greatly contributes to the overall depth impression. The contribution is especially significant for compressed disparities where the contribution of motion parallax and binocular disparity to perceived depth is comparable. In this section, we propose a new disparity mapping operator (Fig. 7.4) which takes an advantage of our motion parallax model. The goal of the technique is to suppress disparities that are supported by motion parallax, and to use freed disparity range to expand disparities in static regions. As our goal is only to reallocate disparities, the technique can be easily combined with other existing disparity mapping approaches by applying them on our input disparity map $\mathcal{D}_d(\mathbf{x})$ beforehand.

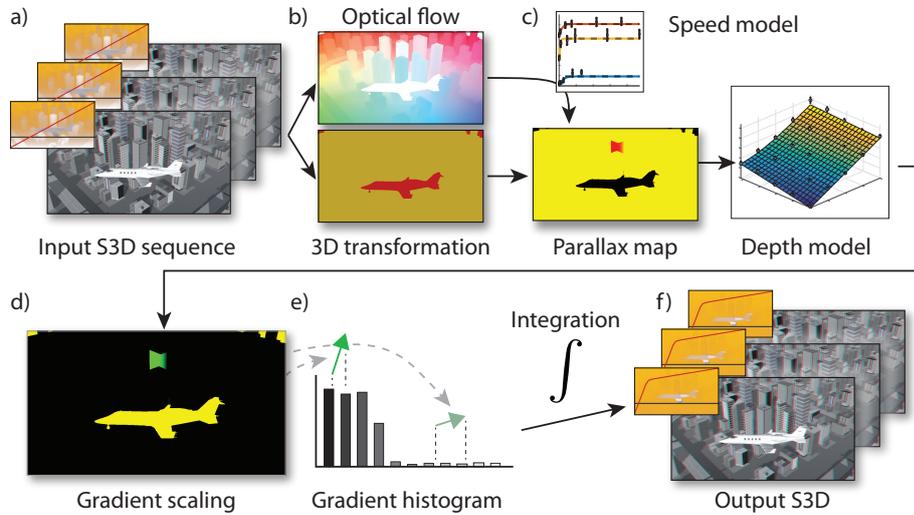


Figure 7.4. Our disparity mapping algorithm. Optical flow and 3D transformations (b) are extracted for the input S3D sequence (a) and used together with our relative motion model to predict perceptibility of motion parallax for X and Y gradients (c). Our model of perceived depth is then sampled to estimate necessary scaling of each disparity gradient (d). Scaled gradients are accumulated in a histogram (e) and integrated to construct a new mapping curve that is applied to produce the final S3D content (f).

7.2.1 Overview

The input to our technique is a stereoscopic sequence with disparities mapped to fit a desired range. In the first step, we extract optical flow and 3D transformation matrices for each pixel of the input S3D sequence (Fig. 7.4a) either by means of computer vision or by directly dumping necessary information from the rendering pipeline (Fig. 7.4b). Second, we detect whether the transformation between neighboring pixels is both rigid and sufficiently fast according to our velocity model. This way the visibility of motion parallax for each pair of neighboring pixels is predicted (Fig. 7.4c). Next, we use our model to estimate contribution of the parallax to the perceived depth, and redistribute depth budget by scaling local disparity gradients accordingly, and storing them in a histogram (Fig. 7.4d). We then, reconstruct a global disparity mapping curve from the histogram (Fig. 7.4e). As the last step, we remap the input disparity, and use an image warping technique to generate a new S3D sequence (Fig. 7.4f). In the following subsection, we describe details of each of these steps.

7.2.2 Parallax map

The parallax map encodes the relative visibility of motion parallax for a pair of pixels (Fig. 7.4c). Such pixels must have both a rigid mutual transformation and a relative motion speed above the visibility threshold.

For the purpose of this technique, we consider all effects in a single-scale fashion, and process local disparities and local parallax as both horizontal and vertical gradients Δx between neighboring pixels \mathbf{x}_A and \mathbf{x}_B . Consequently, we refer to differential measures from Eq. 7.12 using a short notation $\Delta \mathcal{D} = \Delta \mathcal{D}(\mathbf{x}_A, \mathbf{x}_B)$. We process both vertical and horizontal gradients in the same way, and omit this information further on.

The rigidity of two neighboring pixels can be directly evaluated as $\Gamma(\Delta\mathbf{x}) = \Gamma(\mathbf{x}_A, \mathbf{x}_B)$ (Eq. 7.9). However, the visibility of motion between the two neighboring pixels may be too small to qualify for a visible parallax according to $\Psi_V(V)$. This, however, is not in line with motion perception of larger objects. The largest sensitivity for differential motion was detected for targets of 1 deg size [McKee and Nakayama 1984]. Therefore, instead of evaluating rigidity only for neighboring pixels, for each pixel \mathbf{x} , we search a local neighborhood for a rigidly connected location \mathbf{x}' with the highest visibility of motion parallax $\zeta(\mathbf{x}, \mathbf{x}')$ (Eq. 7.10):

$$\mathbf{x}' = \max_{\mathbf{x}'} \zeta(\mathbf{x}, \mathbf{x}'). \quad (7.13)$$

The parallax map M at location \mathbf{x} is then a product of both the local rigidity between neighboring pixels and the maximum motion parallax visibility (Fig. 7.4c):

$$M(\mathbf{x}) = \Gamma(\Delta\mathbf{x}) \cdot \zeta(\mathbf{x}, \mathbf{x}'). \quad (7.14)$$

7.2.3 Disparity scaling

We seek disparity manipulation that provides images such that when both binocular disparities and motion parallax information are combined, they give depth perception specified by the input disparities $\Delta\mathcal{D}_d(\mathbf{x})$. This can be formally expressed by the following constraint:

$$\forall_{\mathbf{x}} \Delta\mathcal{D}_d(\mathbf{x}) = \Phi(\Delta\mathcal{D}_m(\mathbf{x}), \Delta\mathcal{D}'_d(\mathbf{x}), M(\mathbf{x})), \quad (7.15)$$

where $\Delta\mathcal{D}'_d$ are disparities of the output image. Although enforcing this constraint suppresses only disparities that are supported by motion parallax, this increases the part of disparity range dedicated for static parts. It is, therefore, sufficient to rescale the resulting disparity range to the original one at the end of the process. We use Eq. 7.15 to compute target output gradients $\Delta\mathcal{D}'_d(\mathbf{x})$. As the function Φ is monotonic and all other variables are fixed, this can be done by numerically inverting it.

7.2.4 Curve construction

Once the constraint in Eq. 7.15 is enforced, a new vergence map \mathcal{D}'_d can be recovered from disparities $\Delta\mathcal{D}'_d$ by solving a Poisson equation in a similar way as it was done in [Fattal et al. 2002] for luminance. However, such a solution does not guarantee depth ordering preservation and temporal coherence. To address these issues, we propose to construct a global mapping curve based on $\Delta\mathcal{D}_d$ and $\Delta\mathcal{D}'_d$ using a custom histogram of gradients. To this end, we construct a histogram $H: \mathbb{I} \rightarrow \mathbb{R}$ that contains $n = 1024$ bins which split the range of \mathcal{D}_d . Each bin H_i stores information about values between a_i and b_i where:

$$\begin{aligned} a_i &= \min(\mathcal{D}_d) + \frac{i}{n} \cdot [\max(\mathcal{D}_d) - \min(\mathcal{D}_d)] \\ b_i &= a_i + \frac{1}{n} \cdot [\max(\mathcal{D}_d) - \min(\mathcal{D}_d)]. \end{aligned} \quad (7.16)$$

We construct the histogram H by distributing each gradient $\Delta\mathcal{D}'_d$ to all bins covered by the interval between values $\mathcal{D}_d(\mathbf{x})$ and $\mathcal{D}_d(\mathbf{x}) + \Delta\mathcal{D}_d(\mathbf{x})$. Effectively, we add to each of these bins the slope of the future mapping curve $\Delta\mathcal{D}'_d(\mathbf{x})/\Delta\mathcal{D}_d(\mathbf{x})$.



Figure 7.5. Results of our parallax-aware disparity manipulation applied to four video sequences used in our user study (Sec. 7.3.1). Two frames from the input sequences (1st and 2nd columns) compared to our outputs at Time = 1 (3rd column). The mapping curves along with the output disparity maps are shown in the insets. The blue color marks crossed disparity, yellow uncrossed disparity and white zero disparity. The arrows show the local motion direction.

For each gradient, we also accumulate weights w into equivalent histogram $W : \mathbb{I} \rightarrow \mathbb{R}$. They corresponds to the sensitivity of the HVS to the changes in disparities. This favors small gradients to be preserved. Formally, to each gradient $\Delta\mathcal{D}_d$, we assign the following weight :

$$w(\Delta\mathcal{D}_d) = \frac{1}{thr(\Delta\mathcal{D}_d)} \quad (7.17)$$

where $thr(d)$ is a disparity discrimination threshold for pedestal disparity d [Didyk et al. 2011, Sec. 4]. The idea is that compression of large disparities has a smaller effect on the overall depth perception as these are likely to stay noticeable. For our method this also gives an additional freedom for redistributing depth between distant objects.

After all gradients are accumulated in H , we normalize the value in each bin by the accumulated weights in W . Then, we compute the remapping between input pixel disparities \mathcal{D}_d and new disparities \mathcal{D}'_d (both expressed in vergence angles) as a cumulative sum of the histogram, which is normalized to the range of input disparities \mathcal{D}_d . Formally, the remapping function R can be defined as follows:

$$R(d) = [\max(\mathcal{D}_d) - \min(\mathcal{D}_d)] \cdot \frac{\sum_{i=0}^{h(d)} H_i/W_i}{\sum_{i=0}^{n-1} H_i/W_i} + \min(\mathcal{D}_d), \quad (7.18)$$

where the function $h(d)$ provides the index of the bin that corresponds to the value of d . Please note that if no gradients were overlapping in the histogram, the mapping constructed in such a way would provide $\Delta\mathcal{D}'_d$ that fulfills our initial constraint (Eq. 7.15). However, due to a large number of gradients which overlap, the mapping only tries to satisfy all of them. After the vergence information is remapped according to the above equation, we recompute pixel disparity information and use it for modifying the input frames. This last step is done by using image-based warping technique to generate new stereoscopic images [Didyk et al. 2010b]. Examples of the result are presented in Fig. 7.5. The entire algorithm in our non-optimized implementation is executed in real-time for the input resolution 1280×720 pixels on a computer with Intel Xeon E5 and GeForce GTX 980 Ti.

Although histogram-based approaches typically feature better temporal coherence than local processing, we noticed that minor depth oscillation may be observed due to the normalization step. The coherence is also affected by objects leaving the scene. We improve the temporal coherence by an additional filtering of the remapping curves over time. Usually, a window of ten last frames is used. Similar techniques have been applied before in a similar context [Oskam et al. 2011, Didyk et al. 2012].

7.2.5 Application to autostereoscopic displays

One of the main drawbacks of current autostereoscopic displays is a significant angular aliasing when a large depth range is shown on such a screen [Zwicker et al. 2006]. We utilize the depth from motion parallax and instead of enhancing depth in other regions we remove it from the scene completely thus reducing the required display disparity range and thus, the amount of visible aliasing. To this end, we follow the same algorithm as before with the exception of the final histogram construction. Here, we do not map the resulting curve into the fixed depth range, but we directly use the accumulated gradient values that encode the disparity just needed to deliver perception equivalent to a static image. We then construct the mapping curve directly as:

$$R_a(d) = [\max(\mathcal{D}_d) - \min(\mathcal{D}_d)] \cdot \sum_{i=0}^{h(d)} H_i/W_i + \min(\mathcal{D}_d). \quad (7.19)$$

We have tested our approach on a Full HD display Tridality MV2600va that utilizes parallax barrier to deliver 5 views. Refer to the Fig. 7.6 for a captured screen comparison. We compared a linear mapping without and with our manipulation on the top. Disparities of both results are centered around the screen to minimize the aliasing of the display [Zwicker et al. 2006]. While the reference footage suffers from aliasing in the front and far parts of the scene, our output was able to convey a very similar depth impression without violating the usable depth range and introducing visual artifacts due to severe aliasing.

7.3 Validation

7.3.1 Perceptual studies

Our experiment in Sec. 7.1 accounts for the combination of motion parallax and disparity cues but omits other depth cues that are present in complex images [Cutting 1995]. To test the applicability of our model in such conditions we have performed a validation

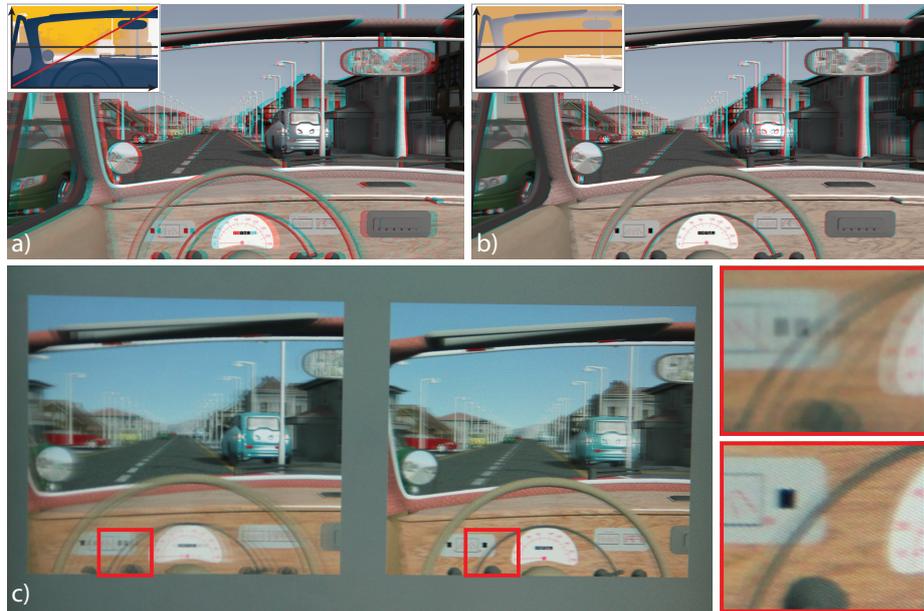


Figure 7.6. Example of disparity compression for autostereoscopic displays. a) The original linear compression with the corresponding disparity map and mapping curve shown in the inset. b) Our manipulation applied on the top of (a) to compress stronger regions that benefit from motion parallax. c) A photo of our display. Aliasing artifacts are strongly present in the content with only linear mapping (insets).

user study. We check if our method increases the 3D appearance of stereoscopic content by showing complex video sequences to study participants.

Method validation

Stimuli Four short video sequences (5-10 seconds) with a camera or scene motion (see Fig. 7.5) were used as stimuli. The stereoscopic videos that had disparities processed by our technique and the original input videos were played in a loop simultaneously side-by-side in a random order.

Task Participants were given an unlimited time to compare both videos and answer a question “Which sequence is more 3D?”. Each pair was shown twice resulting in 8 trials.

Equipment The stimuli were presented using the polarized glasses technology on a 24” Zalman ZM-M240W display with a resolution of 1920×1080 pixels, at a viewing distance of 60 cm under normal, controlled office lighting. We used this display technology as it does not introduce temporal disparity artifacts potentially caused by a time-sequential presentation. A chin rest was employed to prevent the participant from moving away from the optimal viewing angle.

Participants 12 participants (3 F, 9 M, 20 to 30 years old) took part in the study. All of them had normal or corrected-to-normal vision and passed a stereo-blindness test

by describing the content of several random dot stereograms (RDS). The subjects were naïve to the purpose of the experiment.

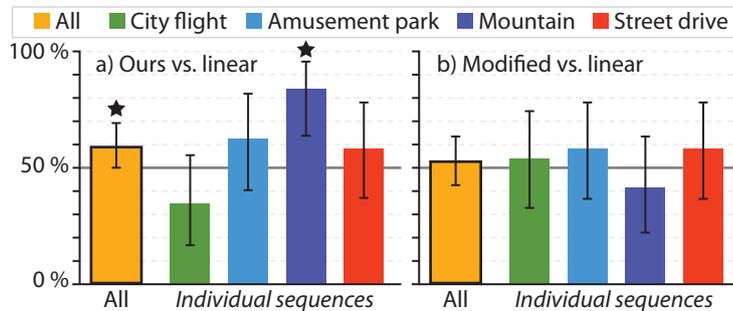


Figure 7.7. Results of our validation study. a) A comparison of linear and our mapping (Sec. 7.3.1). b) A comparison of linear and gradient adjusted mapping (Sec. 7.3.1). Confidence intervals are shown by the error bars and a significance in binomial test for $p < 0.05$ by star symbols.

Results The results of the study are presented in Fig. 7.7a. We observe a statistically significant (binomial test, $p < 0.05$) 60.0% preference of the 3D reproduction in our manipulated results. This confirms that our method succeeds in providing more depth impression by redistributing the same disparity budget. The results are affected by the poor performance in the *City flight* scene (Fig. 7.5, 1st row). We suspect that the relatively small screen coverage of the plane and large attention given to the city in the background caused many people to ignore the depth extension of the plane completely, and instead, focus on detecting small differences in the city. As our method aims to redistribute the disparity but does not change the parallax the total sum of both cues is indeed smaller in the moving part if investigated in isolation which was further facilitated by side-by-side presentation of the stimuli. A saliency predictor could possibly tackle this problem but we decided not to include it into our pipeline to keep it clear and focused.

Modification

As the mapping curve predicted by our method forced some regions of the scenes to be almost flat, we wanted to validate whether such behavior is appropriate. The goal of the following experiment was to see whether the scene can benefit, if a small disparity information is present everywhere. To this end, we repeated the validation study with a modified version of our method. Inspired by Larson et al. [1997], we have adapted their iterative histogram adjustment algorithm, and used their constrain to guarantee that our mapping curve will never be less steep than a linear mapping to a target disparity range of 2.3 arcmin. This choice was motivated by the fact that most people (97.3%) have stereoacuity below this value [Coutant and Westheimer 1993]. This effectively guaranteed that most participants of our experiment should experience at least some stereo impression for such stimuli. In practice, small disparity gradient was added to originally flat regions (Fig. 7.8).

Stimuli, Task, Equipment and Participants Stimuli presentation, task and equipment stayed the same as in Sec. 7.3.1 up to modification to our method described above.

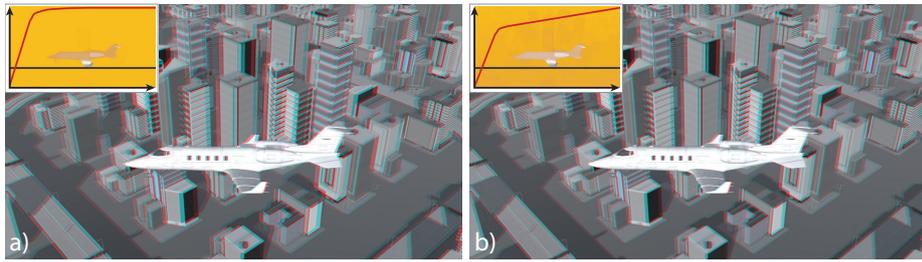


Figure 7.8. A comparison of our proposed algorithm (a) and its modification evaluated in Sec. 7.3.1 (b).

Partially overlapping group of 12 participants (4 F, 8 M, 20 to 30 years old) took part in the study.

Results The results are closer to chance level than previously (Fig. 7.7b). This demonstrates that the very small amount of disparity present in some of our stimuli in the original experiment was justified. For most of the scenes, the depth impression was reduced when the just-noticeable disparity was added to some regions, which effectively reduced this information in other parts of the scene. Our model could capture this balance well.

7.3.2 Discussion and limitations

Our experiment in Sec. 7.1 has been performed for isolated motion parallax and disparity cues, while we apply its outcome for complex images. Following the modified weak cue integration theory [Landy et al. 1995], which conforms well with many experimental data involving motion parallax [Howard and Rogers 2012, Ch. 30.2], the agreement of motion parallax cue with 2D pictorial cues should result in its possible promotion. This makes our model conservative in the sense of predicting the minimal amount of perceived depth enhancement due to motion parallax.

Our model is also conservative with respect to our strong assumption on the transformation rigidity (Sec. 7.1.4), which ignores any motion parallax that can arise from non-rigid relative object motion [Luca et al. 2007]. We relegate as future work the extension of our experimental model (Sec. 7.1) to handle such non-rigid relative motion.

We described our method for detection and modeling depth from motion parallax for cases, where all information about the scene is provided as an input. It is possible to apply our technique also to natural stereoscopic video sequences, which requires deriving all needed information using computer vision methods. For example, pixel disparity can be estimated using standard stereo correspondence methods such as [Brox et al. 2004, Zimmer et al. 2011]. Computing other information, such as depth, positions, motion flow is related to structure-from-motion techniques. While computing this information is non-trivial a number of techniques exist, which could be used to estimate the information [Szeliski 2011]. We relegate deeper investigation of the suitability of such method for our purposes to future work.

Our model has been measured assuming a perfect match between the virtual camera and actual observation setups in terms of the field of view, the observer distance, and so on. In case of significant departs from such measurement conditions, e.g., in cinematographic applications, our model might not optimally predict the magnitude of

perceived depth, and possibly new measurements would be required to accommodate relevant view configurations as well as disparity and motion parallax ranges in such conditions. Nevertheless, as discussed in Sec. 7.1.2 our model is protected against such out of range queries and should always lead to a predictable behavior.

7.4 Conclusions

We presented a method for predicting the impact of motion parallax on scene layout perception. To this end, we first conducted a psychovisual experiment in which we measured depth induced by motion parallax and related it directly to depth obtained from binocular presentation of a static stimuli. Based on these measurements, we proposed a computational model that predicts the induced depth for complex image sequences. To our knowledge, this is the first work that tries to analyze and quantify structure perception from motion parallax for complex image sequences. A big advantage of our model is the compatibility with previously proposed image and video processing techniques. As a result, it can be easily incorporated in those techniques. We demonstrated this on several examples. Our model is also a significant step towards better understanding of perception for new output devices such as head-mounted or lightfield displays where motion parallax is inherent cue obtained from observers' head movements.

Part II

Stereoscopic 3D and HDR

Chapter 8

Disparity perception in photopic vision

When we talk about disparity processing we usually assume that disparity is represented as a single scalar map that assigns a pixel disparity to each element of the image. Such a definition assumes that there is always exactly one pair of corresponding locations in the left and the right image and also that it can be derived from the physical scene depth alone. This is truth for the most simple diffuse opaque material model. However, in the real world materials are often metallic, shiny, transparent or translucent or a mixture of all. Reflections and refractions from such surfaces or even volumes are very important for the overall appearance and realism of an image. This is especially the case in day-like photopic conditions when the vision acuity is high and light magnitude sufficient to fully illuminate the scene.

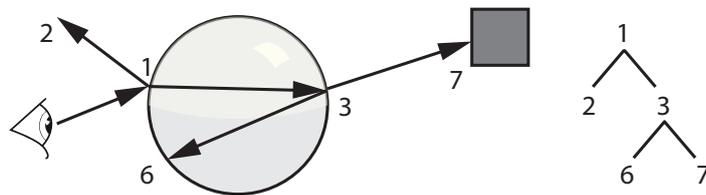


Figure 8.1. Ray tracing tree and our breadth-first search indexing of nodes used for a linear representation of the tree. The first diffuse intersection (1) is followed by both the reflection (2) and the refraction (3). As the reflected ray does not intersect another object it will not contribute to the final sum. The refractive ray reflects (6) and refracts (7) further and the latter intersects a solid diffuse object. Further recursion is omitted for clarity.

Reflections and refractions can be visualized using a ray-tracing tree (Fig. 8.1). Each node contributes by additional luminance information that is composed to the final image. As these effects are view dependent the entire tree with its node can change significantly for viewpoints of both eyes. This way the brain not only receives two different images like in case of diffuse surfaces but also two images that compose of many layers each with its own disparity. This can easily be understood for a planar mirror which works as a virtual extension of the scene. Therefore, objects seen in the mirror will have disparity depth equal to the sum of distance between the viewer and the mirror and the mirror and the object. A similar observation can be made for refraction of light when passing through a transparent surface. This means that disparity is no longer in a simple relation with the scene depth configuration. If the surface is curved the relation is more complicated and the complexity grows when we consider a larger tree depth.

This has several consequences for stereoscopy and the viewing comfort. First, the maximum disparity in the scene is no longer bounded by the maximum depth. Therefore traditional disparity mappers can have difficulty to properly fit the scene into a comfortable range for a given physical screen. Second, although a simple modification could be made if the effective disparity was considered instead of the one computed from surface depths, it is now required to process multiple disparities for each pixel. Third, having multiple disparities means that a user can see multiple layers at once and the disparity between them can be too large even if they all separately fit in the budget of the display. Finally, as the surface curvature can vary, reflective and refractive images can be distorted. Such distortions will typically differ between viewpoints and therefore two distinct images will be delivered to the HVS and fusion will be difficult or impossible. This yields a binocular rivalry which is a source of additional discomfort.

In this chapter we provide a ray tracing inspired model of multi-layer disparity. We propose a computer vision motivated algorithm for deriving individual disparity vectors for input stereoscopic images. We state an optimization problem that finds camera parameters for the scene rendering such that the visual appearance and realism are preserved but the viewing comfort is increased. We validate our approach in a user study. Please refer to our paper [Dąbala et al. 2014] for more details.

8.1 Our approach

Our goal is to generate stereoscopic images with reflections and refraction that are pleasant to look at, maintain a realistic appearance but do not violate stereoscopic rules which could lead to discomfort. In order to do that we first formulate a model of disparity that is suitable for the multi-layer image formation and then propose an algorithm to extract parameters of such model for an input image configuration. We formulate our requirements on the output as a cost function that we use in an optimization of local camera parameters to obtain a prescription for the rendering of the output.

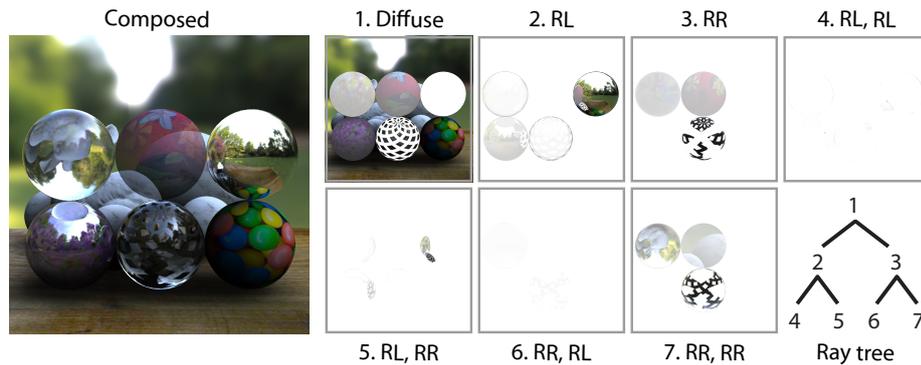


Figure 8.2. Decomposition of the image into layers of a ray tree. The labels encode the path of the ray. *RL* stands for a reflection and *RR* for a refraction. The tree scheme in the bottom right corner has reflections on the left side.

8.1.1 Disparity model

We base our model of disparity on ray tracing. We trace a ray of light for each image pixel in the direction from a camera towards the scene and on each intersection we obtain one disparity value that corresponds to the disparity between identical object locations in the two images (Fig. 8.1). This is analogous to rendering respective layer of the scene (Fig. 8.2), computing diffuse surface shading and evaluating disparities in the usual way. Up to two child rays are then traced, one for reflection and one for refraction. The recursion is stopped after a pre-defined tree depth is reached. This gives a fixed maximum size of disparity vector for each pixel and allows its storage in a 2D array texture where it can be efficiently sampled during further computations.

8.1.2 Disparity extraction

Although simple optical interfaces and their light transport can be described using closed form equations this becomes much harder when curved surfaces along a path containing multiple reflections and refractions are considered. We can illustrate this by a simple case of multiple flat mirrors each reflecting the same object when observed from a fixed camera location. We see that unlike in the diffuse case we observe multiple copies of each object at once. That also means that there are multiple possible matches when compared with an image from another viewpoint.

To solve this problem in a simple and robust way we adapt an approach known from the computer vision as a stereo matching. However, instead of matching often ambiguous color information we use additional information known from the rendering. A position buffer for the i -th tree node holds a 3D position of an object point before it was projected to the screen plane. Such position is unique as no other object can have the same location and be visible at the same time. Thanks to this we only need to match single pixel information to find corresponding object point in the other view image. To address possibility of multiple matches we detect all candidates whose 3D position differs below a threshold and select the one closest in the image coordinate space as it is most likely to be fused by HVS.

We also account for the limitation of HVS. First, we only search a neighborhood corresponding to the panum fusion area and, therefore, we only allow for small vertical disparities. Second, we also check if local patches of luminance in both images match to each other as this is required for depth inference. Finally, we require that the local luminance patch has contrast above the threshold needed to support stereoscopic fusion as observed by Cormack et al. [1991].

We repeat this process for every node of the ray tracing tree thus obtaining the closest match and corresponding disparity map D_i for every reflection and refraction node.

8.1.3 Cost function

To achieve comfortable presentation we need to limit absolute disparities from the screen as well as disparities between individual layers of the image. We also want to avoid rivalry by preventing strong unfuseable distortions of the image. At the same time we need to maintain the depth ordering and a clear layer separation to preserve image realism. Based on these requirements we formulate a cost function.

Absolute disparity term The absolute disparity from the screen has to be kept within a usable range to prevent the vergence-accomodation conflict or an aliasing if autostereoscopic display is used. We enforce this by minimizing all absolute disparities D_i with respect to the screen for each layer i and image location \mathbf{x} as:

$$f_a(\mathbf{x}, i) = |D_i(\mathbf{x})|.$$

Data term At the same time we want to avoid complete flattening of the scene. The data term prevents this by weakly forcing the optimized disparity to follow the original input disparity \hat{D}_i :

$$f_d(\mathbf{x}, i) = |D_i(\mathbf{x}) - \hat{D}_i(\mathbf{x})|.$$

Relative disparity term We want to avoid diplopia when two layers i and j are observed at the same time. To do that we enforce the relative disparity between them to be not larger than a chosen threshold γ :

$$f_p(\mathbf{x}, i, j) = \max(|D_i(\mathbf{x}) - D_j(\mathbf{x}) - \gamma|, 0).$$

We use $\gamma = 3$ arcmin as this is well above disparity perception threshold but at the same time low enough to prevent diplopia even if summed over multiple layers [Didyk et al. 2011].

Rivalry term Finally, we prevent rivalry by enforcing similarity between corresponding patches. We minimize the difference between corresponding luminance patches as:

$$f_r(\mathbf{x}, i) = l(\mathbf{x}_A, \mathbf{x}_B, i),$$

where $l(\mathbf{x}_A, \mathbf{x}_B, i)$ computes a difference of local patches around respective positions of both the left and the right eye images for the layer i using a sum of square differences (SSD). We consider patches of size 7×7 pixels.

8.1.4 Optimization

The final cost function is a weighted sum of all four terms $f = w_a f_a + w_d f_d + w_p f_p + w_r f_r$. The weights can be provided as constants or as an user input from our painting interface. In our results we use $w_a = w_p = 1$, $w_r = 10$ and $w_d = 0.1$ as this enforces proper layer ordering important for the realism but does not get strongly affected by extreme values of the input disparity map.

We optimize left and right eye camera offsets for every layer of the image. For the first (diffuse) layer this corresponds to setting up the camera interaxial distance in classical stereography. The other layers are then relative to the first one, hence zero offsets mean that additional layers collapse to the first diffuse one and all reflections and refractions are flat. The maximum offset that we consider is the one in the input image and represents physically correct rendering. We precompute disparities and luminances for all offsets of all layers and store them in a texture. We use a gradient descent to find the optimum offsets for each of the layers as a slice through our precomputed data that minimizes the cost function f . This is done for every image pixel independently. We enforce a spatial coherence using a depth edge-aware spatial blur of the final camera offsets for each layer i .

Finally we use the optimized camera offsets to render and compose a final stereoscopic image using ray tracing. See Fig. 8.3 for examples. As we base our method on ray tracing we also support non-traditional camera models. We demonstrate this on the example of the Eikonal light transport where the light travels through refractive material with spatially varying index of refraction such as a hot gas (Fig. 8.3, 3rd column). The only required modification is to define the first diffuse surface as the location of the first ray direction change. Then the ray is followed by ray tracing the volume with a constant step.

8.2 Validation

We have validated our approach in a user study. Users were presented with four different versions of the same stereoscopic image at the same time and they were asked to sort them by the ease of fusion, realism and the overall preference.

We have compared our method to the physically correct rendering with maximum camera offsets (see *Physical* in Fig. 8.3), completely flat reflections and refractions with zero offsets in all layers except the first one (*On-surface*) and small constant offsets simulating method of Templin et al. [2012] (*Near-surface*).

As expected our method was marked as significantly more realistic than both the on-surface flat and the near-surface competitors. The difference with respect to the physical rendering was on a chance level. In the case of the fusion difficulty our method performed significantly better than the physical rendering and it was on a chance level with the other two. Finally, the overall preference was significantly shifted towards our results in all three cases. This confirms the achievement of our goals as our results are easier to look at than the physically correct renderings, at the same time they are more realistic than the other two alternative approaches and also preferred in general.

8.3 Conclusions

In this chapter we discussed issues emerging from rendering view dependent effects in stereoscopic 3D. Specifically we proposed a modeling framework and an optimization scheme for the treatment of reflective and refractive materials. Our method has managed to improve the viewing comfort by reducing excessive disparity values both with respect to the screen and between individual layers of the image composition.

A reduction of rivalry then made the fusion of our images much easier than in the physical reality. Thanks to our carefully preserved layer separation we were able to find a good compromise between a fully realistic and a completely diffuse illumination model. By achieving an overall preference over the physical reference we showed that going beyond reality can bring advantages that outweigh the fact that the result is not correct as a simulation. We argue that in the real world perception of highlights differ as free head motion allows for a change of viewpoint and that way reduction of rivalry. Proper reproduction of accommodation in natural conditions also helps to disambiguate individual layers and makes fusion easier. Our method enables us to overcome limitations of the display and deliver perceptually pleasant images.

The largest challenge for future work is certainly a generalization of our approach for a real captured content. A necessary step for this is a computer vision algorithm for a robust multi-layer disparity extraction that does not rely on rendering meta information. Another very interesting direction would be application of the method to rendering of

fully volumetric effects such as clouds. The method could potentially improve the depth appearance of such volumes as the reduction of rivalry would support the HVS disparity fusion mechanism.

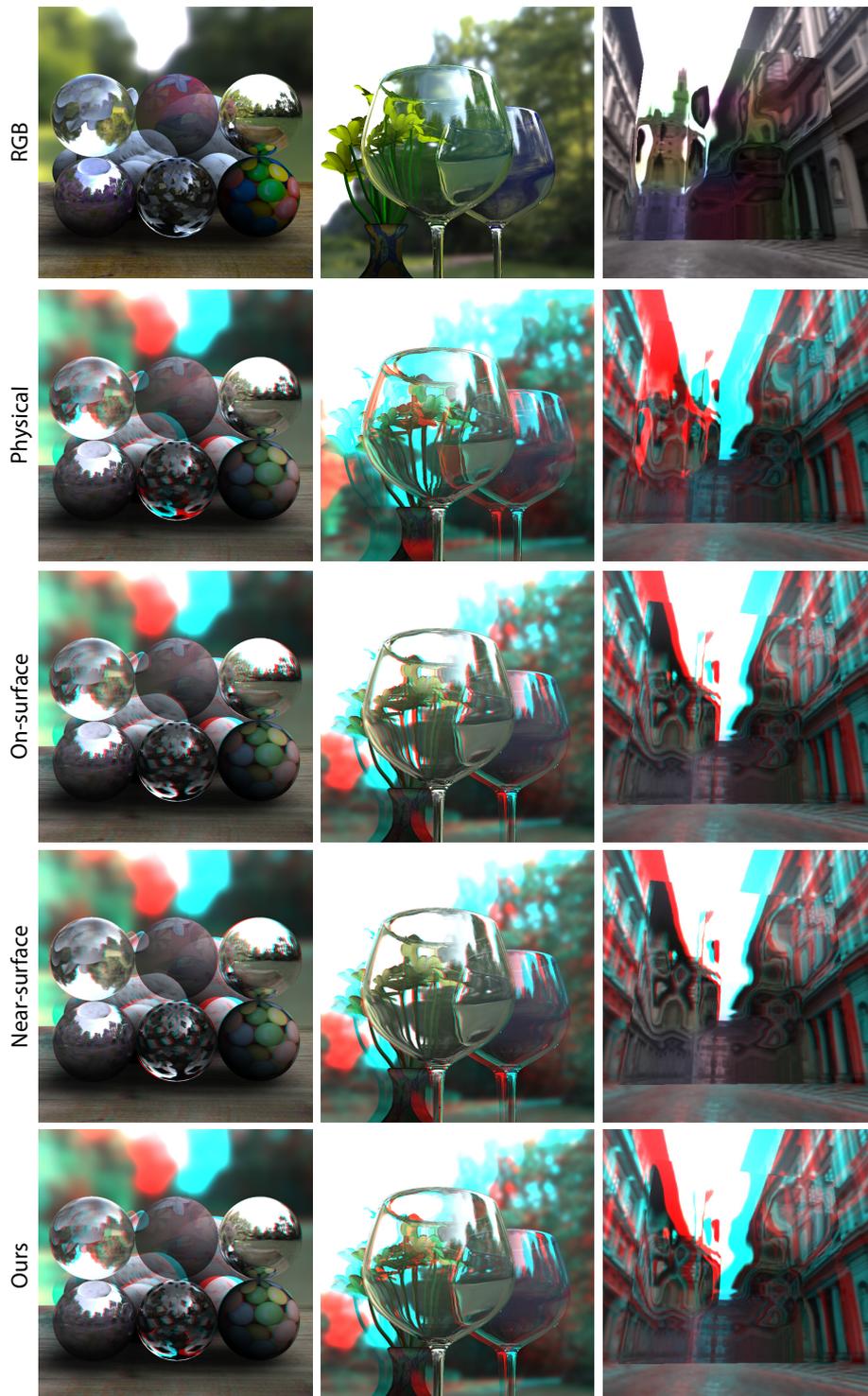


Figure 8.3. Results of our camera offset optimization compared to alternatives tested in our validation study.

Chapter 9

Disparity perception in scotopic vision

Unlike in previous chapters we cannot assume constant performance of the HVS when wide range of perceivable luminance levels is considered. This is not only truth for previously well explored limitations in luminance sensitivity but, as we show in this chapter, similar dependencies can also be observed in performance of the stereoscopic vision.

The change of appearance between the same scene presented once at photopic luminance adaptation levels (daylight) and once at scotopic conditions (night vision) is not limited to a simple decrease of brightness but includes desaturated colors, decreasing spatial details and a shift to blue hues. However, a painting, print or display cannot reproduce scotopic conditions for technical reasons and adaptation to such conditions would require long time. To nonetheless convey a nocturnal mood, artists have employed several tricks to reproduce a nightly impression in painting [Livingstone 2002, Ch. 3] and since its early days in movie making where it is known as the “day-for-night” effect or the “American night”. Several tricks are used to overcome the difficulty that an observer should feel as if at scotopic conditions but at the same time limiting the degrading effect of scotopic vision: an overly blurry and colorless image would not be useful (e.g., in an interactive application such as a game) and therefore not preferable. Consequently, everything that can make an image feel more scotopic without degrading it to become useless is highly desirable.

The physiological processes explaining the phenomena of scotopic perception are well understood [Hess et al. 1990]: the loss of color perception and acuity are due to the change from cone to rod vision while the increased sensitivity towards the shorter wavelengths in the visible color spectrum is known as the Purkinje shift. In computer graphics, physiological models were used to simulate how a scene would be perceived at night, including the course of adaptation [Ferwerda et al. 1996, Pattanaik et al. 2000, Durand and Dorsey 2000, Haro et al. 2006] and color shift [Thompson et al. 2002, Kirk and O’Brien 2011]. Day-for-night tone mapping is a common effect in interactive computer applications such as games, where it is sometimes even a part of the gameplay (e.g., stealth action). In general, reproducing the appearance of one illumination setting using a medium that requires different conditions is a special form of tone mapping [Reinhard et al. 2010].

With the advent of stereoscopic reproduction, the question arises if and how depth perception changes in scotopic conditions and if so, how to account for it, both theoretically by a predictive perceptual model and in practical applications such as tone mapping. If monocular day-to-night tone mapping needs to reduce acuity and shift desaturated colors towards blue in images presented at photopic conditions, how do we need to change a stereo image presented at photopic conditions to most faithfully

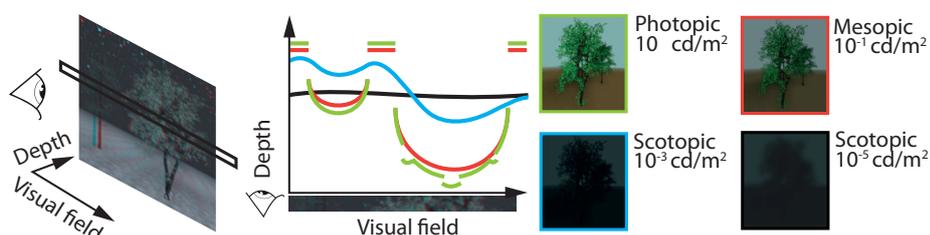


Figure 9.1. A one-dimensional cross section (horizontal axis) through the perceived depth (vertical axis) of a stereo rendering of trees such as the one in Fig. 9.8 for different adaptation states (colors). Under photopic conditions, all stereo details such as leaves are visible and the trees are fully separated. In mesopic conditions, stereo details start to disappear and the range gets smaller. Under scotopic conditions, only rough structures remain, that disappear before the luminance is too low to perceive an image entirely.

reproduce scotopic appearance? We address this question by two contributions:

- Measurement of disparity sensitivity in scotopic conditions
- A combined color-disparity manipulation approach to reproduce scotopic stereo appearance

9.1 Overview

Our approach emulates the appearance of scotopic stereo on a photopic stereo display. As illustrated in Fig. 9.1, stereoacuity is lower in scotopic conditions. Consequently, day-for-night stereo has to degrade the stereo content to match the scotopic experience. To this end we address two questions: how does stereoacuity behave in scotopic conditions and how do we need to process disparity to account for it?

To address the first question, we conduct a series of experiments to relate adaptation luminance to threshold elevation in the discrimination of binocular disparity patterns (Sec. 9.2). This experiment complements the findings of Cormack et al. [1991], who measure the relation of luminance contrast and disparity perception for mostly photopic conditions. In pilot experiments, we identify a well-discernible pattern that limits monocular stereo cues, as well as each participant's individual luminance contrast detection thresholds. A final experiment is conducted by presenting stereo stimuli in different adaptation conditions that range over eight orders of magnitude using an HDR stereo display in combination with neutral density optical filters.

Addressing the second question, we describe how to process disparity to produce a scotopic stereo impression (Sec. 9.4). We use the result of the perceptual experiment to identify stereo details that would not be perceived if the scene actually was scotopic and remove them. The processing is conceptually simple and can be computed at interactive frame rates.

9.2 Experiments

We repeat the experimental procedure of Cormack et al. [1991] to find how the luminance contrast and disparity detection thresholds change in scotopic conditions, which have

not been studied so far. Even existing monocular contrast perception models [Mantiuk et al. 2006] differ significantly with respect to sensitivity measurements under different levels of adaptation luminance.

We perform three experiments. The first two are pilot experiments to establish a reliable monocular baseline which is required for two reasons. First, it is not obvious which luminance pattern to use, so we test several alternatives (Sec. 9.2.2). Second, luminance contrast detection thresholds vary between observers and need to be known to model the effect in a longitudinal study with the same observers under different conditions (Sec. 9.2.3). Finally, the third experiment quantifies the relation of luminance adaptation and disparity detection thresholds (Sec. 9.2.4) and produces the model actually required to perform stereo day-for-night conversion (Sec. 9.3).

9.2.1 Methods

Stimuli The stimuli are random square stereograms (RSSs) composed of a regular grid of randomly black or white squares with equal probability on a gray background (Fig. 9.2b). To create the desired luminance contrast, we ensure equal brightness steps between the black, gray, and white levels used, thereby matching the average brightness of the stimuli to the background. The gray background luminance is also the adaptation luminance.

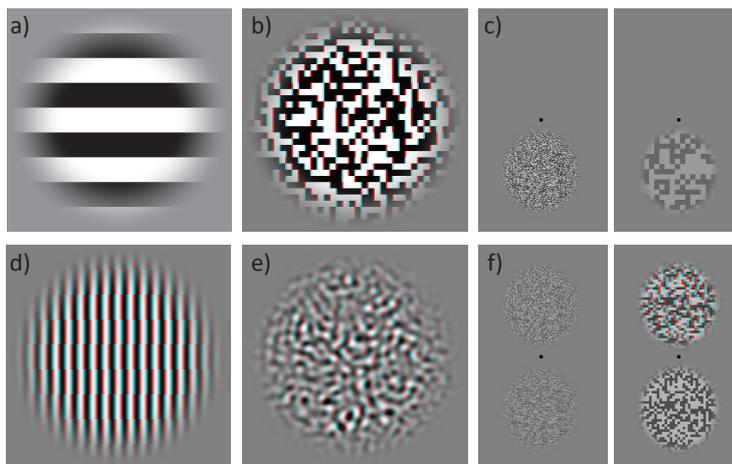


Figure 9.2. (a) Depth map used in all experiment stimuli. (b) Example random square stereogram (RSS) stimulus. Alternative stimuli such as (d) sine wave gratings or (e) bandlimited noise were rejected because of false monocular cues to disparity. Screen layout used in Experiments 1 and 2 (c), and Experiment 3 (f) optimized for various adaptation levels.

The stimuli have horizontal depth corrugations in the form of a square wave created by 8 equally spaced steps in disparity that coincide with some of the horizontal luminance edges between two rows of squares. The luminance contrast is faded out smoothly at the edge of the stimuli to avoid monocular cues due to the disparity shift. The average depth of the stimuli is at the screen plane to avoid vergence-accommodation conflicts.

RSS stimuli were chosen to resemble the stimuli used by Cormack [1991] but at varying, coarser resolutions to account for scotopic conditions. Several alternative luminance patterns were considered, including band-limited noise, Gabor patches, and sine-wave gratings (Fig. 9.2d,e), but as they do not allow alignment of the disparity

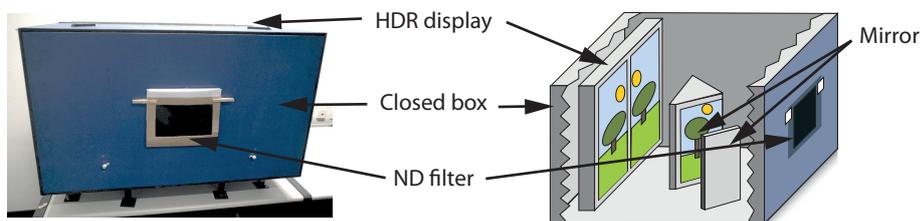


Figure 9.3. Wheatstone HDR stereoscope and filter aperture.

edges with existing luminance edges, and instead each disparity edge creates a new luminance edge, this would result in a monocular depth cue.

Apparatus Stimuli are presented on a 47" SIM2 HDR47E high dynamic range display with 1920×1080 resolution. Each half of the screen is directed to one eye by a Wheatstone [1838] stereoscope enclosed in a black box with a small viewing aperture to avoid any stray light (see Fig. 9.3). The viewing distance to the screen is 90 cm, giving each pixel a size of approximately 2 arcmin. This display system is set up in a dark room and was radiometrically calibrated to accurately and reliably produce any luminance value between 1 and 1000 cd/m^2 . Rosco E-Colour+ neutral density filters with specified optical densities of 0 (no filter), 1.2, 2.4, 3.6, or 4.5 cover the viewing aperture to produce the required scotopic luminances while preserving the displayed luminance contrast. Our own measurements confirmed that the transmittance of the filters was within 4.5 % of the specifications.

Procedure All thresholds are estimated with two-alternative forced-choice (2AFC) adaptive staircase designs using QUEST [Watson and Pelli 1983]. Stimuli are presented simultaneously in two locations, above and below a black fixation dot at the center of the screen (Fig. 9.2c,f). Between trials the screen is blanked to a slightly darker gray level than the gray background used during trials. The staircases are repeated several times and the results of the converged staircases are averaged.

Observers 3 observers participated in the experiments: 2 authors (B and C) and 1 experienced psychophysical observer (A) who was naïve regarding the purpose of the experiment, all male, age 24–31 ($M = 27$, $SD = 3.6$). Participants had normal or corrected-to-normal visual acuity and did not have stereo-blindness or night-blindness. We find 3 observers sufficient as the time needed for complete measurement consisting of many staircases is in the order of several hours and the variance between observers is low. A small number of participants in complex psychophysiological experiments is also common in comparable studies [Cormack et al. 1991, Livingstone and Hubel 1994] that measure similar phenomena.

9.2.2 Experiment 1: Optimal RSS frequency selection

The first experiment is conducted in monocular conditions and identifies the best luminance pattern frequency for each luminance adaptation. The frequency is defined by the inverse of the size of the squares. In this experiment, a stimulus with luminance contrast between 0 and 1 Michelson units is presented in either the top or bottom

location. The other location remains blank. There are no depth corrugations in the RSS stimuli, i.e., all squares are presented at screen depth.

The luminance contrast detection thresholds are estimated at each square size (2, 4, 8, 16, or 32 pixels, or 4, 8, 16, 32, or 64 arcmin) and adaptation luminance level (10, 0.631, 0.04, 0.003, or 0.0003 cd/m^2) by averaging the results of 2 staircases. The lowest threshold occurs for the optimal square size for each adaptation luminance level. The optimal square sizes (2, 8, 8, 16, and 32 pixels) are used in the following experiments in their respective adaptation luminance level.

A single observer (B) determined the optimal size of the squares in the RSS stimuli, which leads to highest sensitivity to luminance contrast detection for each adaptation luminance. Note that we repeat the experiment of Cormack et al. over a wide range of adaptation luminances, where the peak of contrast sensitivity shifts towards lower spatial frequencies with decreasing luminance [Wandell 1995, Fig. 7.21]. Since our extension of the function measured by Cormack et al. will be parametrized by adaptation luminance, and specifies stereoacuity thresholds as a function of contrast scaled in JND units, the goal of this experiment is to derive RSS stimuli that lead to conservative (as small as possible) contrast detection thresholds. This way a common basis in deriving the Cormack function is established, which enables its meaningful interpretation across all lighting conditions.

While the contrast sensitivity function (CSF) might already predict the optimal visible frequency, we decided to perform this calibration experiment, as our stimuli do not consist of sinusoidal gratings of a single frequency with which the CSF was measured (the reason is recalled in Fig. 9.2) but of boxes which combine all frequencies in a way that is complex to model.

9.2.3 Experiment 2: Luminance contrast detection threshold

The same experiment as before is now performed by all observers, using only the optimal spatial frequency of the RSS squares for each adaptation luminance, to determine their luminance contrast detection thresholds by averaging the results of 3 staircases. These thresholds (see Fig. 9.4) will be used as the JND units of luminance contrast in the following experiment.

9.2.4 Experiment 3: Disparity detection threshold

In this experiment, a stimulus is presented in either the top or bottom location with disparity between 0 and 30 arcmin at the depth corrugation edges, which corresponds to 18 cm between the near and far parts of the corrugation. Disparities up to 30 arcmin were included to provide sufficient operational space to the staircase, in particular at the lowest adaptation luminance level. Such horizontal disparity is still within the fusion range [Qin et al. 2006]. No convergence to disparity thresholds higher than 15 arcmin was observed in practice. A stimulus with the same luminance pattern but without any depth corrugations is presented in the other location. The disparity detection thresholds are estimated by averaging the results of 2 staircases at the same adaptation luminance levels as before and at 4 luminance contrast levels starting at 2 or 3 JNDs (as measured in the previous experiment) and increasing logarithmically to the maximum verified contrast of the display (see Fig. 9.5). Stimuli at luminance contrast levels below 2 JNDs are not sufficiently visible to allow disparity detection. The optimal spatial frequency of the RSS squares is used for each adaptation luminance.

9.3 Model of wide-luminance range stereo sensitivity

We will now fit the result of the previous three experiments to a closed-form model of luminance contrast sensitivity and disparity sensitivity.

Luminance contrast sensitivity model First, the results of the contrast detection threshold experiment for all observers are fit to a power function that maps from adaptation luminance L_a to mean luminance contrast detection threshold C_{Thr}

$$C_{\text{Thr}}(L_a) = c_1 L_a^{c_2},$$

where $c_1 = 0.0145$ and $c_2 = -0.314$ (Fig. 9.4; Degree of freedom-adjusted $R^2 = .94$). Note that this function resembles the well-known contrast versus intensity-functions (c. v. i.) [Reinhard et al. 2010, Sec. 10.7.2], here expressed in threshold magnitudes which are inversely proportional to the typically plotted sensitivity. We measure our thresholds for luminance RSS patterns that are directly used for the stereoacuity measurements in Exp. 3. Using threshold C_{Thr} , luminance contrast $C_{\text{JND}}(C_M, L_a) = C_M / C_{\text{Thr}}(L_a)$ (expressed in JND units) is computed for the luminance adaptation L_a and Michelson contrast C_M following the procedure of Cormack et al. [1991].

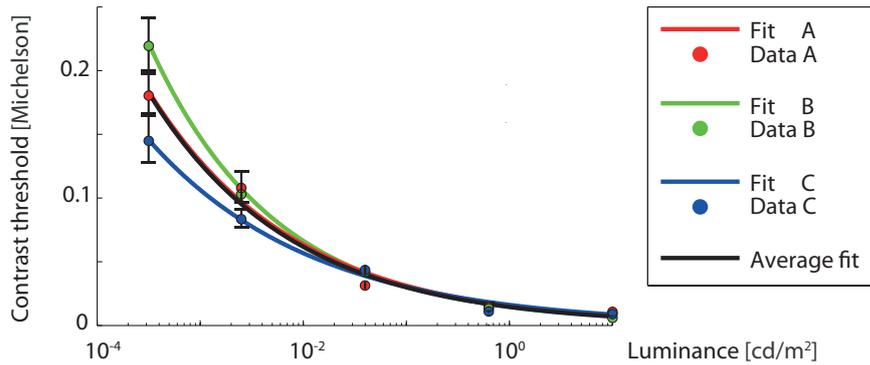


Figure 9.4. Mapping from adaptation luminance L_a to luminance contrast detection threshold C_{Thr} fitted from Exp. 2 in Sec. 9.2.3.

Disparity sensitivity model Next, we compute the disparity threshold D for adaptation luminance L_a and luminance contrast C_{JND} . We provide both a non-linear and a linear fit to the measurements from Exp. 3. The non-linear model is

$$\log_{10} D_{\text{nl}}(L_a, C_{\text{JND}}) = \frac{c_3}{\log_{10} C_{\text{JND}} + c_4} \cdot (L_a^{c_5} + c_6),$$

where $c_3 = 0.485$, $c_4 = 0.893$, $c_5 = -0.181$, $c_6 = 5.21$. A fit of each separate observer and a combination of all observers is shown in Fig. 9.5, bottom (Degree of freedom-adjusted $R^2 = .77$ for combined observers; $R^2 = .83$, $.95$, and $.95$ for individual observers). The nonlinear model saturates thresholds for suprathreshold contrast and adaptation luminance levels. An alternative linear fit is

$$\log_{10} D_{\text{lin}}(L_a, C_{\text{JND}}) = c_7 \log_{10} C_{\text{JND}} + c_8 \log_{10} L_a + c_9,$$

where $c_7 = -0.873$, $c_8 = -0.155$, $c_9 = 2.618$. A fit of each separate observer and a combination of all observers is shown in Fig. 9.5, top (Degree of freedom-adjusted $R^2 = .75$ for combined observers; $R^2 = .78, .88, \text{ and } .91$ for individual observers).

The linear model is simpler and computationally more effective, but provides a lower quality-of-fit in terms of R^2 as it lacks the saturation the non-linear model provides. We did not observe a large influence of model choice on our end results and use the non-linear model $D(L_a, C_{\text{JND}}) = D_{\text{nl}}(L_a, C_{\text{JND}})$ in the following.

Discussion We find that the difference of our scotopic disparity thresholds to the photopic ones is up to 20-fold and on average 12-fold depending on contrast. Such large differences are commonly visualized in logarithmic space (see Fig. 9.5). We applied logarithmic scaling to the adaptation luminance to get closer to the non-linear sensitivity of the HVS and also to keep plots comparable with those of Cormack et al. [1991]. We chose the functions $D(L_a, C_{\text{JND}})$ to model the influence of luminance contrast and adaptation luminance as two independent effects. We performed the least-square fitting in the same logarithmic-logarithmic space to achieve perceptually meaningful fitting errors. As a result, logarithms are present in both alternatives for $D(L_a, C_{\text{JND}})$. In case of non-linear fit it further increases the apparently complicated form of the model, otherwise consisting of a simple mix of rational and exponential function. The logarithm applied to the function C_{JND} was removed by substitution $z^{a \cdot \log_z x} \equiv x^a; z \in \mathbb{R}^+$ turning the exponential function into a power function.

Our photopic disparity thresholds are on average fourfold higher than those reported for similar contrast JNDs by Cormack et al. [1991]. This can be caused by the different procedure used to scale contrast in JNDs or by differences in the disparity threshold measurements themselves. Absolute threshold magnitudes cannot be compared between publications, but relative differences inside a single experiment likely remain valid. Therefore, our contrast-disparity threshold function has a similar shape as the one of Cormack et al. [1991].

9.4 Our approach

In this section, we show how to process stereo content to match luminance adaptation, i.e., to produce a scotopic appearance on a photopic stereo display. The goal is to achieve visually plausible stereoscopic image reproduction by reducing the acuity of disparity perception in a similar way as during true scotopic adaptation. This consequently also decreases the performance for tasks requiring depth understanding or object segmentation, e.g., spotting an enemy in a computer game. To this end, we will change both the disparity (Sec. 9.4.1), and the luminance (Sec. 9.4.2). The input to our stereo day-for-night conversion is stereo content captured at photopic conditions and a desired new luminance adaptation condition. The output is stereo content that resembles a scotopic stereo impression when shown in photopic conditions. We process stereo content in the form of an image pair, in combination with a per-pixel vergence map. Every pixel of this map stores the angle the eyes would form when verging to it. Such a vergence map can be computed from the interocular distance and the screen's distance, size and resolution, in combination with the pixel disparity that is readily output by interactive applications such as computer games, as well as by movie production, and can be approximated from a stereo pair using computer vision approaches.

Some pixels are only visible in one image of the stereo image pair (occlusions). As they cannot be matched with a pixel in the other image of the pair, they require special

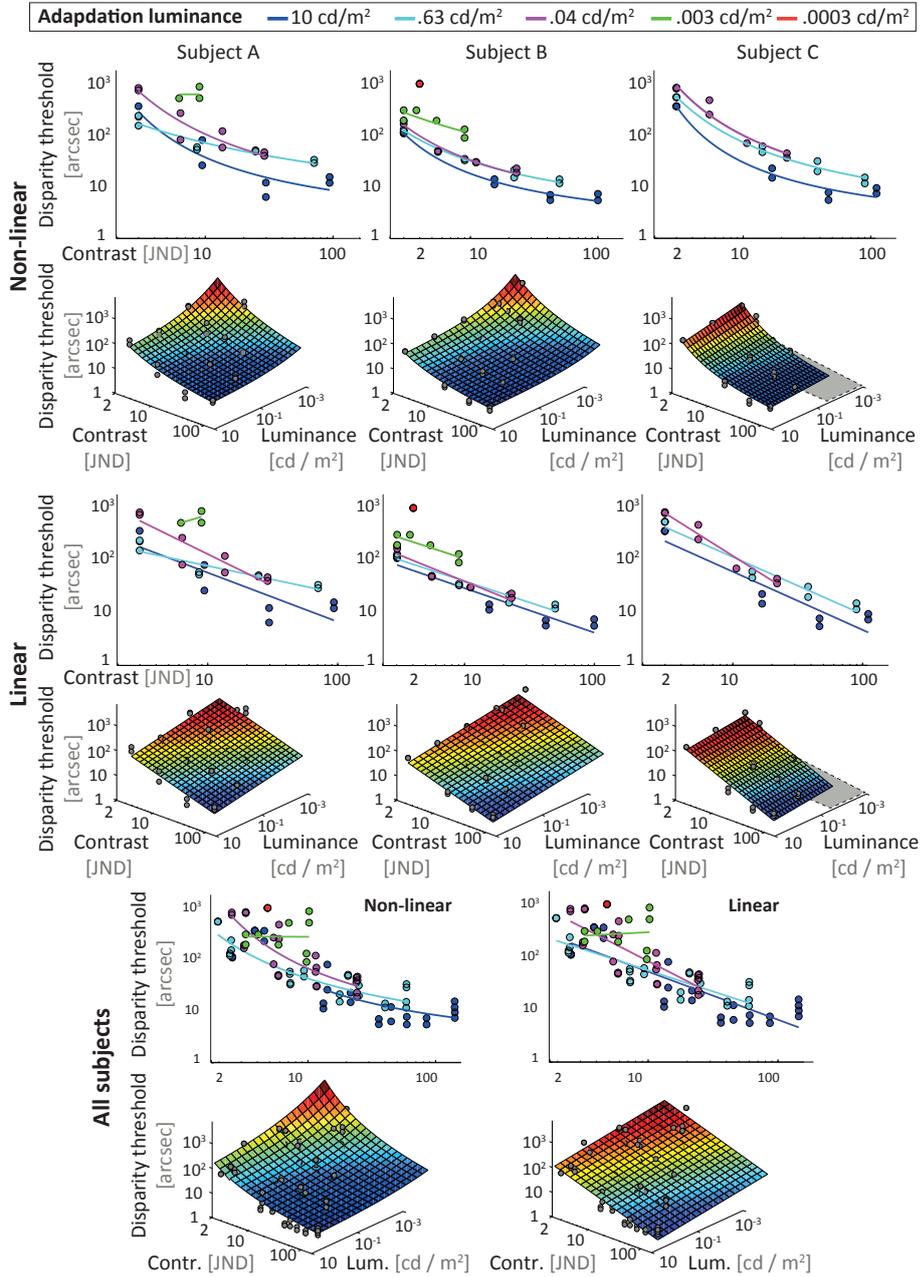


Figure 9.5. Mapping from luminance contrast C_{JND} and adaptation luminance L_a to disparity threshold D fitted to Exp. 3 by a non-linear and a linear function in Sec. 9.2.4.

consideration. For computer-generated content, the depth at such pixels is known. For computer-vision content, filling such pixels with plausible depth values is up to the stereo reconstruction used. Therefore, in both cases, vergence can be computed for every pixel, even if occluded in the other image. The resulting maps are identical in unoccluded regions but differ in occluded regions. While this is not perceptually principled, it allows for a practical solution in which the pipeline is executed twice, with a different vergence map for each image of the stereo pair.

9.4.1 Disparity processing

Disparity processing first filters the vergence map and then warps both images in the stereo image pair to match the filtered result. We will first list the requirements of this processing and give some necessary definitions, before explaining the process itself.

Requirements The input vergence map should change such that disparity that would be perceived weaker or not at all in scotopic conditions is weakened or removed entirely. First, all combinations of physical disparity with luminance contrast below 1 JND in scotopic conditions – and therefore imperceptible – are removed. In particular, fine details with low luminance contrast will disappear [Frisby and Mayhew 1978]. Second, the overall range of disparities will be reduced: While photopic conditions reproduce a wide range of disparity JNDs, in scotopic conditions this range is arbitrarily low, up to the point where luminance patterns are still discernible, but no stereo perception is present.

One could argue that changing the photopic image luminance to emulate scotopic conditions is sufficient to already degrade the stereo perception to meet those requirements. The existence of day-for-night techniques demonstrates that this does not apply in practice: displays are not capable of producing scotopic image conditions, the viewing conditions do not allow for adaptation, and even if they could, adaptation would take a considerable time. Finally, directly changing luminance to produce a scotopic depth perception (reduced contrast, blur) would lead to severe degradation of the image and render the content unusable as seen in Fig. 9.6. Consequently, even if color image



Figure 9.6. *a)*: Directly simulating scotopic vision by degrading luminance results in the desired reduced depth, but produces an overly blurry image that lacks contrast and consequently has become useless in practice. *b)*: Our approach matches the disparity to be similar to scotopic depth perception in combination with luminance that was created using common day-for-night (*c)*), producing results that both have correct depth and visual appeal.

day-for-night reduces the luminance to the low end of the range afforded by the display device, it has to remain photopic and so does its stereo perception. Emulating the image hue and saturation degradation due to scotopic conditions could reduce stereo perception to some extent [Simmons and Kingdom 1997, 2002]. However, common day-for-night luminance, hue, and saturation processing applied to stereo pairs does not match the

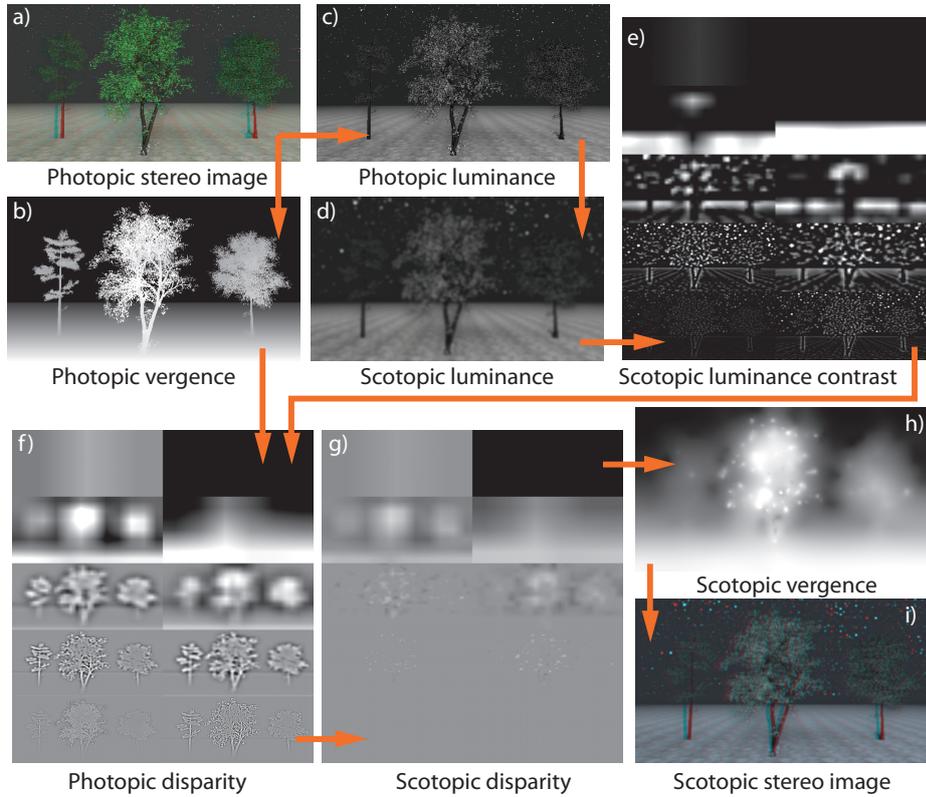


Figure 9.7. Disparity processing steps (orange arrows) in our pipeline explained in Sec. 9.4: *a, i*): Input and output HDR stereo image in physical units shown in anaglyph. *b, h*): Input and output vergence (visualized as gray). *c, d*): Photopic and scotopic luminance. *e*): All bands of the luminance contrast pyramid (decreasing frequency from bottom-left to top-right; brighter is more contrast). *f, g*): Photopic and scotopic disparity (Laplacian of vergence); same layout; black indicates negative, gray neutral and white positive disparity.

reduced stereo perception of real scotopic conditions (see Fig. 9.8). We conclude that disparity itself needs to be altered in order to produce scotopic stereo appearance.

Notation We denote the input and output vergence maps as $d_{\text{in}} \in \mathbb{R}^2 \rightarrow \mathbb{R}$ and $d_{\text{out}} \in \mathbb{R}^2 \rightarrow \mathbb{R}$. Further, we denote the input and output RGB images as $I^{\text{In}} \in \mathbb{R}^2 \rightarrow \mathbb{R}^3$ and I^{Out} . The color image processing is performed twice: once for I^{In} from the left and once from the right eye's view. The disparity processing for d_{in} is performed once and applied to both color images. Finally, the adaptation luminance we would like to emulate is denoted as L_a^{In} , typically ranging from 10^{-4} cd/m² to 10^{-1} cd/m² and the current display condition adaptation L_a^{Out} between 1 and 100 cd/m².

Disparity processing First, the Michelson luminance contrast C_M (Fig. 9.7e) is computed from a Laplacian pyramid [Burt and Adelson 1983] of a luminance image (Fig. 9.7c) which has been mapped for target scotopic luminance L_a^{Out} by filtering the luminance with the CSF cutoff (Fig. 9.7d). The highest spatial frequency of luminance resolvable by the HVS was determined using the acuity function (Eq. 15 from

Ward et al. [1997]). This possibly very blurry image (Fig. 9.7d) is never actually shown to the user, but only serves as a proxy to predict stereo perception. That allows separate processing of luminance and disparity, and therefore achieves both artistic goals in the luminance domain and physiologically correct presentation in the disparity domain.

Next, the vergence map d_{in} (Fig. 9.7b) is decomposed into another Laplacian pyramid (Fig. 9.7f), which effectively contains differences of vergence angles, i.e., disparity δ . Note that both pyramids are fully analogous, which means that the corresponding value of luminance contrast C_M can be immediately found for any spatial frequency band and spatial location where the disparity δ is defined. This is important as our goal is to reduce the disparity perception in the scotopic luminance adaptation L_a^{In} with respect to the photopic condition adaptation L_a^{Out} given the luminance contrast C_M . To perform such a reduction in a visually meaningful way we need to transform the physical disparity magnitude δ into a perceptually linear sensory response space, where the linearized disparity values δ^* better correspond to actually perceived depth changes. (A direct analogy to the CIE $L^*a^*b^*$ color space, where differences are proportional to perceived differences.) The transducer functions $\delta^* = t(\delta)$ serve this purpose [Didyk et al. 2011] by taking into account the increase of disparity discrimination thresholds with increasing pedestal disparity. Due to the compressive nature of the disparity transducer it can be approximated by a simple logarithmic function $\delta^* = 21 \log_{10}(\delta + 0.82)$, which is a fit to the data in [Didyk et al. 2011, Fig. 3.6] for depth corrugations of spatial frequency of 0.3 cpd, where the HVS sensitivity is the highest. We employ this transducer to transform all values in the disparity magnitude pyramid. Since the transduced disparity δ^* models the hypothetical response of the HVS, one can perform linear operations on its values so that the scaling $s \cdot \delta^*$ translates directly into the perceived depth reduction. We define such a scaling as a ratio of perceptually linearized disparity sensitivities $D^* = 21 \log_{10}(D(L_a, C_{\text{JND}}) + 0.82)$ for the photopic and scotopic adaptation luminance L_a^{Out} and L_a^{In} :

$$s(C_M, L_a^{\text{In}}, L_a^{\text{Out}}) = \frac{\log(D(L_a^{\text{Out}}, C_{\text{JND}}(C_M, L_a^{\text{Out}})) + 0.82)}{\log(D(L_a^{\text{In}}, C_{\text{JND}}(C_M, L_a^{\text{In}})) + 0.82)},$$

for suprathreshold of $C_{\text{JND}}(C_M, L_a^{\text{In}}) \geq 2 \text{ JND}$, as

$$s(C_M, L_a^{\text{In}}, L_a^{\text{Out}}) = 0$$

for $C_{\text{JND}}(C_M, L_a^{\text{In}}) \leq 1 \text{ JND}$ and as a smooth ramp between the two on the interval from 1 to 2 JND. The special treatment of luminance contrast below 1 JND was chosen to avoid extrapolation out of the range of our measurements. It models the absence of disparity perception in low contrast regions [Frisby and Mayhew 1978]. The luminance contrast on a given frequency band is defined as the maximum contrast in the sub-pyramid corresponding to equal or higher frequencies [Didyk et al. 2012]. To predict the effect of luminance contrast on disparity, each prior transduced disparity band is multiplied by $s(C_M, L_a^{\text{In}}, L_a^{\text{Out}})$ (Fig. 9.7g). As usually $s(C_M, L_a^{\text{In}}, L_a^{\text{Out}}) < 1$, disparity is effectively compressed, when displaying scotopic luminance on a photopic display ($L_a^{\text{In}} < L_a^{\text{Out}}$).

Finally, the inverse photopic transduction (a simple exponential function is used in this work to invert the logarithmic transducer) converts perceived disparity δ^* back to physical disparity δ [Didyk et al. 2011] and the scotopic vergence map is reconstructed using the inverse Laplacian transform (Fig. 9.7h).

Despite the involved perceptualization steps, the entire disparity processing can be efficiently performed on a GPU in a time linear in the number of pixels. Our

implementation processes a single Full HD vergence map (1920×1080) in 18 ms on an Nvidia Quadro 4000.

Discussion Note that although the resulting dynamic range compression of perceived depth under scotopic conditions clearly is a suprathreshold effect, the luminance contrast-dependent ratio $s()$ of perceivable disparity thresholds in photopic and scotopic conditions is used for this manipulation. Patel et al. [2009] observe that in the presence of blur not only the stereoscopic disparity thresholds are increased, but also the perceived suprathreshold stereoscopic depth is reduced. They attempt to explain their data on the perceived depth reduction using various sensory-perception transducer functions, including the logarithmic compression employed in our work, in which case the gain factor must be reduced. Note that the gain factor directly corresponds to our scaling $s()$, which we use for modeling both the perceived depth reduction and disparity detail suppression in scotopic conditions as a function of local luminance contrast. Unfortunately, only +2D of dioptric blur was considered by Patel et al. [2009], while presenting high contrast bright lines of 30 cd/m^2 that are imposed on a dark background, which makes those data too limited for our purposes. Clearly, further research is needed to investigate the impact of suprathreshold disparity magnitudes on the value of the scaling factor $s()$, which in this work relies on the threshold data. Note that this would require collecting data in a 5D space (adaptation luminance, luminance contrast, luminance spatial frequency, disparity magnitude, disparity spatial frequency), while in this work we focused on the first three dimensions, which are the most relevant for typical day-for-night image manipulations.

Image pair warping From the modified vergence map a stereoscopic image pair needs to be produced. First, the modified vergence is converted into modified pixel disparity assuming a fixed interocular distance, and display distance, size and resolution. The new image pair is created by warping, i.e., deforming the two images of the image pair such that features that had the original pixel disparity before processing, now produce the desired pixel disparity. This can be done in two ways: first, if the scene is sufficiently simple or the depth map is acquired from computer vision, depth image-based rendering (DIBR) [Fehn 2004] is used to produce a new image pair. Second, for rendered scenes, 3D geometry warping [Kellnhofer et al. 2013] achieves superior results, which do not suffer from disocclusions or sampling problems that are inherent for DIBR.

9.4.2 Luminance processing

Both images of the stereo image pair are processed to simulate monocular scotopic vision phenomena such as loss of acuity, the Purkinje shift and noise using the method of Thompson et al. [2002]. We used equation 15 from Ward et al. [1997] for physiologically-based spatial acuity. The luminance processing requires 30 ms on the aforementioned Nvidia Quadro 4000.

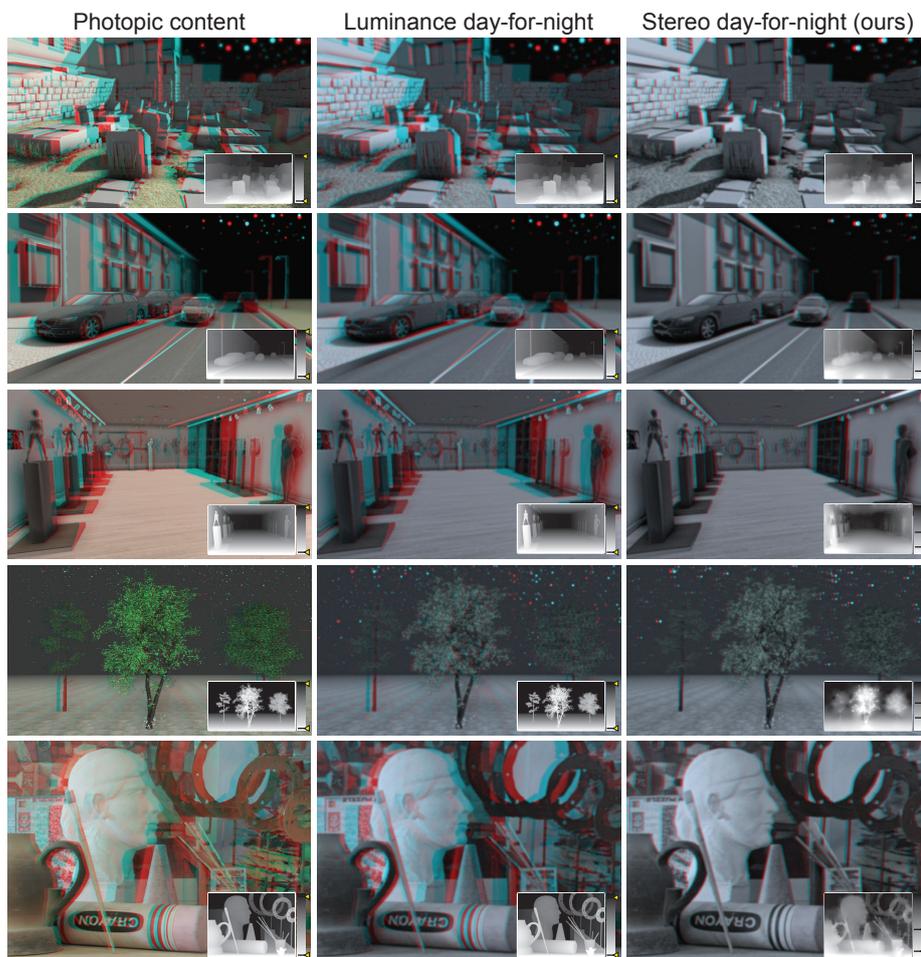


Figure 9.8.  Low-light stereo content (*rows*), processed using: Original (*1st column*), common day-for-night color tone mapping [Thompson et al. 2002] (*2nd column*), our stereo day-for-night (*3rd column*). Insets show the disparity with markers on the bar denoting the actual disparity range used. Common day-for-night tone mapping reproduces the desaturation, loss of acuity and blue shifts, but results in a mismatching depth impression, which is only perceived in photopic conditions. Our approach matches stereo fidelity to the actual scotopic percept. The captured stereo image (*5th row*) is courtesy of the Middlebury Stereo database [Scharstein and Pal 2007].

9.5 Validation

9.5.1 Results and discussion

Typical results of our approach are shown in Fig. 9.8. For each scene, the original photopic scene, the same scene with day-for-night tone mapping but the original stereo, and our combined luminance-stereo day-for-night are shown. Along with each scene, the original and modified disparity maps are shown as an inset.

The first row in Fig. 9.8 shows a virtual scene of a graveyard at night. After stereo day-for-night, small stereoscopic details, such as the vegetation in the foreground are removed, which remained visible in common day-for-night tone mapping. Also the overall depth range is reduced. Using our approach, the individual gravestones in the foreground are still separated, whereas the gravestones in the back – even if their physical distance is the same as in the foreground – do not produce a perceivable distance anymore. In common day-for-night, all gravestones are separated, which overestimates scotopic depth perception. The engravings in the stones and the structure of the back wall in the left have also disappeared after stereo day-for-night, even if they are still clearly perceived in the luminance image. This margin serves as an example for the fact that degrading the luminance alone is not sufficient to produce a scotopic stereo impression.

The second row in Fig. 9.8 shows a night driving stereo rendering. This shows a typical application, where simulation of scotopic stereo performance on a photopic screen is essential. A driving simulation requires real-time feedback, which is possible thanks to our GPU implementation. While the closer car to the left can still be discerned from its surrounding in depth, the remote car to the right has no remaining depth difference in respect to the other car. This effect is increased due to the fact that the car is both distant, has little contrast and has a low luminance.

The third row in Fig. 9.8 shows an art gallery room. Dark indoor environments are another common situation inducing scotopic vision. Here the statues on the far right blend with the surrounding wall while larger statues in the front left are still differentiable.

The fourth row shows a rendered landscape with trees. At night, the range is drastically reduced, but also small details disappear, which are still visible when only degrading luminance. The trees still do have a perceivable depth difference, but their depth details have disappeared.

The fifth and final row shows a captured stereo image from the Middlebury Stereo database [Scharstein and Pal 2007]. Here, dark edges such as the rings or fine structure on the pencil have disappeared, that were still present in the luminance image after common day-for-night tone mapping.

In general all our results exhibit compression of small disparity gradients which corresponds to inability of an observer to distinguish small objects with low luminance contrast from the background. As a result of threshold elevation the absolute disparity range of all scenes is scaled down. One might argue that such scaling is not motivated by real world observations where perceived large scale distances do not change with illumination.

Given the large difference between scotopic and photopic disparity discrimination thresholds, many disparity gradients that are not observable at real scotopic conditions would become visible in photopic LDR display conditions. Using our scaling we detect such events and modify the disparity so that details that would be below 1 JND in scotopic conditions stay below 1 JND in photopic conditions. This ensures the

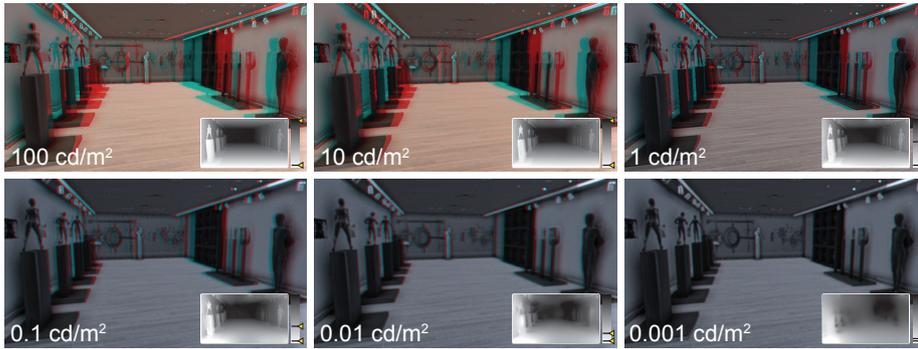


Figure 9.9. Results of applying our stereo day-for-night to an architectural scene at six different adaptation luminance levels L_a^{In} . RGB tone mapping controlled independently and simulated for adaptation level not lower than $0.1 \text{ cd} / \text{m}^2$ to prevent visual degradation.

preservation of detail visibility and invisibility. As a consequence, the total disparity range obtained by integrating such a disparity map is also rescaled. However, the 3D percept is maintained in areas where it would be expected, and the price of global depth range compression is typically acceptable to achieve the given objective. We also argue that in practice physically correct depth is rarely used as most scenes are usually larger than what the accommodation-vergence conflict would allow to present using current display technologies [Hoffman et al. 2008, Lambooj et al. 2009].

Our disparity mapping supports a large range of adaptation luminances (Fig. 9.9), where the luminance and disparity tone mapping could be controlled independently to achieve the desired artistic effect. This way the trade-off between perceptually correct depth detail preservation and physically correct depth range in the scene can be controlled.

9.5.2 User study

Validating our model is particularly challenging: a simple preference study would merely confirm that viewers dislike visual impairment, not whether the model depicts night scenes more realistically. A true validation of this model would require a comparison of depth perception performance between real scotopic conditions and simulated darkness in photopic conditions. The time required to dark-adapt prevents any direct comparison. For this reason, we report only informal user feedback collected as follows:

We invited $N = 9$ observers to dark-adapt for 10 minutes and asked them to memorize the appearance of the Graveyard, Driving, and Art Gallery scenes from Fig. 9.8 in scotopic stereo conditions. Observers then adapted back to office lighting for one minute and were presented the same three scenes in photopic display conditions, using a classic day-for-night luminance-only tone mapping [Thompson et al. 2002] and our approach vertically next to each other for unlimited time.

Observers were asked to answer the following questions and give any additional comment they would like to make.

1. Which image looks most similar to the scotopic image you saw?
2. Which image looks more like realistic night vision?
3. Which image would you prefer in a video game?

Interpreting the direct answers to our questions as 2AFC results in no significant effect (binomial test), except the following statement: The Graveyard scene with classic day-for-night is preferred and more similar to scotopic conditions (both $p < .02$). Observers also made the following comments:

- “Photopic images are generally too bright and detailed.” (Noted by 3 observers)
- “Street lights and gallery lights are expected to be turned on in a realistic night scene.”
- “Too much depth doesn’t look similar to scotopic conditions.”
- “More depth in a video game is always preferred.”
- “Our method looks more similar to scotopic for the Art Gallery.”

As a control group, monocular day-for-night and unprocessed images were shown asking the same questions. The only significant result was that day-for-night was more similar to scotopic conditions than the unprocessed image in all scenes (all $p < .02$). Comments made were:

- “Unprocessed color doesn’t look similar to scotopic.”
- “The blue shift makes it appear more scotopic.”
- “Day-for-night is less blurry than scotopic.”
- “The unprocessed image looks more similar to scotopic, except for color.”

The overall mixed outcome is expected as stereo perception preferences are already highly subjective and even more so in scotopic conditions. While strictly speaking we cannot conclude anything from insignificance, the outcome shows that while day-for-night has been accepted for many decades this did not result in a clear user preference, both for classic monocular and our stereo approach.

9.6 Conclusions

In this chapter we analyzed the relation of luminance adaptation and stereo perception. We conducted perceptual experiments to cover a very large range of luminance in stereo and devised a model that relates luminance contrast and luminance adaptation to disparity threshold elevation. This model was used to process binocular disparity, such that photopic stereo appearance is matched to the scotopic one. This processing can be achieved in real-time and is designed to extend classic day-for-night tone mapping in the luminance domain to stereo, e.g., in a computer game that should convey a nocturnal mood.

The model represents a true-to-life simulation of the visual degradation due to reduced information in scotopic conditions. Viewers may not prefer to be visually impaired in movies or video games, but do like a convincing depiction of scotopic conditions. Our scotopic stereo is a tool at the disposal of the director to make a scene feel more nocturnal.

Our day-for-night technique filters the disparities based on the model fitted to our measurements, essentially removing small depth details and compressing the depth range. As future work we would like to investigate simpler models that could produce similarly

reduced stereo perception with lower implementation complexity. We also would like to better understand the relation of scotopic depth perception and suprathreshold disparity, chroma, the time course of adaptation as well as different disparity frequencies, for which we do not account in this work.

Chapter 10

Luminance perception at absolute threshold



Figure 10.1. We simulate the change of image appearance between photopic conditions (*left*) and appearance in scotopic conditions close to the absolute threshold (*right*), where consistent vision fades into temporally varying (not reproducible in print) noise.

The HDR disparity remapping method presented in the previous chapter was able to simulate binocular depth perception under wide range of luminance conditions starting from very bright photopic all the way to very dark scotopic adaptation levels. However, human perception does not fully stop at such levels and visual percepts are processed even close to absolute darkness where sparse photons arriving to the retina can be modeled as discrete quanta. Performance of the vision in such conditions deteriorates drastically and the fusion of disparity becomes impossible as the noise in signal of each eye reduces their correlation below acceptable level. For this reason we cannot model stereoscopic perception using thresholds as in Chapter 9 as no such thresholds exist. Instead of that, this chapter describes the vision close to its absolute threshold through the appearance modeling of the visual noise characteristic for such conditions.

To complement the disparity perception from the previous chapter let us first discuss how the luminance perception changes with its absolute intensity. The HVS adapts to absolute luminance through several orders of magnitude; we can perceive a bright daylight scene as well as a moonless night. Appearance drastically changes for different absolute levels: at night (scotopic) color and acuity are reduced and a shift towards blue tones is perceived, when compared to the same scene in daylight (photopic) conditions.

In visual arts, cinematography or interactive applications (e.g., games), the change of appearance is often simulated to convey the illusion of a certain adaptation level despite being in a different display condition. A skillful painter is able to depict a scene shown on a photopic canvas as if it actually was scotopic. The same holds for movies, where

the so-called “Day-for-night” effect is used since the early days of cinema. For computer applications, techniques like tone mapping can convey a scotopic impression. In all cases, it is important to point out that adaptation effects are qualitatively reproduced and might differ in quantity: night scenes are blurred only enough to become noticeable and not as much as a strict HVS simulation would require, which would lead to an unpleasant viewing experience.

Computer graphics has now routinely modeled the shift from photopic over in-between mesopic to scotopic conditions [Ferwerda et al. 1996, Pattanaik et al. 1998, Durand and Dorsey 2000, Pattanaik et al. 2000, Thompson et al. 2002, Khan and Pattanaik 2004, Kirk and O’Brien 2011, Wanat and Mantiuk 2014] but the scotopic regime (vision close to its absolute threshold, e.g., a moonless night), has received little attention. Remarkably, the absolute threshold is close to the physical limitations of light itself; most dark-adapted subjects reliably detect flashes of light resulting in as little as 5 to 10 photons total on the retina during an integration time of 100 ms [Hecht et al. 1942]. Appearance under such conditions is substantially different from all other conditions. While scotopic vision can still rely on the retina as a classic noise-free sensor described by scalar ray optics, for close to absolute threshold, receptor noise due to the particle nature of light becomes apparent and requires accounting for quantum statistics.

In this chapter, we complement day-for-night tone mapping to account for the effects encountered close to the absolute threshold. Sec. 2.5 has provided the background in physics, neuroscience, and human vision, and Sec. 3.3 a review of the state of the art in modeling human scotopic perception in computer graphics. Here we propose a neurophysiologically-motivated model of rod receptor noise which adds temporal variations to the image, as expected to be experienced in scotopic conditions close to the absolute threshold (Sec. 10.1). We then present the related computational aspects, involving a photon-accurate retinal image representation, an efficient rod noise generation drawn from image content-dependent distributions, and a temporal rod-signal integration (Sec. 10.2). Using graphics hardware, our model requires 18 ms for an HD image and we compare our results to different alternatives in a perceptual evaluation (Sec. 10.3).

10.1 Model of rod noise

While the source of scotopic noise is well understood in physiology, it has not yet been considered in computer graphics, where noise is added in an ad-hoc way. This section introduces the reader with a computer-graphics background to the physiology of luminance perception at levels close to the absolute thresholds.

Absolute threshold On average, 60 % of flashes with 510 nm wavelength, a duration of 1 ms, emitting 54–148 photons in total towards a retinal area covering 500 receptors located off the fovea (which has no rods) will be detected by dark-adapted subjects [Hecht et al. 1942]. The fraction of photons that actually reach the retina is then only 10%. The key result of this experiment is that, close to absolute threshold, answers can only be given with certain probabilities, not with absolute certainty. In consequence, photon counts are related to detection likelihoods via *receiver operating curves* (ROCs).

Quantization noise The seminal work of Hecht et al. [1942] has shown that the quantization of light into photons has actually a practical perceivable consequence. In conditions close to the absolute threshold, photon count is important as, for rare

discrete random events, noise is to be expected. Such noise can be modeled by a Poisson distribution, which estimates the probability density function P of observing k events given the expected number of such events μ :

$$P(k, \mu) = \frac{\exp(-\mu)\mu^k}{k!}.$$

The probability of observing Θ or more events is P 's tail distribution (complementary cumulative distribution function)

$$F(\Theta, \mu) = \sum_{k>\Theta}^{\infty} P(k, \mu).$$

The probability of seeing a flash of N photons per unit time (integration time) at the cornea is

$$F_{\text{Quant}} = F(\Theta, qN), \quad (10.1)$$

where Θ is the minimal number of photons that can be perceived and q is the quantum efficiency. Such noise is qualitatively multiplicative (i.e., its magnitude depends on the image), as it depends on the actual number N of photons at the cornea. Hecht et al. [1942] have fitted the parameters of their ROC measurements against such a model and found $\Theta \approx 6$ and $q \approx 0.06$. Note, that for $N = 0$ the probability is zero, which cannot explain seeing noise in the absence of light. Furthermore, a quantum efficiency of $q \approx 0.06$ is judged to be too low with respect to other physiological data [Field et al. 2005]. Consequently, the model needs to be extended.

Photon-like noise An alternative source of noise has been identified in spontaneous photo-transduction [Barlow 1956, Baylor et al. 1979, Ashmore and Falk 1977]. Once in two thousand years, a rhodopsin molecule is isomerized without any reason, leading to false-positive responses, which becomes important when explaining the perception of noise in the absence of all light. There are 60,000,000 rods [Jonas et al. 1992] and given 2,000,000,000 rhodopsin molecules in each [Yau et al. 1979], results in 0.032 events per second and rod. While this is little compared to the excitation rates above absolute threshold (high N), it is perceivable close to absolute threshold where N and the rate of such spontaneous photo-transductions become similar. The probability of seeing a flash due to such photon-like chemical events is

$$F_{\text{Dark}} = F(\Theta, D) \quad (10.2)$$

where D is a dark-noise constant, which characterizes the rate of spontaneous photo-transductions. This noise is qualitatively additive (it does not depend on the actual number N of photons but on a constant D) and could explain perceived noise in the practical absence of light. When fitting behavioral data [Teich et al. 1982] to such a model, one can find $\Theta \approx 40$ and $D \approx 50$. The best fit however, is produced by a model that accounts for both quantum and dark noise.

Combined noise Lillywhite [1981] has shown physiological evidence that photo-transduction near absolute threshold is in fact not a Poisson process. A Poisson process assumes that events are statistically independent. This does not hold as bleaching causes a non-linear response of the photoreceptors to consequent photons [Gutierrez et al.

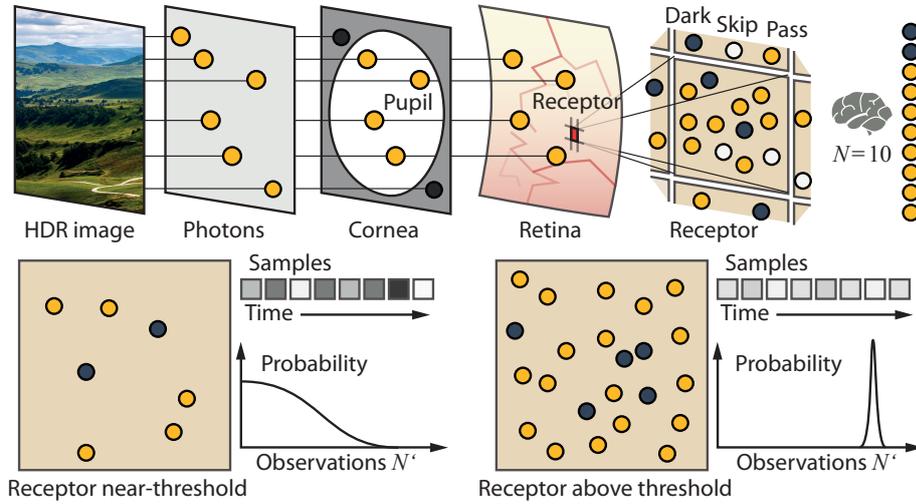


Figure 10.2. Luminance perception at absolute thresholds (*Left to right*). Starting from the input HDR image, we compute the number of photons reaching the retina per unit time. At a single receptor, a photon either contributes to luminance perception (*yellow*) or is skipped (quantum efficiency) (*white*). Additionally, dark photon-like events (*blue*) contribute to the perceived sum, here ten. Near absolute threshold, the probability distribution is wide (*curve*). Luminance samples drawn over time according to this distribution are unstable and vary (*grey squares*). Well above absolute threshold, the probability distribution is peaky. Samples drawn from this distribution are stable and very similar to the expected value.

2005]. A better model can be obtained if the noise is assumed to be a combination of both quantization and photon-like noise. The probability distribution of observing exactly k photons is given by

$$P_{\text{All}}(k, N) = P(k, qN + D)F(\Theta, \alpha k), \quad (10.3)$$

where α is a *constant of growth*. Fitting to behavioral data yields $\alpha = 0.5$, $q = 0.2$, $D = 19$, and $\Theta = 15$ [Field et al. 2005], which is in good agreement with all physiological evidence [Teich et al. 1982].

10.2 Our approach

Overview The input to our system is a sequence of HDR images [Reinhard et al. 2010] storing “retinal” radiance, i.e., after including the eye’s complete optical transfer. The simulated noise is added to an LDR image, produced from the HDR image by a day-for-night tone mapping of choice, leading to changes in chroma, saturation and acuity. Decoupling noise and tone mapping allows us to maintain full control over the appearance. A modular design also leads to easy and efficient integration into existing systems. For all results in this chapter, we used the tone mapping by Thompson et al. [2002].

The output is an LDR image sequence to be displayed on an LDR display at photopic levels, which is perceived as similar to a scotopic experience close to absolute threshold.

Fig. 10.2 summarizes the computational pipeline from photons emitted by an HDR image to the triggered rod responses. First, the HDR input is converted into photon

counts per unit time and area (Sec. 10.2.1). Next, the according retinal response is simulated (Sec. 10.2.2). As light perception depends on an integration period, which is particularly long in night vision (Bloch's law [Bloch 1885]), we also consider eye motion (Sec. 10.2.3).

10.2.1 Photon counts

Close to absolute thresholds, the actual number of photons is important. Hence, we need to convert the image or the frames of an image sequence into photon counts per time and receptor. We follow the derivation of Deering [2005, Sec. 6].

First, we assume that the HDR input image contains scene-referred calibrated values in CIE XYZ color space (refer to [Reinhard et al. 2010, Table 2.9] for transformations from/to other standard RGB color spaces). The derivation is independent between pixels and described for a single pixel in the following.

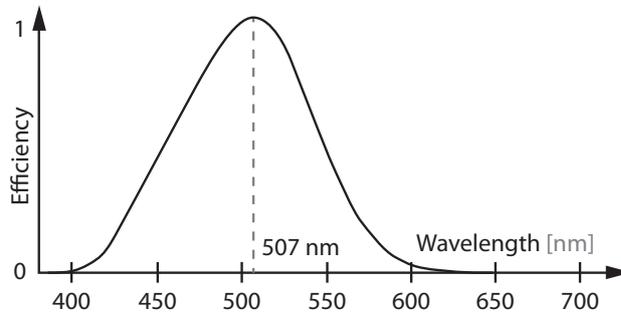


Figure 10.3. The scotopic luminous efficiency function [Wandell 1995].

Given a pixel, its HDR XYZ values are converted into scotopic luminance L following the transformation proposed in [Pattanaik et al. 1998]:

$$L = -0.702X + 1.039Y + 0.433Z.$$

Note that this is merely an approximation derived through a linear regression of the color matching functions $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$ and the scotopic luminous efficiency function $V'(\lambda)$ [Wandell 1995] (see Fig. 10.3), which are defined over the visible spectrum of wavelengths λ . In an unlikely case when per-pixel spectral radiance values $Y_e(\lambda)$ are available (as in [Kirk and O'Brien 2011]), L can be computed directly as:

$$L = 1700 \int_0^\infty V'(\lambda) Y_e(\lambda) d\lambda. \quad (10.4)$$

In both cases, a scotopic luminance L in candela per square meter is obtained. Given the area ΔA of a screen pixel, the scotopic luminous intensity in candela is $I = \Delta A \cdot L$. The luminous flux $\Phi_s = I \cdot \Delta\omega$ arriving at the retina is expressed in lumen, where $\Delta\omega$ is the solid angle of the pupil. This solid angle $\Delta\omega = 2\pi \cdot d_p(d_p^{-1} - (d_p^2 + d_s^2)^{-1/2})$ is approximated as a disk of diameter d_p in distance d_s . The distance to the pupil d_s is given by the distance of the observer to the display. The diameter of the pupil d_p depends on the luminance adaptation state and can be computed for the average value of L [Watson and Yellott 2012].

The data we use in Sec. 10.2.2 were acquired for $\lambda = 507$ nm [Hecht et al. 1942], the peak of V' . We assume that the response to other wavelengths close to absolute threshold

is proportional to their luminous efficiency (refer to Eq. 10.4). In other words, we treat multi-chromatic luminous flux Φ_s as equivalent to its monochromatic analog with $\lambda = 507$ nm. Therefore we can derive the radiant flux Φ_e in Watt as: $\Phi_e = \Phi_s/1700$, i.e., by inverting Eq. 10.4, which holds for any photometric quantity and its radiometric counterpart. The radiant energy for a time interval Δt in Joule is $Q_e = \Phi_e \cdot \Delta t$. The energy of single photon in Joule is $E = \hbar c/\lambda_s \approx 3.918 \cdot 10^{-19}$, where \hbar is the Planck constant, c the speed of light and $\lambda_s = 507$ nm the wavelength of the theoretically luminance-equivalent monochromatic source. Finally, the number of photons entering the pupil is $P = Q_e/E$.

To determine the number of rods that are covered by a pixel projection, we assume that a 24" display with 1920×1200 pixels is observed from the distance $d_s = 0.6$ m. We also assume a density of 100,000 rods/mm². We chose this value as a representative average density, since the rod acuity peak has a density of 150,000 rods/mm² [Jonas et al. 1992] at the eccentricity of 1.5 mm from the fovea center [Mariani et al. 1984].

Based on these assumptions and knowing that the spatial extent of one visual degree corresponds to approximately 0.288 mm on the retina [Drasdo and Fowler 1974], we derive that roughly $\rho = 5$ rods are covered by each pixel. Given an ideal optical focus each rod can only see a single pixel. Therefore no additional compensation such as number of visible pixels has to be considered. Further, assuming a perfect focus in the eye optics, the number of photons per rod can be approximated as $N_r = P/\rho$.

As our model is fitted for a retinal area covering 500 rods, we set the coefficient ϕ for the conversion from the input HDR luminance L to the photon count N as $\phi = 500 \cdot N_r$, hence $N = \phi L$. Using our setup $\phi \approx 1.2 \cdot 10^5$ for 10^{-3} cd/m² and $N \approx 120$ photons.

Discussion The goal of this chapter is not a strict simulation of a complete perceptual pipeline for night vision, although our spatio-temporal model of handling photons at rods could potentially serve as input for higher-order processes. Using a photon-unit scale is a means to offer control over the day-for-night processing.

10.2.2 Simulation

The simulation is performed independently for all pixels in time steps matching the rendering framerate. Relying on our analytical model, this choice is both practical and performance-efficient. In the following, we will discuss the simulation outcome for a single receptor.

Eq. 10.3 is used to sample the number N' of photons perceived, depending on the number N of photons. It is not a Poisson process (contrary to simple shot-noise in Eq. 10.1, or dark noise-model in Eq. 10.2) and analytically drawing samples is not straightforward. As the evaluation is needed for all pixels per frame, an efficient procedure is required. To make sampling tractable, we use an inversion via a lookup table.

First, the values $P_{\text{All}}(k, N)$ of Eq. 10.3 are tabulated for all values 0 to k and all values 0 to N . From this table, a complementary cumulative sum $F_{\text{All}}(k, N) = \sum_{i=1}^N P_{\text{All}}(k, i)$ is created numerically. Note, that each row in Eq. 10.3 is already a PDF and its integral is 1. The inverse of each row, $F_{\text{All}}^{-1}(\xi, N) = \min\{k | F_{\text{All}}(k, N) > \xi\}$, is stored as a look-up table.

The lookup table is constructed offline, but we also provide it on our website¹. To convert the physical photon count N into a photo-transduced photon count N' , a random

¹<http://resources.mpi-inf.mpg.de/DarkNoise>

number $\xi \in [0, 1]$ is generated and used to look up $N' = F_{\text{All}}^{-1}(\xi, N)$ in constant time.

As the values $\alpha = 0.5$, $q = 0.2$, $D = 19$ and $\Theta = 15$ were derived for stimuli that covered 500 receptors, and a duration of 100 ms [Hecht et al. 1942, Lillywhite 1981], the number N' computed above is valid for 500 receptors and 100 ms. However, we would need to apply a conversion to a single receptor, but computing the response of every individual receptor is computationally costly. Further, we actually should consider ρ receptors covered by a pixel ($\rho \approx 5$ for the display in our experiment). To accelerate the computation, we assume that the probability for an observation is uniform in a spatial neighborhood of 500 receptors in a time window of 100 ms and that observation events are independent between different receptors. Under these conditions, the probability that a pixel observes M events is given by the binomial distribution

$$P_{\text{Final}}(M, N') = \binom{M}{N'} (\rho/500)^M (1 - \rho/500)^{N'-M}.$$

Again, a single sample M' is drawn from this distribution using $M' = F_{\text{Final}}^{-1}(\xi, N')$ using the inversion method and a lookup table for each N' .

Finally, the number of transduced photons for this pixel M' is converted back to a displayable value. At this point, we have to account for the factor ϕ that relates luminance L' and photon counts, as well as for the quantum efficiency that reduced the photon count due to the eye optics: $L' = \phi^{-1} \cdot q^{-1} \cdot (\rho/500)^{-1} \cdot M'$. In order to preserve chroma, the tone-mapped RGB values are first converted to YCrCb, the noise is applied to the luminance Y and the resulting Y'CrCb is converted back to RGB. The noise is determined by the ratio of the photon count M' and the expected photon count given by the HDR luminance L . It is applied to Y as a combination of gain and bias, where the gain represents the multiplicative noise from the light quantization and depends on the size of qN , and the bias represents the additive photon-like noise and depends on the dark-noise constant D , as well as a noise baseline K

$$Y' = \frac{M'}{(qN + D)\rho/500} \left[\left(\frac{qN}{qN + D} \right) Y + \left(1 - \frac{qN}{qN + D} \right) K \right].$$

Applying the noise to a toned-mapped image provides fine appearance control, as we can choose the noise intensity K in totally black regions, where no evidence about the absolute scale is available. $K = 0.06$ was used in our results. Photopic and mesopic conditions are practically noise-free and seamlessly covered by our simulation because $L \approx L'$, as dark noise can be neglected when $N \gg D$ and the standard deviation of quantum noise is small for $N \gg 0$. Because both, Poisson and binomial distributions, converge to the normal distribution for sufficiently large samples, we keep only up to 1000 values in the lookup tables and samples from larger distributions are drawn using the Box-Muller method.

10.2.3 Temporal integration

The temporal resolution at scotopic conditions is low [Wandell 1995, Fig. 7.23] and even lower close to absolute threshold [Umino et al. 2008, Fig. 1] (unfortunately, in the range 10^{-6} – 10^{-5} cd/m² we were not able to find the relevant data for human vision). Consequently, noise undergoes filtering of temporal frequencies above 10 Hz. A simple solution would store the last 100 ms and average them. Instead, we use a closed-form solution routinely applied in graphics [Kass and Pesare 2011]: To simulate the current frame, the old frame is blended with the new one weighted by $\alpha = \exp(-2\pi \cdot f_c/f_s)$,

and $1 - \alpha$ [Smith 1997], where $f_c = 0.5$ Hz is the cutoff frequency and f_s is the frame rate. The cutoff of $f_c = 0.5$ was tuned by manual inspection to achieve a result most similar to averaging multiple frames.

As the noise process occurs on the retina, the noise pattern is expected to move with the eye over the image. In the absence of eye tracking, we make the assumption that the eye follows the optical flow for given pixel [Kass and Pesare 2011] during the integration-time period. We warp the noise pixel-by-pixel along this flow, including a depth test if depth is available.

Pixel-by-pixel warping is used, as repeated warping of an image at a high framerate would quickly blur out all high spatial frequencies that are important for our noise. However, this warping, as well as disocclusion (when depth is used), results in holes.

To address this issue, we *jitter* the target position of every pixel by a small random offset of one pixel. Doing so reduces the regular structure, which would become apparent in a smooth flow field. Further, we *fill* the remaining holes with new noise values, but, as they did not undergo temporal integration, a careful choice is needed, otherwise, their brightness statistics would differ from the warped pixels. One approach would be to draw multiple samples over time and average them over the integration period. A more efficient solution is to directly change the distribution from which these hole-filling samples are drawn to match the mean and standard deviation of the temporally-integrated distribution. Such a distribution can be obtained by properly scaling the standard simulation time of 100 ms and the number of photons. Intuitively, larger time leads to higher mean values in the simulation and therefore lower relative noise. As our integration procedure is an exponential smoothing filter and our cumulative distribution behaves like a box filter, we can find a proper scaling for the simulation time by looking for a box filter length such that the corresponding exponential and uniform distribution mean values and standard deviations are equal. The resulting scaling factor is $(1 + \alpha)/(1 - \alpha)$.

10.3 Validation

Acquiring a reference noise for comparison is impossible; it exists solely as a neural representation, for which no imaging technology is available. This section complements the quantitative fit to physiological data, which we have provided so far, with performance evaluations, a qualitative assessment in form of actual images, and a perceptual experiment.

10.3.1 Performance

Our implementation computes an HD noise frame from an input image in 18 ms on a Nvidia Geforce 660 GTX. Most time is spent on the image-warping with respect to the estimated eye movement (9 ms). Producing samples from the distribution is fast when using the lookup tables (1.8 ms).

10.3.2 Results and discussion

Fig. 10.1, Fig. 10.4 and Fig. 10.5 show typical results, compared to other noise models at different scales of absolute luminance. Gaussian noise follows the recommendation by Thompson et al. [2002]; adding Gaussian noise with a small standard deviation, independent of the luminance in the image. We see that the noise does not adapt to the

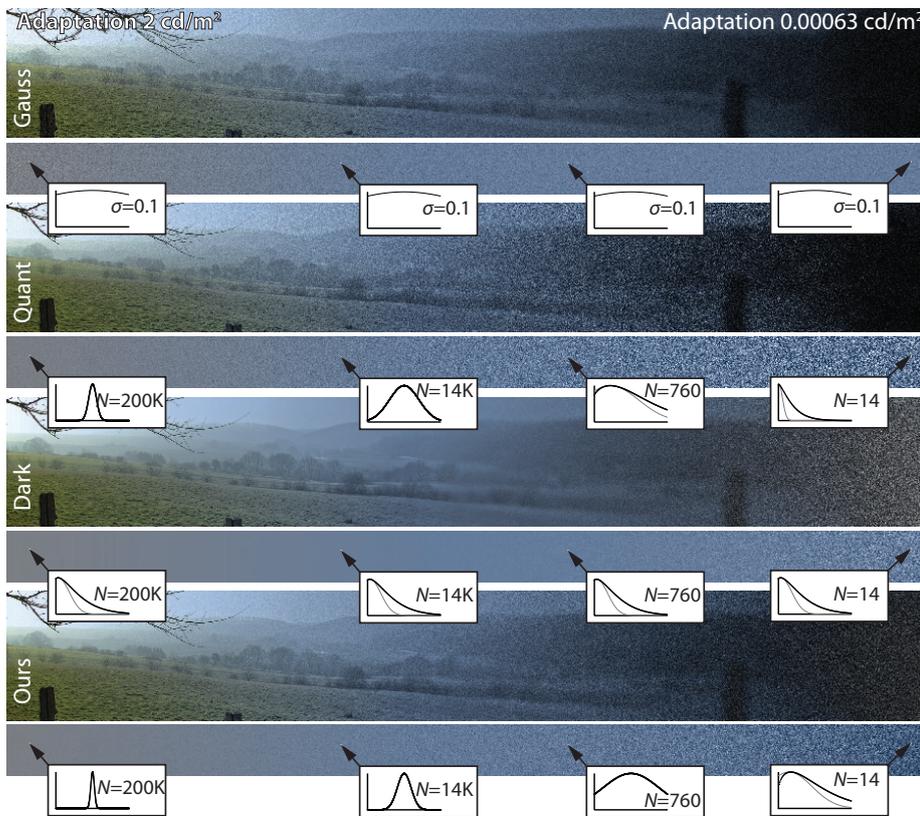


Figure 10.4. Gaussian [Thompson et al. 2002], quantum (Eq. 10.1), dark noise (Eq. 10.2) and our model (top to bottom), applied to an image that contains different absolute scales of luminance (*horizontal*), including a condition close to absolute threshold. Insets show the power distribution of the noise (*black*) and a Gaussian reference (*grey*) at four specific pixels together with their photon count N .

image content and it is unclear how it should be scaled in respect to adaptation. The quantum and dark noise show implementations of Eq. 10.1 and Eq. 10.2 respectively. The quantum noise reacts to image content but lacks noise in dark areas. The dark noise has an inverted behavior. Only our model combines all properties into a consistent omnipresent noise that reacts to luminance. Note, that Fig. 10.1 and Fig. 10.4 span a range of adaptation levels for didactic purposes, while images only have one dominant adaptation level in practice. For animated versions of these two figures please refer to our website.

10.3.3 Perceptual experiment

Motivation Introducing artifacts in images or videos can be useful for artistic purposes. Considering that a large number of artifacts are intentionally introduced in almost all movies or games – such as depth-of-field, motion blur, glare, and in particular the addition of noise (e.g. the motion pictures *Hugo*, *300*, *Pi*, *Planet Terror*, *Saving Private Ryan* or the game *Limbo*) – it is clear that there is a substantial artistic demand. Ultimately, it is up to the artist to decide if noise should be included. But in case the

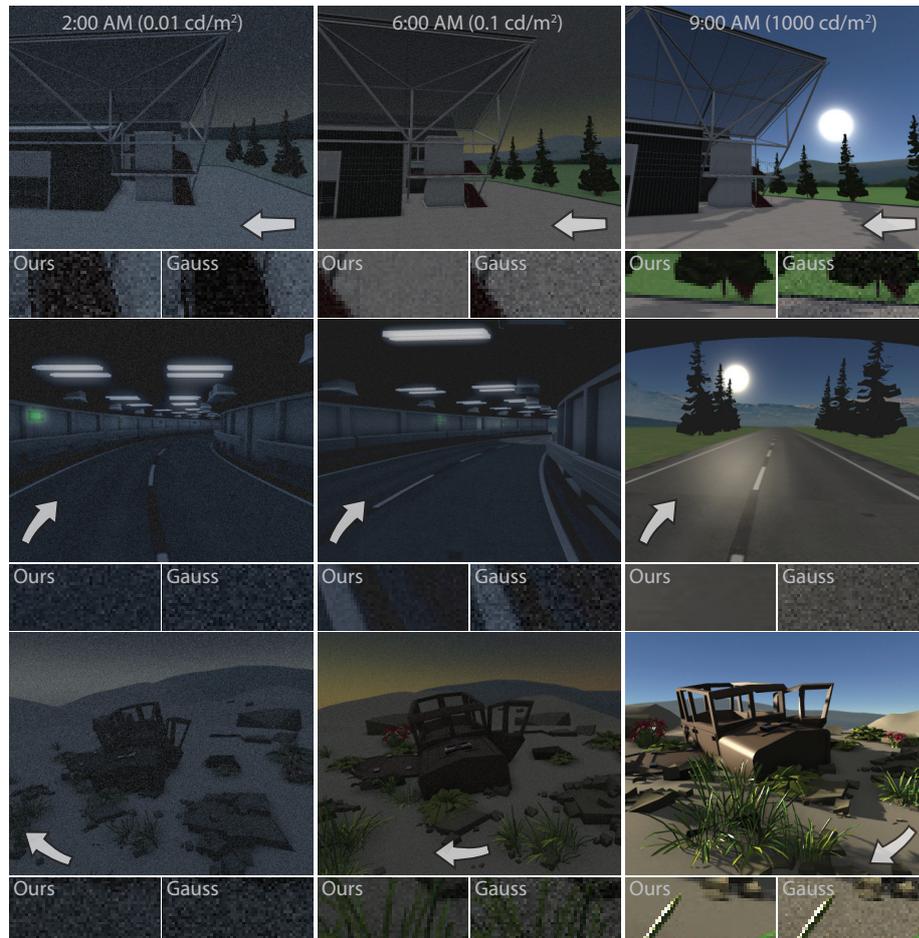


Figure 10.5. Results of our full model applied to different CG animations (*Rows*) with adaptation luminance changing over time (*Columns*). The first row is a time-lapse architectural visualization from night over morning to daylight. The second row is a driving simulation in a tunnel. The last row shows a time-lapse animation of a three-dimensional animated 3D scene with a setting similar to the photography used by Thompson et al. [2002]. The pairs of insets (the left is ours, the right is Gaussian noise) show a magnified part of the frames above. Our approach changes over time, adapts to luminance changes, and does not suffer from the shower-door effect, which is typical of screen-space patterns, such as Gaussian noise, which only works well for a photo with a fixed luminance level.

noise is desired, our work is first to explain how to include it properly and its cause. This is why our study focuses on comparison of four possible methods for simulation of the scotopic noise rather than answering the question whether any presence of noise is subjectively preferred by our particular subjects.

Methodology The key questions using our noise model concern the nocturnal mood impression (realism), viewing comfort, and the observer's overall preference. To this extent, we performed a perceptual experiment, which was preceded by a calibration phase where dark-adapted subjects could possibly experience scotopic noise themselves. Afterwards, subjects were shown videos with day-for-night tone mapping applied, with either (1) white additive noise as suggested by Thompson et al. [2002] or (2) our full model of noise distribution temporarily simulated either as (1) a static noise frame (a single white noise pattern or our full model with a constant random seed), (2) a dynamically changing phenomena. Subjects were shown pairs of videos and asked, which one: (1) depicts the scene more realistically? (2) is more comfortable to look at? (3) is preferred in general?

Stimuli To explore the dynamic aspects of noise and its interaction with the image content the video sequences exhibit camera motion. In total four different short movies (10 s) were used; two computer-generated (ARCHITECTURE, TUNNEL) and two captured (COUNTRYSIDE, CAR). The animation in CAR was produced by horizontally panning across the teaser in Thompson et al. [2002]. Three of the videos (ARCHITECTURE, COUNTRYSIDE, TUNNEL) contain temporal changes of absolute luminance (see the video on our website).

Procedure Ten subjects took part in the experiment comprising a *calibration* and a *query* phase. In the calibration phase, they were instructed to adapt for 10 minutes in a dark room. It was done merely to let them investigate the appearance of several presented objects and at the same time experience the scotopic noise for themselves. While more time is typically required to achieve full scotopic adaptation [Ferwerda et al. 1996, Pattanaik et al. 2000], we found scotopic noise to become apparent already after ten minutes in our setting. Longer adaptation is expected to produce an even stronger effect, but results in fatigue. The query phase was performed in photopic conditions under controlled dim office illumination. Subjects were shown all of the $4 \times 2 \times 2$ combinations of the above stimuli and noise variants in a random order. Stimuli involving changes in distribution and temporal behavior simultaneously were skipped to reduce fatigue. Videos were looped and after three repetitions (30 s), subjects were asked to answer the three questions stated above. The used display was a Dell U2412M with a resolution of 1920×1200 pixels and stimuli were played at 30 Hz. They were shown next to each other at a resolution of 800×800 pixels in front of a dark background, at a distance of 60 cm, at which a pixel covers a visual angle of 1.45 arcmin. Subjects were adapted to dim, photopic office-lighting conditions.

Results Full breakdown of the study results for individual scenes is available in Tbl. 10.1.

First, we compare our full approach to a variant using static instead of dynamic noise (Fig. 10.6a). It is significantly preferred (all significant effects reported are, $p < .05$ binomial test) overall (63.7%, CI [11.5, 10.5]%) and, in particular, in terms of realism (68.8%, CI [11.3, 9.9]%), while no significant effect on comfort was present (52.5%,

Stimuli	Comparison	Realism	Comfort	Preference
ARCHITECTURE	Ours dyn. × White dyn.	* 0.85 CI [0.23, 0.12]	* 1.00 CI [0.17, 0.00]	* 0.95 CI [0.20, 0.05]
	Ours dyn. × Ours stat.	0.65 CI [0.24, 0.20]	0.55 CI [0.23, 0.22]	0.65 CI [0.24, 0.20]
	Ours stat. × White stat.	0.60 CI [0.24, 0.21]	* 0.80 CI [0.24, 0.14]	* 0.75 CI [0.24, 0.16]
CAR	Ours dyn. × White dyn.	0.30 CI [0.18, 0.24]	* 0.20 CI [0.14, 0.24]	* 0.25 CI [0.16, 0.24]
	Ours dyn. × Ours stat.	0.65 CI [0.24, 0.20]	0.55 CI [0.23, 0.22]	0.65 CI [0.24, 0.20]
	Ours stat. × White stat.	* 0.20 CI [0.14, 0.24]	* 0.15 CI [0.12, 0.23]	* 0.15 CI [0.12, 0.23]
COUNTRYSIDE	Ours dyn. × White dyn.	0.60 CI [0.24, 0.21]	* 0.75 CI [0.24, 0.16]	* 0.75 CI [0.24, 0.16]
	Ours dyn. × Ours stat.	0.70 CI [0.24, 0.18]	0.40 CI [0.21, 0.24]	0.60 CI [0.24, 0.21]
	Ours stat. × White stat.	* 0.85 CI [0.23, 0.12]	* 0.75 CI [0.24, 0.16]	0.70 CI [0.24, 0.18]
TUNNEL	Ours dyn. × White dyn.	0.55 CI [0.23, 0.22]	* 0.95 CI [0.20, 0.05]	* 0.95 CI [0.20, 0.05]
	Ours dyn. × Ours stat.	* 0.75 CI [0.24, 0.16]	0.60 CI [0.24, 0.21]	0.65 CI [0.24, 0.20]
	Ours stat. × White stat.	0.60 CI [0.24, 0.21]	* 1.00 CI [0.17, 0.00]	* 1.00 CI [0.17, 0.00]
All	Ours dyn. × White dyn.	0.57 CI [0.12, 0.11]	* 0.72 CI [0.11, 0.09]	* 0.72 CI [0.11, 0.09]
	Ours dyn. × Ours stat.	* 0.69 CI [0.11, 0.10]	0.53 CI [0.11, 0.11]	* 0.64 CI [0.12, 0.10]
	Ours stat. × White stat.	0.56 CI [0.12, 0.11]	* 0.68 CI [0.11, 0.10]	* 0.65 CI [0.11, 0.10]

Table 10.1. Study results with relatives scores and 95 % confidence intervals (CI) for individual stimulus. Scores above 0.5 mark the first noise to be perceived as better according to the given criteria. Stars denote statistical significance ($p < .05$ binomial test).

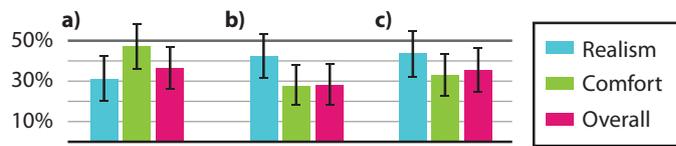


Figure 10.6. Study statistics for three different comparisons (*a-c*) and three qualities (*colors*). Each bar's height denotes preference compared to a reference. Notches denote 95 % confidence intervals (CI). Comparison with a CI not intersecting the 50 %-null hypothesis line are statistically significant.

CI [11.5, 11.3] %). The finding indicates that adding dynamic noise is useful because static noise is perceived unnatural, in particular for dynamic scene.

Second, we compare our full approach to dynamic white noise (Fig. 10.6b). Again it is found to be significantly better overall (72.5 %, CI [11.1, 9.4] %) and in terms of comfort (72.5 %, CI [11.1, 9.4] %), while the improvement in realism is not significant (57.5 %, CI [11.6, 11.0] %). The finding indicates that adding dynamics alone is not preferred over adding it to the appropriate noise distribution. Probably, uncorrelated white noise fluctuations are perceived unnatural as they not adjust to the image content.

The comparison of a static variant of our noise and static white noise (Fig. 10.6c), leads to a significant preference of the static variant of our approach in terms of comfort (67.5 %, CI [11.4, 10.1] %) and preference (65.0 %, CI [11.5, 10.3] %) with no significant effect on realism (56.3 %, CI [11.6, 11.1] %). The finding indicates that besides all dynamics, our choice of physiologically-principled noise is an important factor; only a mix of additive and multiplicative noise, as well as adaption to the actual luminance seems to appear plausible.

The per-scene results in Tbl. 10.1 show that our method is performing the best in the computer graphics stimuli (ARCHITECTURE and TUNNEL). This confirms that the advantages of our model are most prominent in combination with a dynamic content while noise application to a static image is more forgiving (CAR and COUNTRYSIDE).

In summary, the experiments indicate, that previous work adds to the nocturne mood in static images, but might be incomplete for animated imagery. Further, the noise dynamics as a function of scene content is not trivial and the type of noise distribution leads to perceivable differences. Still, extensive noise can reduce viewing comfort and, ultimately, if an artist decides to use noise to depict scotopic conditions, a tradeoff is possible.

10.4 Conclusions

We derived a physiologically-motivated model of noise perception close to the absolute luminance threshold. The model is practical and can be computed efficiently. Our specialized warping maintains noise details and leads to temporal coherence and might be useful in other contexts or other forms of temporally-coherent high-frequency noise. The experimental evaluation shows that our simulated noise is always overall preferred, and more comfortable to watch than previous solutions, which are based on white noise. Our dynamic-noise solution can be potentially less comfortable than its static counterpart, but it consistently improves realism. The artistic intent should be the key factor when choosing between apparent realism and viewing comfort. Our model does not yet account for any higher-level effects. We assume that the day-for-night

tone mapping (chroma change, acuity loss) is independent of and happens before the receptor noise. While this is physiologically plausible, future work could account for higher-level processes in order to reproduce all major scotopic phenomena, including also the Purkinje shift, and the scotopic (temporal) contrast, as a consequence of the noisy retinal signal processing itself.

Chapter 11

Summary

The thesis presented novel contributions in two areas of stereoscopic 3D applications. Here we offer their summary and an outlook on potential directions for a future research in the area.

11.1 Conclusion

Display and entertainment systems undergo a steady development and each year a new exciting technology pushes the viewer experience further. At the same time each such technology presents new challenges for the content production, and a new methodology has to be established to make a good use of its benefits and suppress its possible drawbacks. The stereoscopic 3D was not an exception and the difficulty of its proper application could be seen as it repeatedly emerged and got abandoned in the past. The fact that the recent resurrection of stereoscopic 3D is commercially successful can be explained by both the technological progress and our better understanding of human perception together with the existence of suitable production rules. As no technology lives in isolation this thesis focused on interactions of stereoscopic 3D with other features of modern displays. Perceptual observations formalized as computational models, and their conclusions turned into practical algorithms provide useful guides for authors of new interactive applications or HDR movies.

As our main goal is an improvement of the image or video quality when judged by a human observer, it is vital to understand how human perceives a visual content in various conditions such as static versus moving pictures or bright versus dark environment. All methods presented in this thesis were introduced by an extensive research of the psychophysical background which allowed us to identify those aspects that are important for the selected perceptual quality in our images. Regardless of the vast extent of previous research there is usually no exact match between previous experiments and intended applications in this thesis. We have identified such gaps and filled them with our own experimental data and models built upon them.

This thesis presented a series of such customized or completely novel measurements and models. We have shown that the preservation of disparity discontinuities and their alignment with luminance edges is of a high importance when the overall subjective quality of the stereoscopic footage is in question. We state that the perception of artifacts is significantly more difficult for video sequences than static images. Our experiments with the disparity shift and scale have shown that humans are not extremely sensitive to temporal disparity changes and that its manipulations can remain invisible if performed over a sufficient time span. We have provided the first complete model of the

interaction between disparity and motion parallax as two distinct depth cues in dynamic videos and we have measured their joint contribution to a sensation of depth. Our work with reflective and refractive materials have shown that humans may judge physically incorrect but visually more pleasant renderings as more realistic. We have extended existing models of the disparity sensitivity to high dynamic range of the adaptation luminance and found that the ability to perceive disparity quickly degrades for very dark levels even beyond what was predicted by previous models. Finally, we have shown that the temporal characteristic of the noise present in the extreme darkness vision is important for its realism in a simulation.

Each such finding has a value for learning more about the human vision on its own. However, we have conducted this research with the aim to solve practical computer graphics problems. We have used the psychophysical models either to make design choices for computational algorithms or as an integral part of optimization tasks. The false depth from the Pulfrich effect and from the sequential presentation was eliminated through a temporal compensation of content. We have included a model of motion in depth sensitivity as a part of an optimization task to reconstruct the correct motion. Our seamless gaze-driven disparity mapping was able to convey more depth in regions attended by the viewer without visible artifacts when the fixation moved to a different location. This was achieved using our model of sensitivity to temporal disparity changes tailored for this particular task. We have redistributed the disparity to enhance the overall 3D appearance of a stereoscopic video using our model predicting relative contribution of disparity and motion parallax to the overall depth percept. The relation between contrast and disparity perception, limits of vergence-accommodation conflict tolerance and the mechanism behind binocular fusion was used to steer the rendering of reflective and refractive surfaces. Our model of stereoscopic vision under a wide range of luminance allowed for a computational retargeting of disparity between different adaptation conditions. Finally, we have shown how measurements of statistical properties in visual noise can be turned into a dynamic post-process effect that simulates this phenomenon to increase picture realism.

All together we have thoroughly explored effects that motion and HDR have for stereoscopy. We have visited issues connected with displaying temporarily changing information on the low level of a display, then we explored the perception of motion in depth which is relying on the disparity rather than luminance. Our gaze-driven remapping aimed on self-induced motion based on the decision of a human observer rather than the one pre-defined by the content. As the most complex case of motion we have modeled motion parallax and we have utilized it as an additional depth cue jointly working alongside with disparity. Finally, we have covered the entire range of HDR luminance conditions when we explored rendering of shiny surfaces in photopic conditions, offered a model and a simulation for the transition between photopic and scotopic conditions and concluded this thesis by a description and a simulation of the vision near the absolute threshold.

This way we have shown that looking on individual image properties separately does not uncover the entire story and that focusing on their mutual relations, support or conflicts can lead to an improvement of the overall subjective quality. We believe that findings from our work will be of use for designers of new generation of HDR stereoscopic displays and their content. Such devices should enable an interactive motion response and support of gaze-tracking as expected in the next generations of head mounted displays for virtual or augmented reality. There, stereoscopy is a natural and necessary feature.

11.2 Future work

In this thesis we addressed some of the most important aspects of stereoscopic perception in context of motion and HDR. However, it is clear that many questions still remain opened. Furthermore, motion and HDR are just two of critical requirements for an ultimate display that will be able to reproduce images indistinguishable from the reality. Light field displays are one of directions towards this goal. At the time they are mostly utilized to distribute stereoscopic experience to one or more observers at once without need of an eye-wear. In future, a higher angular density could also allow for a proper reproduction of the accommodation cue and this way provide a natural stereoscopic viewing without the uncomfortable accommodation-vergence conflict. On the other hand, such a display would require a large amount of data and computation. A perceptual modeling similar to ours can be deployed to predict rendering quality requirements as a function of the viewer and the image. Likewise in our methods, one can utilize properties of luminance, motion or gaze information to redistribute the computational effort to locations where it is the most needed.

At this point head mounted displays are of a high research interest as they offer natural interaction and large field of view together with stereoscopy. Unlike traditional displays they do not present a clear display plane because the focal distance is usually modified using lens to lie in infinity. The immersion also alters perception of absolute scales. Unlike on a small screen it is important to match the field of view and the size of objects in the world to the reality. How large disparity can be reproduced and what consequences does its alternation have on the user comfort and performance in interactive tasks requiring orientation in a 3D virtual world is yet not fully researched.

HDR is now just being introduced into consumer TVs. The technology used for stereoscopy makes its combination with HDR difficult at the time as lot of luminance power is lost during the signal multiplexing at LCD panels. In future we expect this to be solved and HDR to become widely adapted as a de facto standard. This thesis focuses mainly on the ability of HDR to reproduce viewing conditions difficult for the human vision. Although the stereo performance was not observed to rise beyond standard photopic conditions it was reported that HDR makes images look more vivid, real and possibly deep [Vangorp et al. 2014]. A further investigation of mechanisms behind such effect and quantification of its contribution to depth are still missing. It could be potentially combined into the disparity processing framework in a similar way as motion parallax in Chapter 7.

As there are many other depth cues in complex images, it would also be interesting to provide such relative scaling to a more exhaustive subset of them. A knowledge of the relative contribution of cues such as texture gradients, defocus blur, aerial perspective, occlusions, scale or shading to the mental depth image in comparison to the disparity would allow for a design of even more powerful disparity mapping and compression algorithms that would be able to display 3D images with relatively shallow disparity and therefore they would avoid both discomfort and artifacts on current displays. The main challenge in doing this is the formulation of a meaningful parameterization for each of them as unlike disparity they are not available directly in the form of a scalar map but they have to be extracted through methods of image processing. We illustrated the tractability of such task by completing these steps for motion parallax where similar challenges had to be met.

Bibliography – Own Work

- Dąbala, Ł., Kellnhofer, P., Ritschel, T., Didyk, P., Templin, K., Myszkowski, K., Rokita, P., and Seidel, H.-P. 2014. Manipulating refractive and reflective binocular disparity. *Computer Graphics Forum (Proc. Eurographics 2014)*, 33(2), 53–62. ISSN 1467-8659. doi: 10.1111/cgf.12290.
- Kellnhofer, P., Ritschel, T., Myszkowski, K., and Seidel, H.-P. 2013. Optimizing disparity for motion in depth. *Computer Graphics Forum (Proc. EGSR 2013)*, 32(4), 143–152. ISSN 1467-8659. doi: 10.1111/cgf.12160.
- Kellnhofer, P., Ritschel, T., Myszkowski, K., and Seidel, H.-P. 2014a. Improving perception of binocular stereo motion on 3D display devices. *Proc. SPIE*, 9011, 901116–901116–11. doi: 10.1117/12.2032389.
- Kellnhofer, P., Ritschel, T., Vangorp, P., Myszkowski, K., and Seidel, H.-P. 2014b. Stereo day-for-night: Retargeting disparity for scotopic vision. *ACM Trans. Appl. Percept.*, 11(3).
- Kellnhofer, P., Leimkühler, T., Ritschel, T., Myszkowski, K., and Seidel, H.-P. 2015a. What makes 2D-to-3D stereo conversion perceptually plausible? In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception, SAP '15*, pages 59–66, New York, NY, USA. ACM. ISBN 978-1-4503-3812-7. doi: 10.1145/2804408.2804409.
- Kellnhofer, P., Ritschel, T., Myszkowski, K., Eisemann, E., and Seidel, H.-P. 2015b. Modeling luminance perception at absolute threshold. *Computer Graphics Forum (Proc. EGSR 2015)*, 34(4), 155–164. ISSN 1467-8659. doi: 10.1111/cgf.12687.
- Kellnhofer, P., Ritschel, T., Myszkowski, K., and Seidel, H.-P. 2015c. A transformation-aware perceptual image metric. *Proc. SPIE*, 9394, 939408–939408–14. doi: 10.1117/12.2076754.
- Kellnhofer, P., Didyk, P., Myszkowski, K., Hefeeda, M. M., Seidel, H.-P., and Matusik, W. 2016a. GazeStereo3D: Seamless disparity manipulations. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 35(4). doi: 10.1145/2897824.2925866.
- Kellnhofer, P., Didyk, P., Ritschel, T., Masia, B., Myszkowski, K., and Seidel, H.-P. 2016b. Motion parallax in stereo 3D: Model and applications. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia 2016)*, 35(6). doi: 10.1145/2980179.2980230.
- Kellnhofer, P., Ritschel, T., Myszkowski, K., and Seidel, H.-P. 2016c. Transformation-aware perceptual image metric. *Journal of Electronic Imaging*, 25(5), 053014. doi: 10.1117/1.JEI.25.5.053014.

- Khosla, A., Krafcik, K., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., and Torralba, A. June 2016. Eye tracking for everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA.
- Leimkühler, T., Kellnhofer, P., Ritschel, T., Myszkowski, K., and Seidel, H.-P. 2016. Perceptual real-time 2D-to-3D conversion using cue fusion. In *Proceedings of the 2016 Graphics Interface Conference, GI '16*. Canadian Information Processing Society.

Bibliography

- Alpern, M. 1971. Rhodopsin kinetics in the human eye. *J Phys.*, 217(2), 447–471.
- Altman, J. 1977. The sensitometry of black and white materials. In *The Theory of the Photographic Process*.
- Ashmore, F. and Falk, G. 1977. Dark noise in retinal bipolar cells and stability of rhodopsin in rods. *Nature*, 270, 69–71.
- Aydin, T. O., Stefanoski, N., Croci, S., Gross, M., and Smolic, A. November 2014. Temporally coherent local tone mapping of hdr video. *ACM Trans. Graph.*, 33(6), 196:1–196:13. ISSN 0730-0301. doi: 10.1145/2661229.2661268.
- Banks, M. S., Gepshtein, S., and Landy, M. S. 2004. Why is spatial stereoresolution so low? *The Journal of neuroscience*, 24(9), 2077–2089.
- Banks, M., Sekuler, A., and Anderson, S. 1991. Peripheral spatial vision: Limits imposed by optics, photoreceptors, and receptor pooling. *J Opt Soc Am A*, 8(11), 1775–87.
- Barlow, H. B. 1956. Retinal noise and absolute threshold. *J. Opt. Soc. Am.*, 46(8), 634–639.
- Baylor, D., Lamb, T., and Yau, K.-W. 1979. Responses of retinal rods to single photons. *J Phys.*, 288(1), 613–34.
- Beard, T. D. 05 1991. Low differential 3-D viewer glasses and method. Patent. http://www.patentlens.net/patentlens/patent/EP_0325019_B1/en/. EP 0325019 B1.
- Becker, W. and Juergens, R. 1975. Saccadic reactions to double-step stimuli: Evidence for model feedback and continuous information uptake. In *Basic Mechanisms of Ocular Motility and their Clinical Implications*, pages 519–527.
- Bernhard, M., Dell'mour, C., Hecher, M., Stavrakis, E., and Wimmer, M. 2014. The effects of fast disparity adjustment in gaze-controlled stereoscopic applications. In *Proc. Symp. on Eye Tracking Research and Appl. (ETRA)*, pages 111–118.
- Bista, S., da Cunha, Í. L. L., and Varshney, A. 2016. Kinetic depth images: flexible generation of depth perception. *The Visual Computer*, pages 1–13.
- Blake, A. and Bülthoff, H. 1990. Does the brain know the physics of specular reflection? *Nature*, 343(6254), 165–168.

- Bloch, A. M. 1885. Experience sur la vision. *C.r. Séanc. Soc. Biol.*, 37, 493–495.
- Borji, A. and Itti, L. 2013. State-of-the-art in visual attention modeling. *IEEE PAMI*, 35(1), 185–207.
- Bradshaw, M. F. and Rogers, B. J. 1999. Sensitivity to horizontal and vertical corrugations defined by binocular disparity. *Vision Res.*, 39(18), 3049–56.
- Bradshaw, M. F. and Rogers, B. J. 1996. The interaction of binocular disparity and motion parallax in the computation of depth. *Vision Research*, 36(21), 3457–3468.
- Bradshaw, M. F., Hibbard, P. B., Parton, A. D., Rose, D., and Langley, K. 2006. Surface orientation, modulation frequency and the detection and perception of depth defined by binocular disparity and motion parallax. *Vision Research*, 46(17), 2636–2644.
- Braunstein, M. L., Hoffman, D. D., and Pollick, F. E. 1990. Discriminating rigid from nonrigid motion: Minimum points and views. *Perception & Psychophysics*, 47(3), 205–214.
- Brenner, E., van den Berg, A., and van Damme, W. 1986. Perceived motion in depth. *Vis. Res.*, 36, 699–706.
- Brookes, A. and Stevens, K. A. 1989. The analogy between stereo depth and brightness. *Perception*, 18(5), 601–614.
- Brooks, K. R. and Stone, L. S. 2004. Stereomotion speed perception: Contributions from both changing disparity and interocular velocity difference over a range of relative disparities. *J. Vis.*, 4(12).
- Brox, T., Bruhn, A., Papenbergh, N., and Weickert, J. 2004. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision (ECCV)*, volume 3024 of *Lecture Notes in Computer Science*, pages 25–36.
- Burt, P. J. and Adelson, E. H. 1983. The laplacian pyramid as a compact image code. *IEEE Trans. Comm.*, 31(4), 532–40.
- Butler, D. J., Wulff, J., Stanley, G. B., and Black, M. J. October 2012. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag.
- Campbell, F. and Maffei, L. 1981. The influence of spatial frequency and contrast on the perception of moving patterns. *Vision Research*, 21(5), 713–721.
- Cavallo, V. and Laurent, M. 1988. Visual information and skill level in time-to-collision estimation. *Perception*, 17(5), 623–32.
- Chamaret, C., Godeffroy, S., Lopez, P., and Le Meur, O. 2010. Adaptive 3D rendering based on region-of-interest. In *Proc. SPIE vol. 7524*, pages 0V–1–12.
- Chapiro, A., Heinzle, S., Aydın, T. O., Poulakos, S., Zwicker, M., Smolic, A., and Gross, M. 2014. Optimizing stereo-to-multiview conversion for autostereoscopic displays. *Computer Graphics Forum*, 33(2), 63–72. ISSN 1467-8659. doi: 10.1111/cgf.12291.

- Cisarik, P. M. and Harwerth, R. S. 2005. Stereoscopic depth magnitude estimation: Effects of stimulus spatial frequency and eccentricity. *Behavioural Brain Research*, 160(1), 88–98.
- Cormack, L. K., Stevenson, S. B., and Schor, C. M. 1991. Interocular correlation, luminance contrast and cyclopean processing. *Vis. Res.*, 31(12), 2195–07.
- Coutant, B. E. and Westheimer, G. 1993. Population distribution of stereoscopic ability. *Ophthalmic and Physiological Optics*, 13(1), 3–7.
- Cumming, B. G. 1995. The relationship between stereoacuity and stereomotion thresholds. *Perception*, 24(1), 105–114.
- Cutting, J. E. 1995. Potency, and contextual use of different information about depth. *Perception of space and motion*, page 69.
- Deering, M. 2005. A photon accurate model of the human eye. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 24(3), 649–58.
- Deering, M., Winner, S., Schediwy, B., Duffy, C., and Hunt, N. 1988. The triangle processor and normal vector shader: a vlsi system for high performance graphics. In *Proc. of ACM SIGGRAPH*, pages 21–30.
- Didyk, P., Eisemann, E., Ritschel, T., Myszkowski, K., and Seidel, H.-P. 2010a. Perceptually-motivated real-time temporal upsampling of 3D content for high-refresh-rate displays. *Computer Graphics Forum (Proceedings Eurographics 2010, Norrköping, Sweden)*, 29(2), 713–722.
- Didyk, P., Ritschel, T., Eisemann, E., Myszkowski, K., and Seidel, H.-P. 2010b. Adaptive image-space stereo view synthesis. In *Vision, Modeling and Visualization Workshop*, pages 299–306, Siegen, Germany.
- Didyk, P., Ritschel, T., Eisemann, E., Myszkowski, K., and Seidel, H.-P. 2011. A perceptual model for disparity. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 30, 96:1–96:10.
- Didyk, P., Ritschel, T., Eisemann, E., Myszkowski, K., Seidel, H.-P., and Matusik, W. 2012. A luminance-contrast-aware disparity model and applications. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 31(6), 184:1–184:10.
- Drasdo, N. and Fowler, C. W. 1974. Non-linear projection of the retinal image in a wide-angle schematic eye. *J Ophthalmol.*, 58(8), 709–714.
- Duchowski, A. T., House, D. H., Gestring, J., Congdon, R., Świrski, L., Dodgson, N. A., Krejtz, K., and Krejtz, I. 2014a. Comparing estimated gaze depth in virtual and physical environments. In *Proc. Symp. on Eye Tracking Res. and Appl. (ETRA)*, pages 103–110.
- Duchowski, A. T., House, D. H., Gestring, J., Wang, R. I., Krejtz, K., Krejtz, I., Mantiuk, R., and Bazyluk, B. 2014b. Reducing visual discomfort of 3D stereoscopic displays with gaze-contingent depth-of-field. In *Proc. ACM Symp. on Appl. Perc. (SAP)*, pages 39–46.
- Durand, F. and Dorsey, J. 2000. Interactive tone mapping. In *Proc. EGWR*, pages 219–230.

- Durgin, F. H., Proffitt, D. R., Olson, T. J., and Reinke, K. S. 1995. Comparing depth from motion with depth from binocular disparity. *Journal of Experimental Psychology: Human Perception and Performance*, 21(3), 679.
- Eilertsen, G., Wanat, R., Mantiuk, R. K., and Unger, J. 2013. Evaluation of tone mapping operators for hdr-video. *Computer Graphics Forum*, 32(7), 275–284. ISSN 1467-8659. doi: 10.1111/cgf.12235.
- Eilertsen, G., Mantiuk, R. K., and Unger, J. October 2015. Real-time noise-aware tone mapping. *ACM Trans. Graph.*, 34(6), 198:1–198:15. ISSN 0730-0301. doi: 10.1145/2816795.2818092.
- Erkelens, C. and Collewijn, H. 1985. Motion perception during dichoptic viewing of moving random-dot stereograms. *Vis. Res.*, 25(4), 583–588.
- Fairchild, M. D. and Johnson, G. M. 2005. On the salience of novel stimuli: Adaptation and image noise. In *IS&T/SID 13th Color Imaging Conference*, pages 333–338.
- Farbman, Z., Fattal, R., Lischinski, D., and Szeliski, R. 2008. Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 27(3), 67:1–67:10.
- Fattal, R., Lischinski, D., and Werman, M. 2002. Gradient domain high dynamic range compression. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 21(3), 249–256.
- Fehn, C. 2004. Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In *Stereoscopic Displays and Virtual Reality Systems XI*, volume 5291, pages 93–104. SPIE.
- Ferwerda, J. A., Pattanaik, S., Shirley, P., and Greenberg, D. 1996. A model of visual adaptation for realistic image synthesis. In *Proc. SIGGRAPH*, pages 249–58.
- Field, G. D., Sampath, A. P., and Rieke, F. 2005. Retinal processing near absolute threshold: from behavior to mechanism. *Annu. Rev. Physiol.*, 67, 491–514.
- Filippini, H. R. and Banks, M. S. 2009. Limits of stereopsis explained by local cross-correlation. *J Vis.*, 9(1), 1–18. doi: 10.1167/9.1.8.
- Fisker, M., Gram, K., Thomsen, K. K., Vasilarou, D., and Kraus, M. 2013. Automatic convergence adjustment for stereoscopy using eye tracking. In *Eurographics 2013-Posters*, pages 23–24.
- Frisby, J. and Mayhew, J. 1978. Contrast sensitivity function for stereopsis. *Perception*, 7, 423–9.
- Geigel, J. and Musgrave, F. K. 1997. A model for simulating the photographic development process on digital images. In *Proc. SIGGRAPH*, pages 135–142.
- Geisler, W. S. and Perry, J. S. 1998. A real-time foveated multiresolution system for low-bandwidth video communication. In *Proc. SPIE vol. 3299*, pages 294–305.
- Gomila, C., Llach, J., and Cooper, J. 2013. Film grain simulation method. US Patent 8,447,127.
- Gray, R. and Regan, D. 1998. Accuracy of estimating time to collision using binocular and monocular information. *Vis. Res.*, 38(4), 499–512.

- Guenter, B., Finch, M., Drucker, S., Tan, D., and Snyder, J. 2012. Foveated 3D graphics. *ACM Transactions on Graphics (Proc SIGGRAPH Asia)*, 31(6), 164.
- Guo, C., Ma, Q., and Zhang, L. 2008. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society. doi: 10.1109/CVPR.2008.4587715.
- Gutierrez, D., Anson, O., Munoz, A., and Seron, F. 2005. Perception-based rendering: eyes wide bleached. In *EG Short Paper*, pages 49–52.
- Hanhart, P. and Ebrahimi, T. 2014. Subjective evaluation of two stereoscopic imaging systems exploiting visual attention to improve 3D quality of experience. In *Proc. SPIE vol. 9011*, pages 0D–1–11.
- Haro, G., Bertalmío, M., and Caselles, V. 2006. Visual acuity in day for night. *Int. J. Comput. Vision*, 69(1), 109–117. ISSN 0920-5691.
- Harris, J. M. and Watamaniuk, S. N. 1995. Speed discrimination of motion-in-depth using binocular cues. *Vision Research*, 35(7), 885–896. ISSN 0042-6989.
- Harris, J. M., McKee, S. P., and Watamaniuk, S. N. 1998. Visual search for motion-in-depth: Stereomotion does not ‘pop out’ from disparity noise. *Nature Neuroscience*, 1(2), 165–168.
- Harris, J. M., Nefs, H. T., and Grafton, C. E. 2008. Binocular vision and motion-in-depth. *Spatial Vision*, 21(6), 531–547.
- Hecht, S., Shlaer, S., and Pirenne, M. H. 1942. Energy, quanta, and vision. *J Gen Phys*, 25(6), 819–840.
- Heckmann, T. and Schor, C. M. 1989. Is edge information for stereoacuity spatially channeled? *Vis Res*, 29(5), 593–607.
- Hess, R., Sharpe, L., and Nordby, K. 1990. *Night Vision: Basic, Clinical and Applied Aspects*. Cambridge University Press.
- Heuer, H. 1987. Apparent motion in depth resulting from changing size and changing vergence. *Perception*, 16(3), 337–50.
- Hoffman, D. M., Karasev, V. I., and Banks, M. S. 2011. Temporal presentation protocols in stereoscopic displays: Flicker visibility, perceived motion, and perceived depth. *Journal of the Society for Information Display*, 19(3), 271–297.
- Hoffman, D., Girshick, A., Akeley, K., and Banks, M. 2008. Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *J. Vision*, 8(3), 1–30.
- Howard, I. P. and Rogers, B. J. 2012. *Perceiving in Depth*. I. Porteous, Toronto.
- Hubona, G. S., Wheeler, P. N., Shirah, G. W., and Brandt, M. 1999. The relative contributions of stereo, lighting, and background scenes in promoting 3d depth visualization. *ACM Trans. Comput.-Hum. Interact.*, 6(3), 214–242.
- Itti, L., Koch, C., and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, 20(11), 1254–59.

- Jacobs, D., Gallo, O., A. Cooper, E., Pulli, K., and Levoy, M. 2015. Simulating the visual experience of very bright and very dark scenes. *ACM Trans. Graph.*, 34(3), 25:1–25:15. ISSN 0730-0301. doi: 10.1145/2714573.
- Janesick, J. R. 2001. *Scientific charge-coupled devices*, volume 83. SPIE press.
- Johnson, G. M. and Fairchild, M. D. 2000. Sharpness rules. In *IS&T/SID 8th Color Imaging Conference*, pages 24–30.
- Jonas, J. B., Schneider, U., and Naumann, G. O. 1992. Count and density of human retinal photoreceptors. *Graefes archive for clinical and experimental ophthalmol.*, 230(6), 505–10.
- Jones, G. R., Lee, D., Holliman, N. S., and Ezra, D. 2001. Controlling perceived depth in stereoscopic images. In *SPIE vol. 4297*, pages 42–53.
- Kane, D., Guan, P., and Banks, M. S. 2014. The limits of human stereopsis in space and time. *The Journal of Neuroscience*, 34(4), 1397–1408.
- Kass, M. and Pesare, D. 2011. Coherent noise for non-photorealistic rendering. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 30(4), 30.
- Khan, S. M. and Pattanaik, S. N. 2004. Modeling blue shift in moonlit scenes by rod cone interaction. *J Vis.*, 4(8).
- Kim, C., Hornung, A., Heinzle, S., Matusik, W., and Gross, M. 2011. Multi-perspective stereoscopy from light fields. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 30(6), 190:1–190:10.
- Kim, T., Park, J., Lee, S., and Bovik, A. C. 2014. 3D visual discomfort prediction based on physiological optics of binocular vision and foveation. In *Asia-Pacific Signal and Information Proc. Assoc. (APSIPA)*, pages 1–4.
- Kirk, A. G. and O’Brien, J. F. 2011. Perceptually based tone mapping for low-light conditions. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 30(4), 42:1–10.
- Komogortsev, O. V. and Khan, J. I. 2008. Eye movement prediction by Kalman filter with integrated linear horizontal oculomotor plant mechanical model. In *Proc. Symp. on Eye Tracking Res. and Appl. (ETRA)*, pages 229–236.
- Kopf, J., Cohen, M., Lischinski, D., and Uyttendaele, M. 2007. Joint bilateral upsampling. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 26(3), 96:1–96:6.
- Koppal, S. J., Zitnick, C. L., Cohen, M. F., Kang, S. B., Ressler, B., and Colburn, A. 2011. A viewer-centric editor for 3D movies. *IEEE Comp. Graph. and Appl.*, 31(1), 20–35.
- Krähenbühl, P., Lang, M., Hornung, A., and Gross, M. 2009. A system for retargeting of streaming video. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 28(5), 126:1–126:10.
- Kral, K. 2003. Behavioural–analytical studies of the role of head movements in depth perception in insects, birds and mammals. *Behavioural Processes*, 64(1), 1–12.
- Krishnan, V., Farazian, F., and Stark, L. 1973. An analysis of latencies and prediction in the fusional vergence system. *Am. J. Optometry and Arch. Am. Academy of Optometry*, 50, 933–9.

- Kuang, J., Johnson, G. M., and Fairchild, M. D. 2007. iCAM06: A refined image appearance model for HDR image rendering. *J Vis Comm Image Repr.*, 18(5), 406–414.
- Kulshreshth, A., Schild, J., and LaViola, J. J., Jr. 2012. Evaluating user performance in 3D stereo and motion enabled video games. In *Foundations of Digital Games*, pages 33–40.
- Kurihara, T., Manabe, Y., Aoki, N., and Kobayashi, H. 2008. Digital image improvement by adding noise: An example by a professional photographer. In *Image Quality and System Performance V*, volume 6808 of *SPIE*, pages 1–10.
- Lagae, A., Lefebvre, S., Cook, R., DeRose, T., Drettakis, G., Ebert, D. S., Lewis, J. P., Perlin, K., and Zwicker, M. 2010. State of the art in procedural noise functions. In *EG 2010 - State of the Art Reports*.
- Lambooij, M., IJsselsteijn, W., Fortuin, M., and Heynderickx, I. 2009. Visual discomfort and visual fatigue of stereoscopic displays: A review. *J. Imaging Sci. Technol.*, 53(3), 1. ISSN 10623701.
- Landy, M. S., Maloney, L. T., Johnston, E. B., and Young, M. 1995. Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research*, 35(3), 389–412.
- Lang, M., Hornung, A., Wang, O., Poulakos, S., Smolic, A., and Gross, M. 2010. Non-linear disparity mapping for stereoscopic 3D. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 29(4), 75:1–75:10.
- Larson, G. W., Rushmeier, H., and Piatko, C. 1997. A visibility matching tone reproduction operator for high dynamic range scenes. *IEEE Trans. Vis. and Comp. Graph.*, 3(4), 291–306.
- Legge, G. and Gu, Y. 1989. Stereopsis and contrast. *Vis Res*, 29(8), 989–1004.
- Lillywhite, P. 1981. Multiplicative intrinsic noise and the limits to visual performance. *Vis. Res.*, 21(2), 291–296.
- Lit, A. 1949. The magnitude of the pulfrich stereophenomenon as a function of binocular differences of intensity at various levels of illumination. *The American Journal of Psychology*, 62(2), pp. 159–181. ISSN 00029556.
- Lit, A. 1959. Depth-discrimination thresholds as a function of binocular differences of retinal illuminance at scotopic and photopic levels. *J. Opt. Soc. Am.*, 49(8), 746–752.
- Lit, A. and Hamm, H. D. 1966. Depth-discrimination thresholds for stationary and oscillating targets at various levels of retinal illuminance. *J. Opt. Soc. Am.*, 56(4), 510–514.
- Liu, S. and Hua, H. 2008. Spatialchromatic foveation for gaze contingent displays. In *Proc. Symp. on Eye Tracking Res. and Appl. (ETRA)*, pages 139–142.
- Livingstone, M. 2002. *Vision and art: the biology of seeing*. Harry N. Abrams.
- Livingstone, M. S. and Hubel, D. H. 1994. Stereopsis and positional acuity under dark adaptation. *Vis. Res.*, 34(6), 799–802.

- Loschky, L. C. and Wolverton, G. S. 2007. How late can you update gaze-contingent multiresolutional displays without detection? *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(4), 7:1–7:10.
- Luca, M. D., Domini, F., and Caudek, C. 2007. The relation between disparity and velocity signals of rigidly moving objects constrains depth order perception. *Vision Research*, 47(10), 1335 – 1349. ISSN 0042-6989. doi: <http://dx.doi.org/10.1016/j.visres.2006.10.029>.
- Mantiuk, R., Bazyluk, B., and Tomaszewska, A. 2011a. Gaze-dependent depth-of-field effect rendering in virtual environments. In *Proceedings of the Second International Conference on Serious Games Development and Applications, SGDA'11*, pages 1–12, Berlin, Heidelberg. Springer-Verlag. ISBN 978-3-642-23833-8. doi: 10.1007/978-3-642-23834-5_1.
- Mantiuk, R., Myszkowski, K., and Seidel, H.-P. 2006. Lossy compression of high dynamic range images and video. In *Human Vision and Electronic Imaging XI, IS&T/SPIE Symposium on Electronic Imaging*, page 60570V. IS&T/SPIE.
- Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W. 2011b. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 30(4), 40:1–40:14. ISSN 0730-0301.
- Mariani, A. P., Kolb, H., and Nelson, R. 1984. Dopamine-containing amacrine cells of Rhesus monkey retina parallel rods in spatial distribution. *Brain Res.*, 322(1), 1 – 7.
- Masia, B., Wetzstein, G., Aliaga, C., Raskar, R., and Gutierrez, D. 2013a. Display adaptive 3D content remapping. *Computers & Graphics*, 37(8), 983–996.
- Masia, B., Wetzstein, G., Didyk, P., and Gutierrez, D. 2013b. A survey on computational displays: Pushing the boundaries of optics, computation, and perception. *Computers & Graphics*, 37(8), 1012 – 1038.
- Matsumiya, K. and Ando, H. 2009. World-centered perception of 3D object motion during visually guided self-motion. *J Vis.*, 9(1).
- McConkie, G. W. and Loschky, L. C. 2002. Perception onset time during fixations in free viewing. *Behavior Research Methods, Instruments, & Computers*, 34(4), 481–490.
- McKee, S. P. and Nakayama, K. 1984. The detection of motion in the peripheral visual field. *Vision research*, 24(1), 25–32.
- Mendiburu, B. 2009. *3D movie making: stereoscopic digital cinema from script to screen*. Focal Press.
- Merkle, P., Morvan, Y., Smolic, A., Farin, D., Müller, K., de With, P. H. N., and Wiegand, T. 2009. The effects of multiview depth video compression on multiview rendering. *Signal Processing: Image Communication*, 24(1-2). doi: <http://dx.doi.org/10.1016/j.image.2008.10.010>.
- Murphy, H. and Duchowski, A. T. 2001. Gaze-contingent level of detail rendering. *Eurographics Short Presentations*.

- Muryy, A. A., Welchman, A. E., Blake, A., and Fleming, R. W. 2013. Specular reflections and the estimation of shape from binocular disparity. *Proc. of the National Academy of Sciences*, 110(6), 2413–2418.
- Nawrot, M. and Stroyan, K. 2009. The motion/pursuit law for visual depth perception from motion parallax. *Vision Research*, 49(15), 1969–1978.
- Nawrot, M., Ratzlaff, M., Leonard, Z., and Stroyan, K. 2014. Modeling depth from motion parallax with the motion/pursuit ratio. *Frontiers in psychology*, 5.
- Neinborg, H., Bridge, H., Parker, A., and Cumming, B. 2005. Neuronal computation of disparity in V1 limits temporal resolution for detecting disparity modulation. *J. Neurosci*, 25, 10207–19.
- Ono, M. E., Rivest, J., and Ono, H. 1986. Depth perception as a function of motion parallax and absolute-distance information. *Journal of Experimental Psychology: Human Perception and Performance*, 12(3), 331.
- O’Shea, R. P., Blake, R., and Wolfe, J. M. 1994. Binocular rivalry and fusion under scotopic luminances. *Perception*, 23(7), 771–784.
- Oskam, T., Hornung, A., Bowles, H., Mitchell, K., and Gross, M. 2011. OSCAM - optimized stereoscopic camera control for interactive 3D. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 30, 189:1–189:8.
- Pajak, D., Herzog, R., Mantiuk, R., Didyk, P., Eisemann, E., Myszkowski, K., and Pulli, K. 2014. Perceptual depth compression for stereo applications. *Computer Graphics Forum (Proc. Eurographics)*, 33(2), 195–204.
- Palmer, S. E. 1999. *Vision science: Photons to phenomenology*, volume 1. MIT press Cambridge.
- Patel, S. S., Bedell, H. E., Tsang, D. K., and Ukwade, M. T. 2009. Relationship between threshold and suprathreshold perception of position and stereoscopic depth. *J Opt Soc Am A*, 26(4), 847–861.
- Pattanaik, S. N., Ferwerda, J. A., Fairchild, M. D., and Greenberg, D. P. 1998. A multiscale model of adaptation and spatial vision for realistic image display. In *Proc. SIGGRAPH*, pages 287–98.
- Pattanaik, S. N., Tumblin, J. E., Yee, H., and Greenberg, D. P. 2000. Time-dependent visual adaptation for fast realistic image display. In *Proc. SIGGRAPH*, pages 47–54.
- Peli, E., Hedges, T. R., Tang, J., and Landmann, D. 2001. A binocular stereoscopic display system with coupled convergence and accommodation demands. In *SID Symposium Digest of Technical Papers*, volume 32, pages 1296–1299.
- Portfors-Yeomans, C. and Regan, D. 1996. Cyclopean discrimination thresholds for the direction and speed of motion in depth. *Vision Research*, 36(20), 3265–3279. ISSN 0042-6989. doi: [http://dx.doi.org/10.1016/0042-6989\(96\)00065-X](http://dx.doi.org/10.1016/0042-6989(96)00065-X).
- Portfors-Yeomans, C. and Regan, D. 1997. Just-noticeable difference in the speed of cyclopean motion in depth and the speed of cyclopean motion within a frontoparallel plane. *J Exp. Psych.: Human Perception and Performance*, 23(4), 1074–1086.

- Proffitt, D. and Banton, T. 1999. Perceived depth is enhanced with parallax scanning. *University of Virginia-Cognitive Science Department*.
- Qin, D., Takamatsu, M., and Nakashima, Y. 2006. Disparity limit for binocular fusion in fovea. *Optical Review*, 13(1), 34–38. ISSN 1340-6000. doi: 10.1007/s10043-006-0034-5.
- Ramanarayanan, G., Ferwerda, J., Walter, B., and Bala, K. 2007. Visual equivalence: towards a new standard for image fidelity. *ACM Trans. Graph. (Proc. SIGGRAPH)*.
- Rawlings, S. C. and Shipley, T. 1969. Stereoscopic acuity and horizontal angular distance from fixation. *J. Opt. Soc. Am.*, 59(8), 991–993.
- Regan, D. and Beverley, K. 1979. Binocular and monocular stimuli for motion in depth: Changing-disparity and changing-size feed the same motion-in-depth stage. *Vis. Res.*, 19(12), 1331–1342.
- Reinhard, E., Ward, G., Debevec, P., Pattanaik, S., Heidrich, W., and Myszkowski, K. 2010. *High Dynamic Range Imaging*. Morgan Kaufmann Publishers, 2nd edition.
- Richards, W. Jul 1972. Response functions for sine- and square-wave modulations of disparity. *J. Opt. Soc. Am.*, 62(7), 907–911.
- Richardt, C., Stoll, C., Dodgson, N., Seidel, H.-P., and Theobalt, C. 2012. Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos. *Comp. Graph. Forum*, 31(2).
- Ritschel, T. and Eisemann, E. 2012. A computational model of afterimages. *Comp. Graph. Forum (Proc. EG)*, 31(2), 529–534.
- Riva, C. and Petrig, B. 1980. Blue field entoptic phenomenon and blood velocity in the retinal capillaries. *J. Opt. Soc. Am.*, 70(10), 1234–1238.
- Robinson, A. E. and MacLeod, D. I. A. 2013. Depth and luminance edges attract. *Journal of Vision*, 13(11). doi: 10.1167/13.11.3.
- Rogers, B. and Graham, M. 1982. Similarities between motion parallax and stereopsis in human depth perception. *Vision Research*, 22(2), 261–270.
- Scharstein, D. and Pal, C. June 2007. Learning conditional random fields for stereo. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. doi: 10.1109/CVPR.2007.383191.
- Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., and Westling, P. 2014. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition*, pages 31–42. Springer.
- Semmlow, J. and Wetzell, P. 1979. Dynamic contributions of the components of binocular vergence. *J Opt Soc Am A*, 69, 639–45.
- Semmlow, J., Hung, G., and Ciuffreda, K. 1986. Quantitative assessment of disparity vergence components. *Invest. Ophthalmol. Vis. Sci.*, 27, 558–64.
- Seymour, M. 2011. Case study: How to make a Captain America wimp. *fxguide*.

- Shade, J., Gortler, S., He, L.-w., and Szeliski, R. 1998. Layered depth images. In *Proc. of ACM SIGGRAPH*, pages 231–242.
- Sherstyuk, A., Dey, A., Sandor, C., and State, A. 2012. Dynamic eye convergence for head-mounted displays improves user performance in virtual environments. In *Proc I3D*, pages 23–30.
- Shibata, T., Kim, J., Hoffman, D. M., and Banks, M. S. 2011. The zone of comfort: Predicting visual discomfort with stereo displays. *J. Vision*, 11(8), 11. doi: 10.1167/11.8.11.
- Shinya, M. 1993. Spatial anti-aliasing for animation sequences with spatio-temporal filtering. In *Proc. SIGGRAPH*, pages 289–96.
- Shlaer, S. 1937. The relation between visual acuity and illumination. *J Gen Phys*, 21, 165–188.
- Simmons, D. R. and Kingdom, F. A. A. 1997. On the independence of chromatic and achromatic stereopsis mechanisms. *Vision Research*, 37(10), 1271–1280.
- Simmons, D. R. and Kingdom, F. A. A. 2002. Interactions between chromatic- and luminance-contrast-sensitive stereopsis mechanisms. *Vision Research*, 42(12), 1535–1545.
- Smith, S. W. 1997. *The scientist and engineer's guide to digital signal processing*. California Technical Pub.
- Speranza, F., Tam, W. J., Renaud, R., and Hur, N. 2006. Effect of disparity and motion on visual comfort of stereoscopic images. In *SPIE*, volume 6055, pages 94–103.
- Stephenson, I. and Saunders, A. 2007. Simulating film grain using the noise-power spectrum. In *Theory and Practice of Computer Graphics*, pages 69–72.
- Stroyan, K. 2010. Motion parallax is asymptotic to binocular disparity. *arXiv preprint arXiv:1010.0575*.
- Subr, K., Bradbury, G., and Kautz, J. 2012. Two-frame stereo photography in low-light settings: A preliminary study. In *Proc. CVMP*, pages 84–93. ACM.
- Szeliski, R. 2011. *Computer vision: algorithms and applications*. Springer.
- Teich, M., Prucnal, P. R., Vannucci, G., Breton, M. E., and McGill, W. J. 1982. Multiplication noise in the human visual system at threshold. *J. Opt. Soc. Am.*, 72(4), 419–31.
- Templin, K., Didyk, P., Ritschel, T., Myszkowski, K., and Seidel, H.-P. 2012. Highlight microdisparity for improved gloss depiction. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4), 92.
- Templin, K., Didyk, P., Myszkowski, K., Hefeeda, M. M., Seidel, H.-P., and Matusik, W. 2014a. Modeling and optimizing eye vergence response to stereoscopic cuts. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 33(4).
- Templin, K., Didyk, P., Myszkowski, K., and Seidel, H.-P. 2014b. Perceptually-motivated stereoscopic film grain. *Comp. Graph. Forum (Proc. Pacific Graphics)*, 33 (7).

- Thompson, W. B., Shirley, P., and Ferwerda, J. A. 2002. A spatial post-processing algorithm for images of night scenes. *J. Graph. Tools*, 7(1), 1–12.
- Thornton, W. A. 1999. Spectral sensitivities of the normal human visual system, color-matching functions and their principles, and how and why the two sets should coincide. *Color Research & Application*, 24(2), 139–156. ISSN 1520-6378.
- Tsirlin, I., Allison, R. S., and Wilcox, L. M. 2008. Stereoscopic transparency: Constraints on the perception of multiple surfaces. *J. Vis.*, 8(5).
- Turner, J., Braunstein, M. L., and Andersen, G. J. 1997. Relationship between binocular disparity and motion parallax in surface detection. *Perception & psychophysics*, 59(3), 370–380.
- Tyler, C. W. 1975. Spatial organization of binocular disparity sensitivity. *Vision Research*, 15(5), 583–590.
- Tyler, C. W., Likova, L. T., Atanassov, K., Ramachandra, V., and Goma, S. 2012. 3D discomfort from vertical and torsional disparities in natural images. In *Proc. SPIE*.
- Tyler, C. 1971. Stereoscopic depth movement: Two eyes less sensitive than one. *Science*, 174, 958–961.
- Ullman, S. 1983. Maximizing rigidity: The incremental recovery of 3-d structure from rigid and rubbery motion. *Perception*, 13, 255–74.
- Umino, Y., Solessio, E., and Barlow, R. B. 2008. Speed, spatial, and temporal tuning of rod and cone vision in mouse. *J. Neur.*, 28(1), 189–98.
- v3© Imaging. 2015. www.inv3.com.
- Vangorp, P., Mantiuk, R. K., Bazyluk, B., Myszkowski, K., Mantiuk, R., Watt, S. J., and Seidel, H.-P. 2014. Depth from HDR: Depth induction or increased realism? In *Proc. ACM SAP 2014*, pages 71–78.
- Vinnikov, M. and Allison, R. S. 2014. Gaze-contingent depth of field in realistic scenes: The user experience. In *Proc. Symp. on Eye Tracking Res. and Appl. (ETRA)*, pages 119–126.
- Wald, G. 1945. Human vision and the spectrum. *Science*, 101, 653–58.
- Wallach, H. and O’Connell, D. N. 1953. The kinetic depth effect. *Journal of Experimental Psychology*, 45(4), 205.
- Wanat, R. and Mantiuk, R. 2014. Simulating and compensating changes in appearance between day and night vision. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 33(4).
- Wandell, B. A. 1995. *Foundations of Vision*. Sinauer Associates.
- Wang, Y.-S., Lin, H.-C., Sorkine, O., and Lee, T.-Y. 2010. Motion-based video retargeting with optimized crop-and-warp. *ACM Trans. Graph.*, 29, 90:1–90:9.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4), 600–12.

- Ward, G., Rushmeier, H., and Piatko, C. 1997. A visibility matching tone reproduction operator for high dynamic range scenes. *IEEE Trans. Vis. and Comp. Graph.*, 3(4), 291–306.
- Watson, A. B. and Pelli, D. G. 1983. QUEST: a bayesian adaptive psychometric method. *Perception & Psychophysics*, 33(2), 113–120.
- Watson, A. B. and Yellott, J. I. 2012. A unified formula for light-adapted pupil size. *J Vis*, 12.
- Weinshall, D. 1989. Perception of multiple transparent planes in stereo vision. *Nature*, 341(6244), 737–739.
- Wetzstein, G., Lanman, D., Hirsch, M., and Raskar, R. July 2012. Tensor displays: Compressive light field synthesis using multilayer displays with directional backlighting. *ACM Trans. Graph.*, 31(4), 80:1–80:11. ISSN 0730-0301. doi: 10.1145/2185520.2185576.
- Wheatstone, C. 1838. Contributions to the physiology of vision.—Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Phil Trans. Royal Society of London*, 128, 371–394.
- Wikipedia. 2015a. Parallax scrolling — Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Parallax_scrolling [Online; accessed 2-June-2015].
- Wikipedia. 2015b. Wiggle stereoscopy — Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Wiggle_stereoscopy [Online; accessed 2-June-2015].
- Woods, A., Docherty, T., and Koch, R. 1993. Image distortions in stereoscopic video systems. In *Stereoscopic Displays and Applications*.
- Woods, A. J. and Rourke, T. 2004. Ghosting in anaglyphic stereoscopic images. In *Proc. SPIE 5291*.
- Yan, T., Lau, R., Xu, Y., and Huang, L. 2013. Depth mapping for stereoscopic videos. *International Journal of Computer Vision*, 102, 293–307.
- Yang, X., Zhang, L., Wong, T.-T., and Heng, P.-A. 2012. Binocular tone mapping. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4), 93:1–93:10.
- Yang, Z. and Purves, D. 2003. A statistical explanation of visual space. *Nature Neuroscience*, 6(6), 632–640.
- Yano, S., Ide, S., Mitsuhashi, T., and Thwaites, H. 2002. A study of visual fatigue and visual comfort for 3D HDTV/HDTV images. *Displays*, 23(4), 191 – 201.
- Yau, K., Matthews, G., and Baylor, D. 1979. Thermal activation of the visual transduction mechanism in retinal rods. *Nature*, 279, 806–7.
- Young, M. J., Landy, M. S., and Maloney, L. T. 1993. A perturbation analysis of depth perception from combinations of texture and motion cues. *Vision Research*, 33(18), 2685–2696.
- Zhang, J. and Sclaroff, S. 2013. Saliency detection: a Boolean map approach. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*.

- Zhang, L., Vazquez, C., and Knorr, S. 2011. 3D-TV content creation: Automatic 2D-to-3D video conversion. *IEEE Trans. Broadcasting*, 57(2), 372–83.
- Zilly, F., Kluger, J., and Kauff, P. 2011. Production rules for stereo acquisition. *Proc. IEEE*, 99(4), 590–606.
- Zimmer, H., Bruhn, A., and Weickert, J. 2011. Optic flow in harmony. *International Journal of Computer Vision*, 93(3), 368–388.
- Zwicker, M., Matusik, W., Durand, F., and Pfister, H. 2006. Antialiasing for auto-multiscopic 3D displays. In *Proceedings of the 17th Eurographics Conference on Rendering Techniques*, pages 73–82. Eurographics Association.