Perceptual Display:

Exceeding Display Limitations by Exploiting the Human Visual System

Piotr Didyk
Max-Planck-Institute Informatik
Campus E1.4
66123 Saarbrücken, Germany

A thesis for obtaining the title of Doctor of Engineering
of the Faculties of Natural Science and Technology of Saarland University.

14th June, 2012
Saabrücken, Germany

Supervisors – Betreuende Hochschullehrer
Prof. Dr.-Ing. Karol Myszkowski, MPI Informatik, Saarbrücken, Germany
Prof. Dr. Hans-Peter Seidel, MPI Informatik, Saarbrücken, Germany

Reviewer – Gutachter
Prof. Dr. Elmar Eisemann, Télécom ParisTech (CNRS-LTCI), Paris, France
Prof. Dr.-Ing. Karol Myszkowski, MPI Informatik, Saarbrücken, Germany
Prof. Dr. Hans-Peter Seidel, MPI Informatik, Saarbrücken, Germany

Dean – Dekan
Prof. Dr. Mark Groves, Universität des Saarlandes, Saarbrücken, Germany

Examination – Kolloquium

Date – Datum:
20th August, 2012

Chair – Vorsitzender:
Prof. Dr.-Ing. Philipp Slusallek, Universität des Saarlandes, Saarbrücken, Germany

Examiners – Prüfer:
Prof. Dr. Elmar Eisemann, Télécom ParisTech (CNRS-LTCI), Paris, France
Prof. Dr.-Ing. Karol Myszkowski, MPI Informatik, Saarbrücken, Germany
Prof. Dr. Hans-Peter Seidel, MPI Informatik, Saarbrücken, Germany

Reporter – Protokoll:
Dr. Tobias Ritschel, MPI Informatik, Saarbrücken, Germany

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, 14.06.2012

Abstract

Existing displays have a number of limitations, which make it difficult to realistically reproduce real-world appearance; discrete pixels are used to represent images, which are refreshed only a limited number of times per second, the output luminance range is much smaller than in the real world, and only two dimensions are available to reproduce a three-dimensional scene.

While in some cases technology advanced and higher frame rates, higher resolution, higher luminance, and even disparity-based stereo is possible, these solutions are often costly and, further, it is challenging to produce adequate content.

On the other hand, the human visual system has certain limitations itself, such as the density of photoreceptors, imperfections in the eye optics, or the limited ability to discern high-frequency information. The methods presented in this dissertation show that taking these properties into account can improve the efficiency and perceived quality of displayed imagery. More precisely, those techniques make use of perceptual effects, which are not measurable physically, that will allow us to overcome the physical limitations of display devices in order to enhance apparent image qualities.

Kurzfassung

Aktuelle Anzeigegeräte haben eine Reihe von Einschränkungen, die die wirklichkeitsnahe Abbildung der realen Welt begrenzen. Sie verwenden diskrete Pixel, um ein Bild das mehrmals pro Sekunde aktualisiert wird darzustellen. Weiterhin sind der Kontrast und die Leuchtkraft des dargestellten Bildes viel niedriger als das, was der Mensch in der realen Welt wahrnehmen kann und es sind nur zwei Dimensionen verfügbar, um eine dreidimensionale Szene zu simulieren.

Während in einigen Bereichen, durch technologische Fortschritte, höhere Bildwiederholungsraten, höhere Auflösungen, höhere Leuchtkraft und sogar disparitätsbasiertes Stereo möglich wurden, sind diese Lösungen oft kostspielig und es bleibt schwierig Inhalte zu produzieren, welche die technischen Gegebenheiten am besten ausnutzen.

Andererseits ist auch das visuelle System des Menschen Einschränkungen unterworfen: z.B. der endlichen Dichte der Photorezeptoren, und der Imperfektion des Auges, oder die begrenzte Fähigkeit, hochfrequente Informationen zu erkennen. Die in dieser Dissertation vorgestellten Techniken zeigen, dass, wenn diese Erkenntnisse in der Bildsynthese genutzt werden, sowohl die Effizienz als auch die wahrgenommene Qualität der angezeigten Bilder verbessert werden kann. Die Methoden, welche in dieser Dissertation vorgestellt werden, nutzen unterschiedliche Wahrnehmungseffekte, welche physikalisch nicht messbar sind, aber Einfluss auf die Wahrnehmung haben. Dadurch wird es möglich, physische Einschränkungen von Anzeigegeräten zu umgehen und die Bildqualität substantiell zu verbessern.

Summary

The continuous need for better image quality forces display manufactures to unceasingly improve contrast, brightness, color and spatial as well as temporal resolution. Although the design of these devices already accounts, in many cases, for limitations of the human visual system (HVS), such as luminance adaptation, density of photoreceptors in the retina, imperfections in the eye optics or disability of discriminating high-frequency temporal light fluctuations, they are still not able to reproduce an appearance that we are used to from looking at the real world. The task of realistic image reproduction on display devices is getting even more difficult in the presence of new display technology. For example, the quickly-developing 3D technology, displays stereo images on flat screens. This may lead to a conflict between eye vergence and accommodation which are strongly coupled. Also "frozen in time" discrete frames (e. g., for LCD display) may result in perceived blur due to the fact that eyes track objects moving on the screen in a continuous way while the content is presented to the viewer as a sequence of static images that are displayed for an extended period of time.

In this dissertation, we propose several methods that rather refer to perceptual effects than to physical effects, which means that we can experience but not measure them physically. In particular, we are aiming at the exploitation of perceptual effects to help overcome physical limitations of display devices in order to enhance apparent image qualities. To this end, we present perceptually-motivated temporal upsampling which, by exploiting the temporal integration of the HVS, reduces the so-called hold-type blur for computer generated content in an efficient way. We also show that temporal integration can be successfully used in the context of apparent resolution enhancement. Those two techniques allow for better retargeting of the presented content between devices of different temporal or spatial resolution. Such retargeting has become even more crucial in the context of 3D stereo technology. In this dissertation we show that taking human perception can improve perceived quality and even overcome certain limitations of current 3D display technology.This is achieved by developing perceptual models for disparity along with a number of disparity manipulation techniques. These methods, besides improving stereo content, can be applied at interactive rates which is enabled by our technique for efficient stereo-image creation. This technique can produce regular stereo images based only on one available view. All the here-presented techniques are evaluated via psychophysical experiments, which show the significant advantages when our techniques are used. In the following paragraphs we shortly summarize our methods.

Perceptually-motivated Real-time Temporal Upsampling of 3D Content for High-refresh-rate Displays

High-refresh-rate displays (e. g., 120 Hz) have recently become available on the consumer market and quickly gain on popularity. One of their aims is to reduce the perceived blur created by moving objects that are tracked by the human eye. However, an improvement is only achieved if the video stream is produced at the same high refresh rate (i. e., 120 Hz). Some devices, such as LCD TVs, solve this problem by converting low-refresh-rate content (i. e., 50 Hz PAL) into a higher temporal resolution

(i. e., 200 Hz) based on two-dimensional optical flow. In our approach, we show how rendered three-dimensional images produced by recent graphics hardware can be upsampled more efficiently, which results at the same time in a higher quality. Our algorithm relies on several perceptual findings and preserves the naturalness of the original sequence.

Apparent Display Resolution Enhancement for Moving Images

The limited spatial resolution of current displays makes the depiction of very fine spatial details difficult. We propose a novel method applied to moving images that takes the HVS into account and leads to an improved perception of such details. To this end, we display images rapidly varying over time along a given trajectory on a high-refresh-rate display. Due to the retinal integration time the information is fused and yields apparent super-resolution pixels on a conventional-resolution display. We discuss how to find optimal temporal pixel variations based on linear eye-movement and image content and extend our solution to arbitrary trajectories. This step involves an efficient method to predict and successfully treat potentially visible flickering.

Perceptual Models for Disparity

Binocular disparity is an important cue for the human visual system to recognize spatial layout, both in reality and simulated virtual worlds. In this dissertation we introduce a perceptual model of disparity that is used to predict the human response related to complex stereo images. Such a model has a number of applications. It allows us to define a metric to compare a stereo image to an alternative stereo image and to estimate the magnitude of the perceived disparity change. It can also be used to assess the effect of disparity in order to control the level of undesirable distortions or enhancements (introduced on purpose). Besides the prediction of perceived differences, other applications include compression, and re-targeting. We also present novel applications in form of disparity optimization, hybrid stereo images and backward-compatible stereo. The latter minimizes disparity in order to convey a stereo impression if special equipment is used but it produces images that appear almost ordinary to the naked eye. This is achieved by exploiting the Craik-O'Brien-Cornsweet illusion, which, when used skillfully, can enhance depth perception. We also present a technique, which predicts the HVS response to a disparity signal while accounting for an underlying luminance pattern. This is the first technique that is able to capture this interaction. Accounting for luminance while processing disparity extends the list of possible applications of the disparity model to joint luminance-disparity manipulation or a disparity optimization for multi-view autostereoscopic display.

Adaptive Image-space Stereo View Synthesis

Stereo vision is becoming increasingly popular in feature films, visualization and interactive applications such as computer games. However, computation costs are doubled when rendering an individual image for each eye. Our technique allows us to generate two individual images for the left and right eye using an image-based method which uses only single image, together with a depth map as an input. The resulting method computes a high-quality stereo pair for roughly half the cost when compared

to the traditional method. We achieve this result via an adaptive-grid warping that also involves information from previous frames in order to avoid artifacts.

Zusammenfassung

Der ständige Bedarf an besserer Bildqualität führt dazu, dass Kontrast, Helligkeit, Farbe, sowie räumliche und zeitliche Auflösung von Bildschirmen, andauernd verbessert werden. Obwohl das Design von neuen Geräten bereits vielfach die Einschränkungen des menschlichen visuellen Systems berücksichtigt (wie die Dichte der Photorezeptoren in der Netzhaut, Unvollkommenheiten in der Augenoptik oder Wahrnehmungseinschränkungen von hochfrequenten Informationen in Zeit und Raum), sind Hersteller noch nicht in der Lage, das Aussehen der realen Welt exakt zu reproduzieren. Realistische Bildwiedergabe wird sogar zunehmend schwieriger für einige der neueren Bildschirmtechnologien. Zum Beispiel werden für die 3D-Visualisierung Stereobilder auf Flachbildschirmen präsentiert. Dies kann z.B. zu einem Konflikt zwischen der Augenvergenz und Akkommodation, welche beide stark miteinander gekoppelt sind, führen. Auch Bilder, die "in der Zeit eingefroren" dargestellt werden (z.B. für LCD-Displays) können in Unschärfe resultieren, da dass Auge Objekte auf dem Bildschirm kontinuierlich verfolgt, während in Wirklichkeit eine Sequenz an statischen Bildern präsentiert wird.

In dieser Arbeit schlagen wir mehrere Methoden vor, die auf Wahrnehmungseffekten, welche sichtbar, aber nicht physikalisch messbar sind, beruhen. Insbesondere sind wir an der Ausnutzung dieser Wahrnehmungseffekten interessiert, um die physikalischen Grenzen von Anzeigegeräten zu überschreiten und damit eine Verbesserung der "wahrgenommenen" Bildqualität zu erzielen. Zu diesem Zweck stellen wir ein perzeptuelles temporales Upsampling vor, welches die zeitliche Integration der menschlichen visuellen Wahrnehmung ausnutzt, um die sogenannte "Hold-Typ Unschärfe" von computergenerierten Inhalten zu reduzieren. Wir zeigen auch, dass durch die zeitliche Integration auch die "wahrgenommene" physikalische Bildauflösung "virtuell" erhöht werden kann. Diese Techniken erlauben eine bessere übertragung von Inhalten zwischen verschiedenen Geräten mit unterschiedlicher zeitlicher oder räumlicher Auflösung, was insbesondere für die 3D Stereotechnologie wichtig ist. In dieser Doktorarbeit wird gezeigt, dass, unter Berücksichtigung der menschlichen Wahrnehmung, Verbesserungen der wahrgenommenen Qualität, sowie das überwinden bestimmter Einschränkungen von aktuellen 3D Displaytechnologien möglich wird. Wir erreichen dieses Ergebnis durch die Entwicklung von Wahrnehmungsmodellen und einer Reihe von Techniken zur Manipulation von Stereo-Disparität. Neben der Verbesserung von Stereoinhalten, untersuchen wir auch die effiziente Erzeugung von Stereobildern in Echtzeit. Diese Technik kann Stereobilder aus nur einem einzigen zur Verfügung stehenden Bild erzeugen. Alle hier vorgestellten Verfahren werden mittels psychophysischer Experimente ausgewertet, und wir illustrieren darin die deutlichen Vorteile unserer Techniken. In den folgenden Abschnitten werden wir diese kurz zusammenfassen.

Wahrnehmungsbasiertes Echtzeit-Upsampling von 3D Inhalten für Hochfrequenzbildschirme

Hochfrequenzbildschirme (z.B. 120 Hz) sind seit kurzem weit verbreitet und gewinnen zunehmend an Popularität. Eines der Ziele dieser Arbeit ist es, die wahrgenommene Unschärfe von bewegten Objekten zu reduzieren. Allerdings kann eine Verbesserung

nur erreicht werden, wenn der Video-Stream mit einer hohen Aktualisierungsrate produziert wird (120 Hz). Einige Geräte, wie zum Beispiel LCD-TVs, lösen dieses Problem durch die Umwandlung von wenigen Eingangsbildern (z.B. 50 Hz PAL) in eine höhere Bildfrequenzrate (z.B. 200 Hz) durch die Berechnung von zweidimensionalem optischen Fluss. In unserem Ansatz zeigen wir, wie synthetische dreidimensionale Bilder, unter Benutzung moderner Hardware, effizienter in eine qualitativ hochwertige Sequenz umgewandelt werden können. Unser Algorithmus basiert auf verschiedenen Wahrnehmungsbefunden und bewahrt die Natürlichkeit der ursprünglichen Sequenz.

Scheinbare Auflösungserhöhung für bewegte Bilder

Die begrenzte Bildauflösung aktueller Bildschirme macht die Darstellung sehr feiner Details schwierig. Wir schlagen eine neue Methode vor, die für bewegte Bilder, unter Berücksichtigung des menschlichen visuellen Systems, zu einer verbesserten Wahrnehmung solcher Details führt. Zu diesem Zweck zeigen wir Bilder mit zeitlich schnell variierenden Inhalten entlang einer vorgegebenen Strecke auf einem Bildschirm mit hoher Bildwiederholrate. Aufgrund der zeitlichen Integration des so erzeugten Signals auf der Netzhaut, werden die Informationen verbunden und ein hochauflösendes Ergebnis auf einem niedrig aufgelösten Bildschirm erreicht. Wir zeigen, wie man die optimale zeitliche Variation von Pixeln, abgestimmt auf eine lineare Augenbewegung, bestimmt und erweitern danach unsere Lösung, auf beliebige Augenbewegungen. Dieser Schritt beinhaltet eine effiziente Methode zum Vorhersagen und Vermeidung von potenziell sichtbarem Flickern.

Perzeptuelle Modelle für Stereo-Disparität

Binokulare Disparität ist ein wichtiger Reiz, den das visuelle System des Menschen verarbeitet, um die räumliche Anordnung in der Realität, wie auch in der simulierten Welt, zu verstehen. In dieser Dissertation stellen wir Wahrnehmungsmodelle vor, die verwendet werden, um die menschliche Antwort auf komplexe Stereobilder vorherzusagen. Solche Modelle verfügen über eine Reihe von Anwendungen. So ist z.B. möglich, eine Metrik zu definieren, um ein Stereobild mit einem alternativem Stereobild zu vergleichen und die Größe der wahrgenommenen Unterschiede abzuschätzen. Die Metrik kann auch verwendet werden, um die Wirkung des Unterschieds zu bewerten und unerwünschte Veränderungen (eventuell zu einem gewissen Grad absichtlich hinzugefügt) zu messen und zu steuern. Andere Anwendungen sind ebenfalls möglich, z. B. Komprimierung und kontrollierte Übertragungen zwischen Geräten. Außerdem präsentieren wir neue Anwendungen in Form von Disparitätsoptimierung, hybriden Stereobildern und rückwärts-kompatiblem Stereo. Letzteres minimiert die Disparität, so dass Bilder fast gewöhnlich erscheinen, wenn man sie mit bloßem Auge betrachtet, aber einen Stereoeindruck vermitteln, falls spezielle Ausrüstung verwendet wird. Dies wird durch die Ausnutzung der Craik-O'Brien-Cornsweet Illusion erreicht, die, geschickt eingesetzt, die Tiefenwahrnehmung erhöhen kann. Außerdem stellen wir eine Technik vor, die die menschliche Reaktion auf ein Disparitätssignal unter Berücksichtigung des zugrundeliegenden Farbmusters vorhersagt. Dies ist die erste Technik, welche in der Lage ist, solche Wechselwirkungen zu erfassen. Die Berücksichtigung der Leuchtkraft liefert eine Vielfalt an neuen Anwendungen für unser Disparitätsmodell, wie z.B. Bildmanipulation, oder die Optimierung für autostereoskopische Bildschirme.

Adaptive Bild-Raum-Stereo-Ansicht Synthese

Stereo Vision wird immer beliebter in Spielfilmen, bei der Visualisierung von Daten und für interaktive Anwendungen wie Computerspiele. Allerdings verdoppelt sich der Aufwand der Bildsynthese, da jeweils ein Bild pro Auge erzeugt werden muss. Unsere Technik ermöglicht das Erstellen von zwei einzelnen Bildern (für das linke und das rechte Auge) durch ein bildbasiertes Verfahren, welches nur ein einzelnes Eingabebild (inklusive Tiefenbild) verwendet. Die resultierende Methode reduziert somit die Kosten für ein hochwertiges Stereo-Bildpaar, im Vergleich mit traditionellen Methoden, auf in etwa die Hälfte. Wir erreichen dieses Ergebnis durch ein adaptives Gitter, welches die Inhalte verformt. Desweiteren nutzen wir auch Informationen aus vorhergehenden Bildern, um Artefakte zu vermeiden.

Acknowledgments

Completing my PhD degree was probably the most challenging thing that I have ever achieved. This, however, would never happen if not a number of people that I had a chance to meet and work with during the last 4 years.

First, I would like to thank my supervisor Karol Myszkowski for introducing me to the topic of perception in computer graphics. He has been a continuous inspiration and always has enough time and patience for his students. I am proud of being part of his group and will remember him as a person with a great heart and deep knowledge.

I would also like to thank Hans-Peter Seidel for giving me the opportunity to work at Max-Planck-Institute and creating an environment where people can do what they are passionate about.

This dissertation would not be possible without Elmar Eisemann who was my second mentor. He shaped the way I approach research problems and taught me how to turn good ideas into interesting ideas and how to present them to others. His innovative ideas made a significant contribution to my dissertation and will further influence my research.

My research was made more efficient and much more extensive thanks to Tobias Ritschel. I appreciate particularly his constant readiness to discuss and try new things, his creativity and how he guided me through GPU programming. The work described here would not be possible without his programming framework which he made available for me.

I am grateful for having the opportunity to work with Wojciech Matusik who invited me to MIT. It was a very enriching experience to work in such a great group and be exposed to new problems of computer graphics.

Also, I need to thank Rafał Mantiuk, who was the first person to show me computer graphics from the research point of view. He helped me later with making my first steps as a researcher in this field.

I highly appreciate the time that Krzysztof Templin and Oliver Klehm spent on proof-reading parts of this dissertation. Additionally, I would like to thank Krzysztof for a number of inspiring discussions and all projects on which we worked together.

My dissertation would not be possible without the whole computer graphics group at the institute, which inspired me during my stay there and provided a perfect environment for pursuing my research. I will not forget the help of Sabine Budde, Ellen Fries and all the people working in administration department who always took care of things that I am simply not as good at as they are. My studies also required patience of many people who took parts in experiments and I am thankful for their time.

Finally, I shall not forget the most important and unremitting support - my parents Elżbieta and Krzysztof, my sister Ania, and whole family who stood by me and encouraged me during my whole life. Above all, I would like to thank my wife, Eliška, who supports me in all what I do and is always with me. Without her unwavering love my life would be much less colorful than it is now.

Contents

1

Introduction

This dissertation is inspired by the existing gap between capabilities of current display devices and the richness of the real world. Although continuous development is being made by display manufacturers, we are still far from being able to faithfully reproduce the real world on a screen. This is mostly due to the high cost and physical limitations of the display designs as well as different viewing condition. We propose several techniques that, instead of designing new and better hardware, rely on properties of the human visual system. Considering perception enabled us to improve the quality of the perceived content. Surprisingly, our methods allow to improve the quality of images even beyond physical capabilities of the display devices.

This chapter provides a detailed motivation as well as a description of our novel contributions. We also describe the organization of this dissertation.

1.1 Motivation

Existing display devices introduce a number of physical constraints, which make it difficult to realistically reproduce real-world appearance. For example, a direct reproduction of the luminance range of a moonless night to the intensity of the sun is technically out of reach. Similarly, the continuous nature of spatial and temporal information does not directly match the discrete notions of pixels and frames per second.

The human visual system (HVS) has its own limitations, which to a certain extent reduce the requirements imposed on display devices. For example, through a luminance-adaptation process (that can be extended in time) our eyes can operate both in dark-night and sunny-day conditions, however, simultaneously only 4–5 log-10 units of luminance dynamic range can be perceived at once. Similarly, the limited density of photoreceptors in the retina (in the foveal region the size of cones amounts to 28 arcsec) as well as imperfections in the eye optics limit the spatial resolution of details that can be perceived. In the temporal domain the critical flickering frequency (CFF) limits the ability to discern temporal signals over 60 Hz.

All such HVS-imposed limitations are taken into account, when designing display devices, but still a significant deficit of reproducible contrast, brightness, and spatial pixel resolution can be observed, which fall short with respect to the HVS capabilities. Moreover, unfortunate interactions between technological and biological aspects lead to new problems, which are non-existing in real-world observation conditions. For example, the conflict between the eye accommodation adjusted to the display screen

and the eye-ball vergence driven by depth (disparity) reproduced on 3D stereo displays. This mismatch imposes limitations on the depth range that can be comfortably observed. Also, "frozen in time" discrete frames (e.g., for LCD displays) result in perceptual issues. While the entire sequence might appear smoothly animated, each frame is actually static for an extended period of time. When the eye tracks dynamic objects (to keep their steady projection in the fovea), the static image is traversed smoothly and values crossed by the eye start to integrate on the retina, which results in a perceived hold-type blur. Note that such blur does not exist in the physical space (i.e., in displayed images) but it is a purely perceptual effect. Nonetheless, hold-type blur can degrade the impression of perceived image quality in a similar way as physical blur introduced to images.

Even though hardware is constantly evolving such limitations still persist. In order to surmount the physical limitations of display devices, modern algorithms started to exploit characteristics of the human visual system (HVS) such as apparent image contrast [Purves, Shimpi and Lotto 1999] based on the Cornsweet Illusion or apparent brightness [Zavagno and Caputo 2001] due to glare. Inspired by those methods, we rather refer in our work to the *perceptual effects*, which we can only experience but not measure physically, than to *physical effects*. In particular, we exploit perceptual effects to help overcome physical limitations of display devices in order to enhance apparent image qualities.

In this dissertation, we want to argue that quality consideration need to take human perception into account, in order to take full advantages of new display designs.

1.2 Novel contributions

This dissertation is inspired by the fact that by exploiting human visual perception, perceived quality of content shown on a screen can be significantly improved. We do not achieve our goal by improving existing hardware or designing a new one, which is a common approach used by the industry. Instead, we rely on skillfully designed software techniques that are based on properties of the human visual system. Therefore, techniques presented in this dissertation arise from well-studied findings in the field of human perception.

By analogy to *computational photography*, all techniques that involve additional processing to extend or enhance the capabilities of display devices are called *computational display*. Our methods achieve such goals by taking the advantage of certain HVS properties, therefore, we refer to them as *perceptual display* which forms a subgroup of computational display techniques.

The ideas described in this dissertation have already been published in international journals and presented at various conferences. Overviews of our techniques have also been included in a book chapter [Didyk et al., 2012a] as well as SIGGRAPH Asia and Eurographics courses [Banterle et al. 2011; Banterle et al. 2012]. Here, we present an extended description of our techniques under the common concept of improving image quality by exploiting properties of the human visual system. Our main contributions can be summarized as follows:

- **Perceptually-motivated temporal upsampling**

  We propose a temporal upsampling scheme, tailored towards the requirements of high-refresh-rate hold-type displays and the capabilities for modern graphics hardware at the same time. We exploit the information on depth, occlusion and three-dimensional structure, made available via the GPU as well as perceptual findings to outperform pure image-based upsampling. We diversify subsequent frames in terms of spatial-frequency content to make use of image-fusion characteristics in the HVS. The required steps are simple, do not introduce any lag in the video stream and can extrapolate one or even multiple frames. We show in a psychophysical study that the appearance of the final animation produced by our technique matches 120 Hz sequences closely and outperforms sequences with lower framerate. Similar findings are obtained when considering task performance. The advantages of our technique make it a cheap and more-suited alternative to producing 120 Hz content directly. [Didyk et al., 2010b]

- **Apparent display resolution enhancement**

  Here, we turn the temporal integration of the HVS, which normally leads to perceived blur in standard displays, to our advantage and propose a novel technique, which shows that by taking the temporal eye integration into account, we can optimize the presented images, such that the perceived resolution is enhanced. This allows us to create the impression of looking at images whose resolution is higher than the physical resolution of a screen. The basic idea of this method is to show slightly different information in consecutive frames and to let the HVS fuse it. To this end, we propose an apparent-resolution model for moving images and show a method for optimal subimages derivation. A temporally-varying signal can in some cases produce visible flickering. To avoid such artifacts we also developed an efficient technique that reduces this effect but still improves resolution. We validate our resolution-enhancement technique in a perceptual study, which shows that significant improvements can be achieved, both, for computer-generated images and photographs. [Didyk et al., 2010a], [Didyk et al., 2011a]

- **A perceptual model for stereo 3D disparity**

  We introduce a first disparity model for 3D stereo content that can quantify the perceived depth. In order to build such a model, we first conduct an experiment where we measure the performance of the HVS in discriminating depth differences and later propose a computational framework that allows processing complex images. We show that our model has a number of applications such as perceived disparity metric, depth signal compression or perceptually-based retargeting of stereo content. Based on our model, we also present a novel backward-compatible stereo technique, which minimizes disparity in order to convey a stereo impression when adequate equipment is used but produces images that appear almost ordinary to the naked eye. This is achieved by using a perceptual effect called *Cornsweet Illusion*, which is known from luminance perception but can also be used for depth. [Didyk et al., 2011b], [Didyk et al., 2012b], [Didyk et al., 2012c]

- **A luminance-contrast-aware stereo 3D disparity model**

  Disparity perception is also affected by the underlying luminance pattern. In many cases, when the luminance pattern does not exhibit certain properties, even

a strong disparity signal does not create a good depth impression. In order to capture this behavior, we perform measurements of perceived disparity changes for stimuli with different luminance and disparity patterns and propose a disparity-perception model accounting for RGB image content, which is the first to allow for a joint luminance and disparity handling. This model, besides improving previous applications, enables new ones, such as joint luminance-disparity manipulation or optimization of content for multi-view auto-stereoscopic display. Our results as well as the importance of taking luminance pattern into account while processing a disparity signal are validated in a user study. [Didyk et al., 2012d]

- **Adaptive image-space stereo view synthesis**

  All techniques for 3D stereo content manipulations, which are proposed in our work, can run at interactive rates. This requires a challenging step which is an efficient resynthesis of stereo image pairs after disparity manipulation. Here, we present a method that allows the transformation of a stream of monocular images with depth information into a stream of stereo image pairs, by exploiting modern GPUs and human perception. The reconstruction of a stereo image pair requires only a couple of milliseconds. We achieved this result by extending the grid-based approach of our temporal upsampling techniques coupled with an interleaving technique to render the left and right view depending on the camera path. [Didyk et al., 2010c]

1.3 Dissertation organization

This dissertation is structured as follows. In the following, Chapter 2 describes the perceptual background, where we present properties of the HVS that are later exploit in our techniques. Next, in Chapter 3, we give an overview of related previous work. In Chapter 4, we present a solution for reducing blur in LCD displays. The here-presented technique shows how high-quality frames can be interleaved with low-quality frames, and lead to an overall improvement of the appearance. Afterwards, in Chapter 5, we demonstrate how the high quality of all frames can improve apparent spatial resolution. In Chapter 6, we show the role of perception in the context of stereovision and accommodation/vergence-conflict reduction by presenting two perceptual models for disparity. This chapter is followed by Chapter 7, where we present a number of perceptually-motivated disparity manipulations that take advantage of the disparity models. In Chapter 8, we describe our image-based technique for stereo view synthesis which enables all our disparity manipulation techniques to perform at interactive rates. Finally, we conclude and give indications for future work in Chapter 9.

2

Background

All techniques presented in this dissertation exploit certain properties of the human visual system. This chapter provides an overview of the properties as well as limitations of the HVS in the context of current display technologies. More precisely, we are interested in quality limits coming from both, the HVS as well as the display technology side. Such analysis provides a better overview and understanding of current challenges as well as identifies room for possible improvements. We start in Section 2.1 with discussion of spatial resolution limits imposed by HVS as well as display technology. Next, in Section 2.2, we give an overview of the HVS properties related to temporal resolution which are also discussed in the context of current displays. The last part (Section 2.3) provides a basis for depth perception, including the limits of HVS for perceiving depth as well as the viewing comfort aspect in 3D displays.

2.1 Spatial Resolution Limits

Human capabilities of perceiving high spatial frequencies are crucial in everyday live. Besides reading text or distinguishing small details, they also enable discrimination and identification of objects that are far away. This is important from a survival point of view and has been developed through the human evolution process. The daily experience of highly detailed real word encourage engineers to design not only devices which are capable of capturing this richness but also devices that can faithfully reproduce the unlimited details of the real world.

2.1.1 Physical Limitations

The main limitations of spatial acuity of the HVS come from the way the light coming to our eyes is processed. It first goes through the cornea and anterior chamber to later reeve the pupil. Next, the light passes through the lens and eye's interior to be at the end projected on the retina (Figure 2.1). At this moment the visual acuity is limited by blur caused by the diffraction in the pupil in bright viewing conditions and imperfections of optics in the lens and cornea at low light intensities, when the diameter pupil is much larger [Wandell 1995, p. 37]. Next, located on the retina photoreceptors are stimulated by the projected light, which after conversion to a signal, is transmitted further via nerve fibers for later processing in the brain. Although the visual acuity is here limited by the density of photoreceptors, it turns out that the eye's optical low-pass

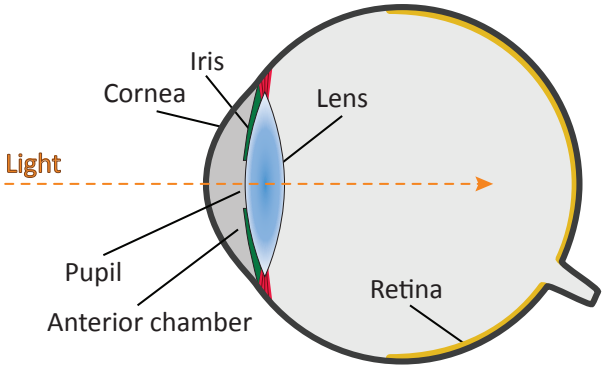


Figure 2.1: Before light that comes to the eye is turned into a signal transported to brain, it goes through different eye components which affect the signal quality.

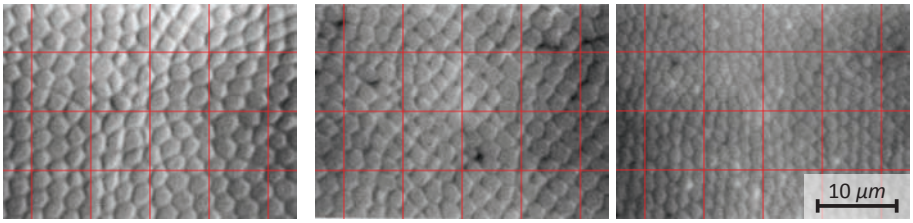


Figure 2.2: Optical sections of foveal cones mosaic for three subjects [Curcio et al. 1990]. The red grid illustrates an estimated pixel grid of 22-inches full-HD display viewed from 50 cm distance projected on the retina.

filtering perfectly matches the foveal photoreceptor density. This way spatial aliasing is avoided.

The limit of the spacial acuity is determined by the smallest anatomic spacing between cones in the fovea. For the average observer it is estimated to be 28" (arc seconds) [Curcio et al., 1990] which, according to the Nyquist's theorem, enables an observer to distinguish 1D sine gratings of roughly 60 $^{\text{cycles}}/_{\text{deg}}$ resolution. However, as shown in Figure 2.2, the cone spacing varies across subjects. For some of them spatial acuity, which is estimated based on cones density, exceeds even 80 $^{\text{cycles}}/_{\text{deg}}$.

2.1.2 Hyperacuity

The story, however, does not end here. Interestingly, the HVS is still able to interpolate a feature position with an accuracy higher than 20 % of the distance between cones in the fovea. Current studies showed so far that this works when the position of one image element is located relative to another, e. g., slightly shifted lines in the vernier acuity task [Wandell 1995, p. 239]. This suggests that it is more a *localization* than a *resolution* task (Figure 2.3). However, it shows that physical limitations of the eye may be compensated to a certain extend by other mechanisms which enhance perceived resolution.

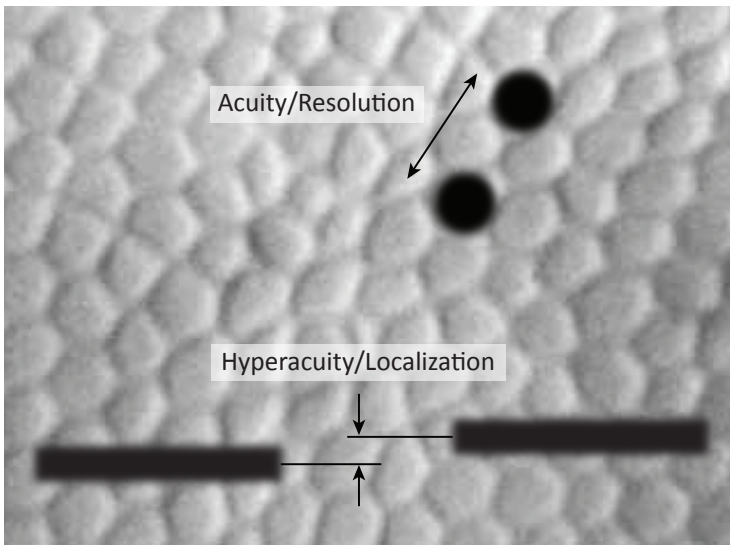


Figure 2.3: Resolution: Two dots (top) projected on the retina can be resolved only if the separation between them leaves space for at least one more active photoreceptor. Otherwise, the patter will be indistinguishable from one elongated dot. Hyperacuity: Two objects (bottom) can be relatively localized using information from many photoreceptors, which enables estimating center of each object with accuracy below cones spacing.

2.1.3 Display Limitations

When discussing HVS capabilities in resolving spatial details it is interesting to confront them with the quality reproducible using current display devices [Deering 2005].

Assuming a pixel size of a typical full-HD desktop display, such as a 120 Hz Samsung SyncMaster 2233 and 50 cm distance from an observer to the screen, the pixel area amounts to roughly 1.5' (arc minutes of visual angle). Comparing it to the estimated density of photorespetors as 28", this means that for average observer 1 pixel covers roughly 9 cones (Figure 2.2). Similar results can be obtained when we compute the density of pixels projected on the center of human retina. According to Wandell [1995], it is around 17,000 px/mm^2. At the same time it has been shown [Curcio et al. 1990], that the center of the retina, which covers 1 deg of viewing angle, contains from 80,000 to 200,000 $cones/mm^2$. Note that those estimates are valid only for the central fovea region. The cone density drops quickly with the eccentricity [Curcio et al., 1990] (Figure 2.4), however, we need to remember that while building overall quality impression this region is most crucial.

Those calculations account only for the standard situation of viewing a computer monitor. However, in many situations an observer might actually be closer to the screen, as this is the case for hand-held devices, big screens or tiled displays [VisBox, Inc.]. In such situations the number of photoreceptors corresponding to one pixel can be significantly higher. This observation resulted in a current trend in display design which aims producing displays with resolution matching the resolution of photoreceptors on the retina. Steps in this direction have been made by Apple which

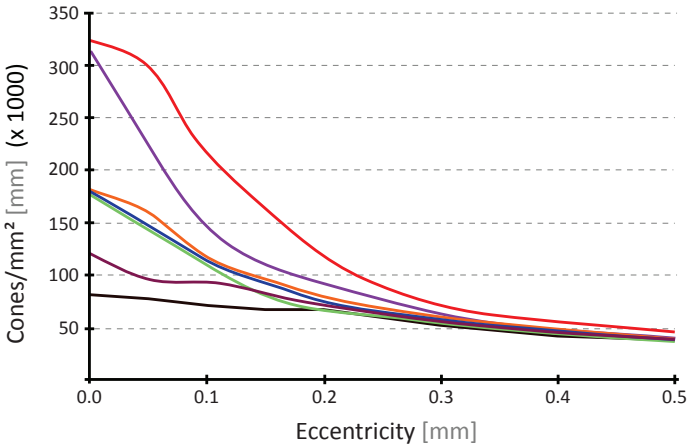


Figure 2.4: Cones density measured for seven subjects as a function of the distance from the center of fovea (Redrawn from [Curcio et al. 1990]).

designed *"Retina"* display or Sharp who has recently presented 85-inches display which offers 8K resolution. However, such screens, especially those whose size allows using them for home cinema applications, are still extremely expensive.

2.2 Temporal Perception

Although described so far aspects are important while analyzing limits of the human visual system, they assume perfect stabilization of an image on the retina as well as constancy of the image presented to the observer. Those conditions, however, are almost never fulfilled in reality as objects in the real world are usually in motion. Also the observer, in order to gain as much information as possible, usually does not look at one point for extended period of time but rather changes the point of interest constantly via fast saccades or stabilize the image of moving objects via tracking. Crucial in this context is also the fact that the way images are registered by our vision system is similar to a time–averaging sensor. This together with omnipresent motion has a huge influence on perceived images in reality but even bigger in the case of display devices which show images in a discrete rather than continuous way.

2.2.1 Temporal Integration and Flickering

Similarly to limitations in spatial resolution, the human visual system is limited in perceiving high frequencies of temporal light fluctuations. This is due to the fact that the response of photoreceptors on our retina is not instantaneous [van Hateren, 2005]. Also higher-level vision processing further lowers the sensitivity of the human visual system to time-varying patterns. To the point where the HVS stops discriminating temporal fluctuations it starts averaging upcoming signal over time.

One of the basic findings from *temporal integration* is Bloch's law [Gorea and Tyler 1986]. It states that the detectability of stimuli with similar spatial characteristics

depends solely on their energy, i. e., the product of luminance and exposure time. In practice it means that the perceived luminance of the signal presented over a given period of time is the same as if the duration time of this signal was halved but the intensity was doubled. It is often assumed that temporal integration of information by the HVS follows this law. However, it is valid only up to some critical duration (around 40 ± 10 ms depending on spatial frequency [Gorea and Tyler, 1986]). For longer durations only the intensity of the signals influences the perceived brightness. Therefore, modeling of temporal integration using Bloch's law is limited to only high frequency temporal fluctuations of signal. In fact, temporal averaging is much more complicated phenomenon and for full understanding, other parameters than time duration and energy of signal need to be taken into account.

From a practical point of view, it is however interesting to know when the HVS sees temporarily varying signal and when the signal is interpreted as constant. Lack of the perceived fluctuations is either due to high temporal frequency light modulation, where Bloch's law holds, or small amplitude of fluctuations that cannot be detected. Signal that appears as constant is defined by *critical flicker frequency* (CFF) [Kalloniatis and Luu, 2009], over which any temporal modulations are imperceptible. In such a case presented intensities are fused and a steady appearance is reached. Flickering perception is complex and the CFF depends on many factors such as the adaptation luminance, spatial extent of flickering pattern (Figure 2.5), and retinal region (the fovea or periphery) at which this pattern is projected. The CFF rises roughly linearly with the logarithm of time-averaged background intensity (the Ferry-Porter law). The specific CFF values for different adaptation luminance have been measured as the *temporal contrast sensitivity function* [de Lange 1958] for stimuli of the spatial extent of 2° (angular degrees). One important observation is that the CFF is significantly reduced for smaller stimuli [McKee and Taylor 1984; Mäkelä, Rovamo and Whitaker 1994] and that the CFF is the highest in the fovea, except for bright adaptation conditions and large stimuli, when flickering is better perceived at the periphery.

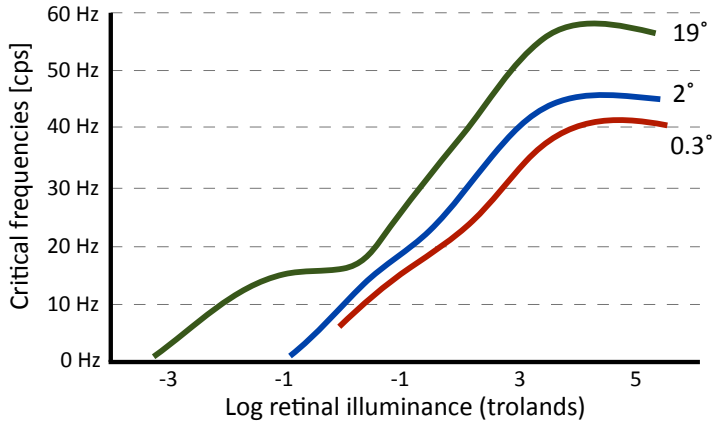


Figure 2.5: Critical flickering frequencies as a function of size and retinal illuminance of test stimulus [Graham 1965].

2.2.2 Perception of Moving Images

The temporal integration nature of the HVS system has a huge influence on perceived image quality once motion is present (e. g., moving objects or eye motion). Mismatch between motion of objects in the scene and human eyes can lower the perceived image quality by e. g., introducing *motion blur*. Surprisingly, in some cases, motion can also improve perceived quality. Therefore, understanding how the HVS creates images of moving objects is crucial for current display design where interaction between continuous eye motion and spatially as well as the temporally discrete nature of screens leads to phenomena that are not observed in the real world.

The perception of motion, where information on objects moving in a 3D world is inferred from 2D retinal images is a complex process [Wandell 1995, Chapter 10]. One of the most important mechanisms that allows humans a good perception of moving objects is *smooth pursuit eye motion* (SPEM). This mechanism compensates for the fact that the visual channel in the HVS specialized in precise object identification is tuned to high spatial frequencies and low temporal frequencies (e. g., food selection), however, it has a poor temporal response [Burr 1981]. Thus, it is crucial for the HVS to stabilize moving objects on the retina in order to be able to recognize them and see details. SPEM is essentially an ability of tracking objects in motion within large range of velocities, so that its projection onto retina is centered in the fovea featuring the highest density of photoreceptors. Through this, images are stabilized on the retina and can be perceived as sharp, which would be impossible without this mechanism. As confirmed in an eye tracking experiment [Laird et al. 2006] such a stabilization is almost perfect for steady linear motion with velocities in the range of 0.625–2.5 $^{deg}/_{s}$. The performance stays very good up to 7 $^{deg}/_{s}$.

SPEM can be also efficiently performed for images involving more complex motion as we experience in real live or during TV or movie watching. It turns out that the initialization of SPEM typically requires 100-120 ms as measured for a completely random direction and velocity of the moving target [Krauzlis and Lisberger 1994]. Anticipatory effects and cognitive strategies as well as the presence of target on the screen before its motion starts can reduce the initialization phase by 30 ms. For comparison, the saccade, which is performed while scanning a visual scene, may require up to 200 ms to initialize and may last 20-200 ms [Krauzlis and Lisberger 1994] when the vision is suppressed without any visible effect on the continuity of seeing. All these observations suggest that switching the eye pursuit between different targets can be effortlessly and efficiently done. While the eye undergoes additional fixational eye movements, such as tremors, drifts, and microsaccades, these are similar to static fixation, and it is believed that the HVS suppresses their influence on perception [Martinez-Conde, Macknik and Hubel 2004].

Although eye tracking is very efficient in many scenarios, it also has its limitations [Daly, 1998]. For low angular velocities below 0.15 $^{deg}/_{s}$ the drift eye movement interferes with the smooth pursuit eye motion. Similarly, for velocities higher than 80 $^{deg}/_{s}$ tracking becomes impossible.

In the next two parts of this section we show how motion can affect perceived image quality. First, we discuss perception of blur, which is caused by the mismatch between SPEM and motion present in the scene. Next, we present cases, where motion can improve quality of perceived images.

2.2.3 Perception of Blur

Image sharpness is an important factor, which decides upon perceived image quality [Janssen, 2001]. Perceptual studies clearly indicate [Calabria and Fairchild 2003; Lin, Gai and Kassim 2006] that people prefer images with increased contrast at edges. This is often achieved by applying image sharpening techniques such as unsharp masking. Not surprisingly, blur in the image is considered an artifact and is particularly annoying when present in the regions of interest that attract visual attention such as moving objects. Blur perception is a complex phenomenon, which is affected by characteristics of the HVS such as temporal integration in the retina, eye motion, and visual illusions.

In the reality the most common kind of blur that our eyes see is *motion blur*. It is created when retinal images of objects move relatively to the retina, which may be caused by the actual object motion, eye motion, or both. Putting it differently, it happens always when SPEM does not work, i. e., velocity of objects is too high to enable good tracking or there are multiple objects in the scene with different velocities and only one of them can be tracked. For all such objects the retinal images acquired in such conditions is blurred, since photoreceptors in the retina integrate signal over time by an analogy to the finite exposure time in cameras.

Blur perception becomes even more complex when we start considering a content shown on a screen, where, due to discrete nature of display devices, moving objects are not perfectly reproduced in terms of motion smoothness. This means that, in contrast to reality, signal transitions at retinal photorespetors do not follow real-world observation conditions. This has significant influence on perceived quality on today's predominant *hold-type* LCD displays. They exhibit two prominent forms of blur: *response time* blur and *hold-type* blur [Pan, Feng and Daly 2005]. Both are not present in *impulse-type* CRT displays, for which other drawbacks exist, such as flickering, lower brightness, and reduced contrast [Klompenhouwer and Velthoven 2004].

Response time blur results from the inability of the LCD display to switch between intensity levels instantaneously, but its contribution to the overall blur is relatively low. Pan et al. [2005] report that only 30 % of blur is a consequence of the response time, even for slow, 16 ms displays. For modern displays, the response time of 2–4 ms becomes negligible and *overdrive* algorithms can even lead to further reductions [Feng 2006].

Hold-type blur is a purely perceptual effect arising from an interaction between the HVS and hold-type displays [Pan, Feng and Daly 2005]. The blur is not physically present in the image and cannot be measured with e. g., a high-speed camera. Hold-type blur can be seen as the inverse of motion blur: In motion blur, the eye is fixed and an object moves, leading to blur, while for the hold-type effect, the image is held constant while the eye moves. Figure 2.6 explains the mechanisms causing this kind of blur.

Hold-type blur can be modeled as a convolution with a box filter oriented in the object motion direction: As the eye moves, the retinal projection moves and, therefore, is spread across a constant box-shaped profile [Klompenhouwer and Velthoven 2004]. The box-filter support size, and thus the strength of blur, depends on the moving pattern, velocity, and the frame-hold duration. The faster the motion, the longer the distance in terms of pixels and consequently, this leads to more blur. With increasing refresh rates, the hold-type blur is reduced because the hold time itself is getting shorter. Figure 2.7 illustrates the amount of perceived blur, introduced by hold-type displays

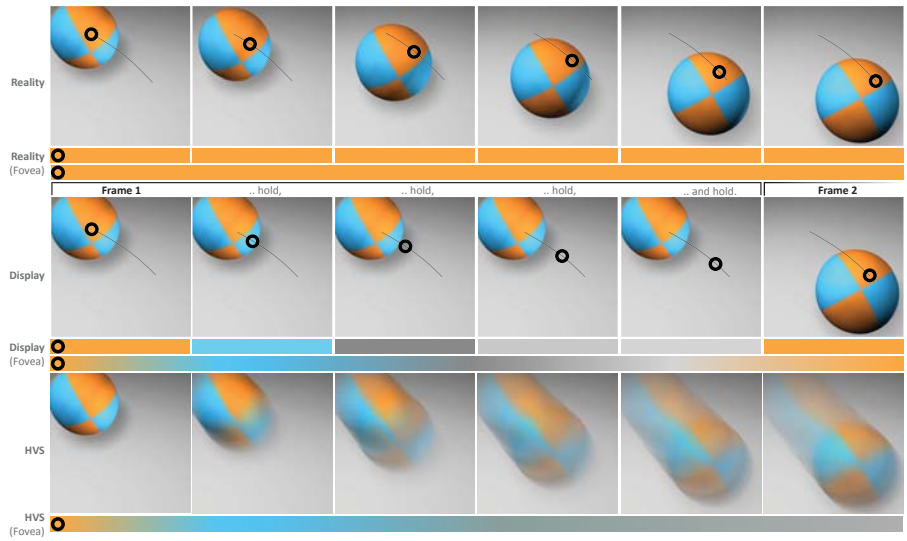


Figure 2.6: A depiction of hold-type blur for a ball moving with a translational motion of constant velocity. In the top row we show six intermediate positions at equal time intervals taken from a continuous motion. The empty circles denote the eye fixation point resulting from a continuous smooth-pursuit eye motion that tracks some region of interest. For each instance of time, the same relative point on the ball is projected to the same location in the fovea, which results in a blur-free retinal image. The central row shows the corresponding hold-type display situation. Here, the continuous motion is captured only at the two extreme positions. Frame 1 is shown during a finite amount of time, while the eye fixation point follows the same path as in the top row. This time, different image regions are projected to the same point on the retina. Temporal integration registers an average color leading to perceived blur as shown in the bottom row.

when reducing the refresh rate from 120 Hz to 60 Hz. In the limit, smaller hold times can remove hold-type blur completely, but this requires feeding displays at higher frame rates. Interestingly, the problem of hold-type blur was not prominent in old CRT displays, where the light from a cathode tube was flashed only for a very short moment of time. Therefore, frames were not kept in the same position for extended period of time which causes blur in LCD displays.

2.2.4 Image Quality Improvements via Motion

It is expected that motion can potentially degrade the perceived image quality, e. g., by introducing motion blur. Interestingly, it has been shown that in many cases presence of motion can also improve image quality. For example, Schütz et al. [2008] reported a 16 % increase of visual sensitivity during SPEM for foveally presented luminance stimuli of medium and high spatial frequencies compared to the static case. This HVS mechanism serves towards a better recognition of tracked object, which contributes to human survival skills. A similar increase of sensitivity has been observed for isoluminant chromatic patterns. Also visual hyperacuity is maintained for moving

Figure 2.7: Simulation of hold-type blur. An animation sequence with the sample frame as shown on the left is displayed simultaneously with 60 and 120 Hz refresh rates on a Samsung 2233RZ 120 Hz display. The effective velocity of horizontal motion as seen on the screen is the same in both cases. The user's task is to adjust the blur in the sequence refreshed with 120 Hz until the level of blur matches the 60 Hz sequence. The average outcome of such an experiment is shown on the right. In other words, the sequence of blurred frames (right) at 120 Hz are visually equivalent to the sequence of sharp frames (left) displayed at 60 Hz.

targets at uniform velocity in the range 0–4 deg/s [Fahle and Poggio, 1981]. Moreover, an illusory displacement can be observed when displaying two parts of a line with a few milliseconds delay [Burr, 1979] because for both targets the HVS assumes a smooth motion and their different perceived locations are correlated with the delay between their exposure. Fahle and Poggio [1981] stress the role of the constant velocity assumption as an important constraint in the target position interpolation by the HVS.

An interesting visual illusion is the so-called *motion sharpening* [Ramachandran, Rao and Vidyasagar, 1974]. Surprisingly, the HVS seems to be equipped with a motion deblurring mechanism which may cause moving blurred images to appear sharper than their static counterpart. Westerink and Teunissen [1995] have observed that for velocities higher than 15–20 deg/s the perceived sharpness of images blurred with a 6-pixel-wide filter appears similar to the original sharp images undergoing the same motion. Takeuchi and De Valois [2005] investigated the motion sharpening effect by interleaving sharp and blurred frames. The viewers could not see any difference in the video sharpness even if two thirds of the images had been blurred, but they complained about flickering for low refresh rates. This idea has successfully been exploited in video compression and transmission applications [Fujibayashi and Boon, 2008], where selected frames have been filtered off-line to reduce the required bandwidth.

Surprisingly, also time averaging related to hold-type blur, which most of the time reduces image quality, can also improve it. Hara and Shiramatsu [2000] observe that a linear image movement at a specific velocity across the display extends the pass band of the image spectrum for some special pixel color mask mosaic configurations. They show that for some spatial subpixel configurations, such as an *RGGB*-mosaic, the perceived image quality can be improved, in particular by reducing detail discoloring. However, they conclude that the extension of the pass band does not improve the image quality for the standard $|RGB|RGB|\ldots$ arrangement, which is predominant in LCD displays, including the ones considered in this work.

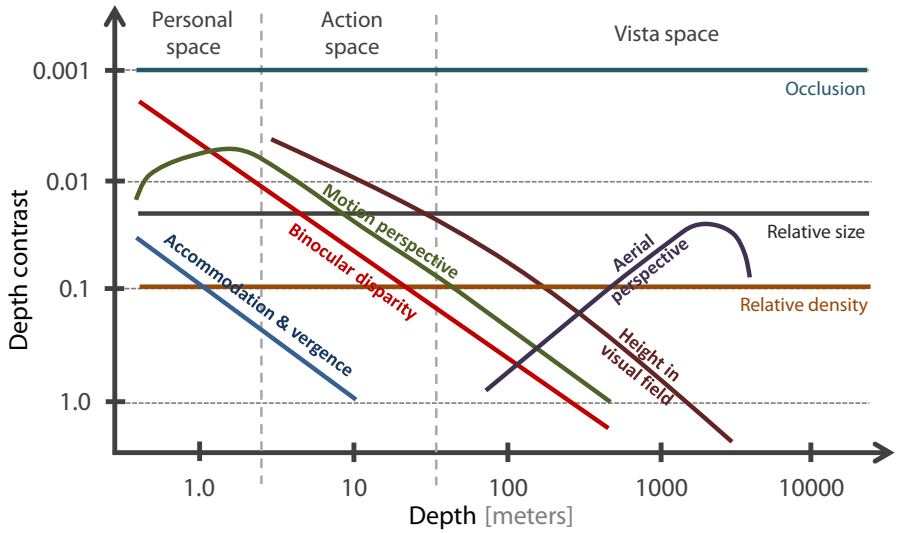


Figure 2.8: Sensitivity to different depth cues. Redrawn from [Cutting and Vishton 1995]

Besides the here-described qualities that can be improved when motion is present in the scene, it has been shown that motion can also improve depth impression for moving objects or moving observers [Rogers and Graham 1979]. Above examples show that although the perception of image motion is quite involved, it does not contradict but under certain conditions often improves quality and more interestingly detail visibility.

2.3 Depth Perception

The quality dimensions described above , i. e., spatial and temporal resolution, are crucial when considering adopting content for different display devices. Such retargeting is recently receiving significant attention in the context of 3D stereo due to the growing importance of this technology in many areas (e. g., video games, feature films and TV production). Although 3D movies, 3D games or even first 3D TV channels are accessible to a wide range of customers, many challenges exist when aiming to produce stereo content that is perceptually convincing. Therefore, it is necessary to account for properties of the HVS, not only in order to create high quality 3D content but also to assure viewing comfort. In the following part, we discuss the foundations of depth perception.

2.3.1 Depth Cues

In order to obtain best layout perception the HVS relies on a large number of different mechanisms that allow us to perceive depth in the real world. They are known as *depth cues* and can be categorized [Palmer 1999] as pictorial (occlusion, perspective foreshortening, relative and familiar object size, texture and size gradients, shadows,

aerial perspective), dynamic (motion parallax), ocular (accommodation, vergence) and stereoscopic (binocular disparity). The HVS exhibits different sensitivity to them (Figure 2.8), which mostly depends on the distance between the observer and the observed objects [Cutting and Vishton 1995]. The HVS is also able to combine the information coming from different cues even if they contradict each other [Palmer 1999, Chapter 5.5.10]. Dominant cues may prevail or a compromise 3D scene interpretation is achieved. An extensive overview of how different cues interact with each other and how those interactions can be modeled has been presented in [Howard and Rogers 2002, Chapter 27].

2.3.2 Disparity (Stereopsis) Perception

Stereopsis is one of the strongest and most compelling depth cues, where the HVS reconstructs distance by the amount of lateral displacement (binocular disparity) between the object's retinal images in the left and right eye [Palmer 1999, Chapter 5.3]. Through *vergence* both eyes can be fixated at a point of interest (e. g., F in Figure 2.9), which is then projected with zero disparity onto corresponding retinal positions. For any degree of vergence exists a set of points featuring zero disparity which is called the *horopter*. All points in front of the horopter lead to non-zero *crossed* (negative) disparity, which increases as their distance to the observer is reduced. Similarly, all points behind the horopter, such as point P in Figure 2.9, feature *uncrossed* (positive) disparity, which increases with the distance to the observer. The disparity at P for the fixation point F is measured as the difference of vergence angles $\omega - \theta$ (Figure 2.9). Note that this is different from the computer vision meaning of this word, where, disparity describes the lateral distance (e. g., in pixels) of a single object inside two images (Figure 2.9). In this dissertation we will use "disparity" in the sense of perception literature and data, while "pixel disparity" refers to the vision definition.

Depending on direction two kinds of disparities exist: horizontal and vertical. Although latter can contribute to depth perception [Howard and Rogers 2002, Chapter 20.3], the contribution is not as big as in the case of horizontal disparities. Therefore, usually only these are considered while vertical disparities are often avoided to assure viewing comfort.

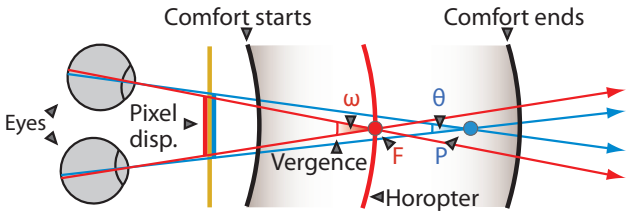


Figure 2.9: Binocular vision.

Binocular Fusion

Although objects in a scene, in most cases, create misaligned retinal images, the HVS is able to fuse them in order to create a single image, instead of perceiving copies of the objects. Usually, it is assumed that this is only possible in a certain

distance/region around the horopter, called *Panum's fusional area*. Outside of this area double vision (*diplopia*) is experienced. Binocular fusion is, however, much more complex process and it depends on many factors such as individual differences, stimulus properties (better fusion for small, strongly textured, well-illuminated, static patterns), and exposure duration. For example Tyler [1973] measured diplopia limits for sinusoidal wave corrugations depending on their spatial frequency. He showed that for such corrugations the HVS is able to fuse low spatial frequencies (0.03 cpd) for very large disparities (~300 arcmin) whereas for high frequencies (3 cpd) double vision is reached very quickly (~2 arcmin). Similar measurements repeated later for square wave corrugations [Tyler 1975] showed that for this kind of corrugations fusion limits decreases significantly, especially for lower frequency patterns. This suggests that binocular fusion operates much more efficiently on gradual spatial depth changes then abrupt once. It is important to remember that the same as stereoacuity, fusion limits vary depending on the subject. It was also shown that for a brief stimulus exposure (200 ms) fusion limits could drop below 27 arcmin for crossed and 24 arcmin for uncrossed disparity, while for longer exposures (2 s) eye-vergence responses have been executed (motoric fusion) that increased the disparity limits to 4.93° for crossed and 1.57° uncrossed disparity [Yeh and Silverstein 1990]. Even though beyond those limits double vision is experienced, perception of depth differences is preserved much longer. It can be observed [Tyler 1973] that depth limits, which are influenced by similar factors to those that influence double vision, are a couple of times higher than diplopia limits.

Disparity Sensitivity

As stereopsis, a low-level depth cue, is one of the strongest cues that is used by the HVS for depth differences discrimination, it is interesting to study how sensitive the HVS is to this kind of signal. This sensitivity can be conveniently studied in isolation from other depth cues by means of *random-dot stereograms* as proposed by Julesz [1971]. It turns out that disparity shares a number of properties with brightness and contrast perception [Brookes and Stevens 1989; Lunn and Morgan 1995; Bradshaw and Rogers 1999]. One of the most important findings from this area is *contrast sensitivity function* (CSF) which defines visibility of different luminance stimuli depending on their spatial frequency. In disparity perception analogous function is called *disparity sensitivity function* (DSF) and it describes visibility of differently corrugated in-depth patterns. This function, similarly to CSF, is found to also depend on the spatial frequency and exhibit similar characteristic. It has familiar inverse "u"-shape with a cut-off frequency around 3 cpd with a peak sensitivity around 0.3–0.5 cpd (cycles-per-degree) where stereoacuity falls into the range of 2–6 arcsec (Figure 2.10). In luminance contrast perception, spatial frequency starts to have neglectable influence on visibility for high contrast values (*contrast constancy*). Similarly to this, for larger-amplitude (suprathreshold) depth corrugations [Ioannou et al. 1993], the minimal disparity changes that can be discriminated (*discrimination thresholds*) are less dependent on spatial frequency [Howard and Rogers 2002, Figure 19.24 d]. Those thresholds exhibit a Weber's Law-like behavior and increase with the amplitude of corrugations [Howard and Rogers 2002, Figure 19.24 d]. Also analogous to luminance maladaptation, where the HVS can hardly adopt to rapidly changing illumination conditions, disparity perception is subject to a similar mechanism. Disparity detection and discrimination thresholds are increasing when corrugated patterns are moved away

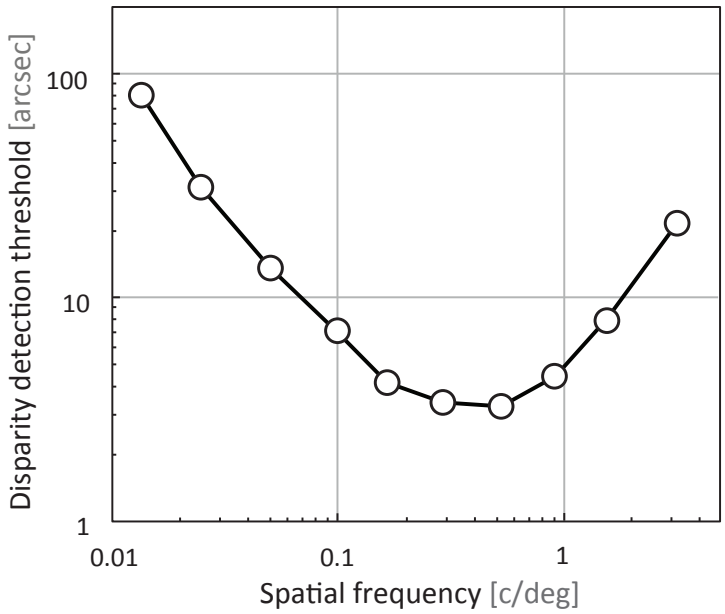


Figure 2.10: Disparity sensitivity as a function of corrugation spatial frequencies. Redrawn from [Bradshaw and Rogers 1999].

from the zero-disparity plane [Blakemore 1970, Figure 6]. The larger the pedestal disparity (i. e., the further the pattern is shifted away from zero-disparity) the higher are such thresholds.

Disparity vs. Pixel Disparity

Another interesting fact from disparity perception is that apparent depth is dominated by the distribution of disparity rather than absolute pixel disparity [Brookes and Stevens 1989]. This is again similar to apparent brightness which is governed by contrasts rather than absolute luminance. While the precise relationship between apparent depth and disparity features is not fully understood, the HVS is most sensitive to regions containing a second-derivative component of disparity and depth is perceived most effectively at surface discontinuities and curvatures, where the second order differences of disparity are non-zero. This means that binocular depth triggered by constant disparity gradients (as for slanted planar surfaces) is weak and such regions are scaled with respect to bordering disparity discontinuity. In fact, those regions are mostly dominated by the monocular interpretation [Brookes and Stevens 1989]. This suggests that from the perception point of view it is more interesting to consider disparity (i. e., pixel disparity changes) rather then absolute pixel disparity.

Cornsweet Effect

This similarity to luminance perception exists also in regards to different illusions related to disparity and luminance contrast. For example, Craik-O'Brien-Cornsweet

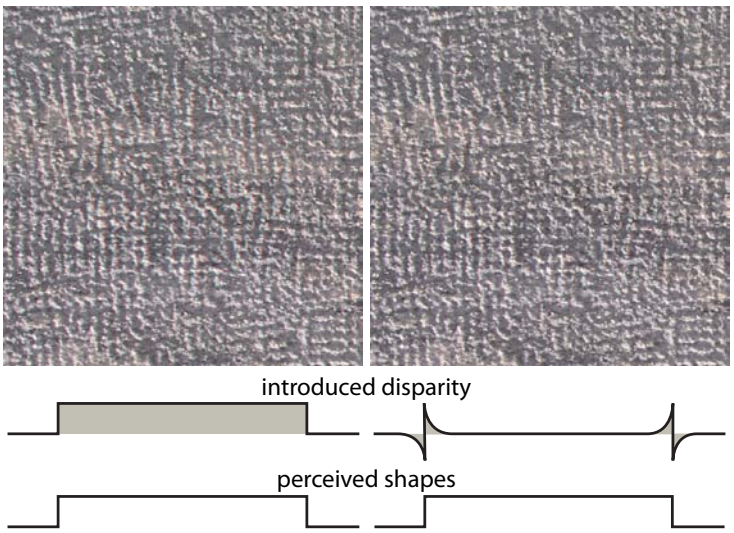


Figure 2.11: Top: A circle with depth due to disparity and apparent depth due to Cornsweet disparity profiles in anaglyph. Texture was required to provide disparity cues. Bottom: The corresponding disparity profiles as well as perceived shapes. The solid area depicts the total disparity, which is significantly smaller when using the Cornsweet profiles.

illusion is a well-known luminance-contrast phenomenon where two regions with the same luminance are separated by a sharp discontinuity with luminance gradually decaying towards equiluminant regions [Kingdom and Moulden 1988]. The two different lightness levels at the discontinuity are propagated by the filling-in mechanisms of the HVS which results in the impression that one region is brighter. Thus, the illusion creates an apparent brightness difference between both regions, which leads to similar appearance as the introduction of physical differences by means of a step function separating the regions, but without the loss of dynamic range [Pratt 1991; Krawczyk, Myszkowski and Seidel 2007]. Different shapes/profiles can be used to produce such a local contrast [Kingdom and Moulden 1988].

It was shown that the Cornsweet effect holds for quite different signals such as perceived line lengths or texture pattern density [Mackay 1973]. Anstis et al. [1978] found that Cornsweet Illusion for depth exists and a depth Cornsweet profile adds to the perceived depth difference between real textured surfaces. The strong apparent depth impression arises at sharp depth discontinuities and is maintained over regions where depth is actually decaying towards equidistant ends. Rogers and Graham [1983] confirmed the effect for random-dot stereograms and concluded that the gradual decay is mostly not noticeable, whereas the visible discontinuity is propagated but the HVS over both regions. This effect is illustrated in Figure 2.11 for textured stereograms, where the screen disparity is directly modulated accordingly to the Cornsweet profile (a symmetric double-spur shaped profile) as shown at the bottom of the figure. Rogers and Graham observed that the induced depth difference over the whole surfaces amounted up to 40% with respect to the depth difference at the discontinuity, which was roughly twice larger than in experiments conducted by Anstis et al. for real surfaces. They further measured that the effect is stronger along the horizontal (i. e., eye separation)

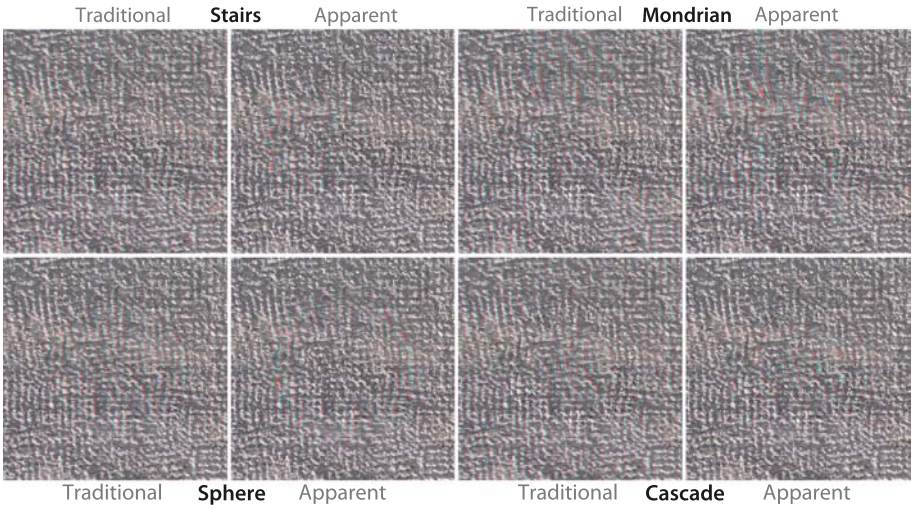


Figure 2.12: Four stereo images in anaglyph are considered where no stereo cue other than stereopsis is present, similar to random dot stereograms [Julesz 1964 Science]. Note that cascading the local Cornsweet profiles still conveys a consistent impression of discrete depth changes, while in the traditional approach disparity accumulation is required for proper stereoscopic effect.

direction, but recent results indicate no significant difference with respect to the orientation [Sato 2004].

The great advantage of the Cornsweet disparity is its locality that enables depth cascading (refer to the depth Cascade and Mondrian in Figure 2.12) without accumulating screen disparity as it would usually be required.

Recently, it was found that other effects associated with lateral inhibition of neural responses (such as Mach bands, the Hermann grid, and simultaneous contrast illusions) that are mostly known from luminance perception, can be also readily observed for disparity contrast [Lunn and Morgan 1995].

Visual Channels for Disparity

The fact that disparity as well as luminance perception exhibit similar characteristics, is an evidence for existing similar mechanisms that luminance and disparity signal undergo. Techniques used in spatial contrast vision, such as masking and adaptation, show that the CSF shape is an envelope of responses for a number of independent channels, which are tuned to different spatial frequencies [Daly 1993]. The same conclusion can be drawn when similar techniques are employed with respect to disparity (refer to [Howard and Rogers 2002, Chapter 19.6.3d] for the survey of relevant experiments). The independent channel bandwidth for disparity modulation has not been clearly established, but existing estimates suggest the range of 1–3 octaves. This imply that disparity and luminance perception could be modeled in a similar way.

2.3.3 Joint Disparity and Luminance Perception

The properties of the human disparity perception which are described so far assume that luminance patterns in the image have no influence on the perceived disparity or that the underlying luminance maximizes disparity experience. Disparity detection and discrimination thresholds are routinely found by applying different depth corrugations to carefully textured images that have good contrast and clearly visible structure. However, such ideal conditions are hardly found in real images, where luminance is often band-limited and contrast can be low. In those conditions it is expected that sensitivity of the HVS to disparity signal will be reduced. In this part of the dissertation we will present the perception background related to the influence of luminance patterns on stereoacuity.

Spatial band-pass channels

Although it is often assumed that correspondence matching in stereo between image patterns in both eyes is achieved at the level of luminance edges, there is direct evidence that band-pass limited channels in the luminance domain play an important role in disparity processing [Heckmann and Schor 1989]. The observation is not surprising since contrast processing in the HVS follows such principles and contrast is required for stereo matching. Hence, one can expect a strong correlation between stereoacuity and contrast characteristics such as the compressive contrast nonlinearity at suprathreshold levels [Wilson 1980] and the contrast sensitivity function (CSF) [Barten 1989], which we discuss next.

Contrast magnitude

Legge and Gu [1989] and Heckmann and Schor [1989] investigated stereoacuity for luminance sine-wave gratings and found that perceivable disparity thresholds decrease with increasing contrast, which can be modeled using a compressive power function with exponents falling into the range from -0.5 to -0.7. Similar results have been obtained for narrow-band-filtered random-dot stereograms by Cormack et al. [1991] (Figure 2.13). They observed a significant reduction of stereoacuity for low contrast (below tenfold contrast detection threshold) which relates to the lower reliability of edge localization in stereo matching due to a poorer signal-to-background-noise ratio in band-pass luminance channels [Legge and Gu 1989]. For contrast at suprathreshold levels, stereoacuity is little affected.

Spatial contrast frequencies

Legge and Gu [1989] measured the necessary luminance-contrast thresholds to detect a fixed disparity for sinewave gratings of various spatial frequencies. They neglect disparity magnitude, but derive a CSF for stereopsis, whose shape is similar to the luminance-CSF shape. Monocular-luminance thresholds are usually assumed to be 0.3–0.4 log units smaller than the luminance contrast needed for stereovision. Pulliam [1981] measured detection thresholds of sinusoidal disparity-luminance corrugations. He found that sensitivity to disparity increases with increasing luminance frequencies from 0.3 to 7 cpd.

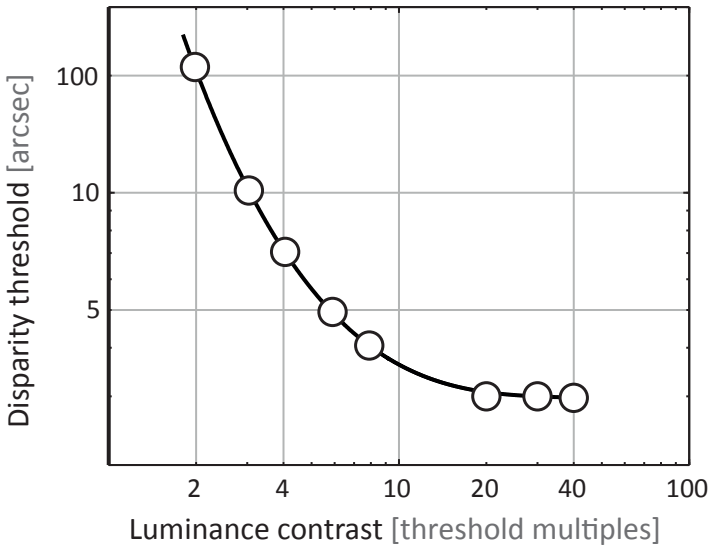


Figure 2.13: Stereoacuity as a function of luminance contrast. Redrawn from [Cormack, Stevenson and Schor 1991].

Lee et al. [1997] measured the impact of luminance frequency on disparity perception for band-pass-filtered random-dot stereograms. They showed that the relationship between disparity sensitivity and luminance frequency exhibits a band-pass characteristic with the maximum located at a luminance frequency of 4 cpd, which is shifted for lower-frequency disparity modulation below 0.25 cpd to around 3 cpd. Their conclusion was that the observed differences in sensitivity result from the stimulation of different visual channels, which are tuned to different spatial modulations of luminance and disparity. They also observed that there is a mostly weak influence of luminance frequency on disparity sensitivity at suprathreshold disparities, except for high luminance frequencies as well as low disparity frequencies. Lee et al. considered relatively narrow ranges of luminance frequency 1–8 cpd , corrugation frequency 0.125–1.0 cpd, and disparities up to 4 arcmin. The results by Lee et al. have been challenged by Hess et al. [1999], who experimented with randomly-positioned Gabor patches with modulated disparity. Hess et al. found that low-frequency disparity modulations were detected equally well for low and high-luminance frequencies. However, for high-frequency disparity corrugations, perception of depth was enhanced when a high-frequency luminance pattern was used, which improves its localization and thus facilitates stereo matching.

Independent-channels hypothesis

As described above, stereoacuity is influenced by luminance spatial frequency as well as luminance contrast. However, the measurements presented so far were performed only for isolated spatial frequencies and contrast values of the luminance pattern. Therefore, it is unclear how sensitive the HVS is to disparity corrugation when the luminance pattern consists of different luminance spatial frequencies and contrasts.

A theory that explains this is the *independent-channels hypothesis* for disparity

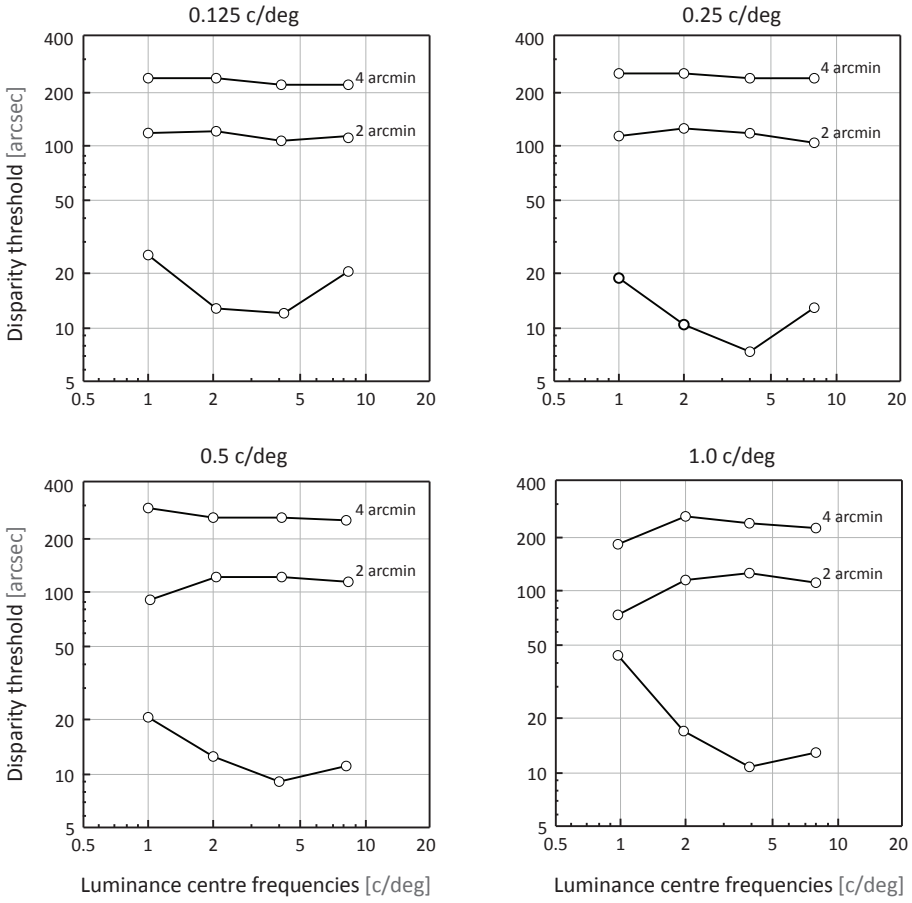


Figure 2.14: Disparity detection thresholds and supra-thresholds for two disparity amplitudes (2 and 4 arcmin) plotted as a function of luminance center spatial frequency. The four panels correspond to different corrugation frequency indicated on top of each panel. Redrawn from [Lee and Rogers 1997].

processing which was presented by Marr et al. [1979]. It implies that stereoacuity is determined by the most sensitive channel and remains uninfluenced by others. This hypothesis has been confirmed in psychophysical studies where stereoacuity has been investigated for independent, as well as summed up sine-wave stimuli of different luminance-contrast frequencies and magnitude [Heckmann and Schor 1989]. It turns out that the phase relationship of sine-wave components, which affects also the local shape of the resulting luminance gradients, is not utilized in stereoacuity. What matters are mostly peak-to-through luminance gradients. Even more convincing is that the thresholds obtained for sinusoidal luminance gratings, for which stereoacuity is best (in the range of 3–10 cpd), are the same as those obtained for multi-frequency square-wave luminance stimuli [Legge and Gu 1989, Fig. 3].

Asymmetry effect

An interesting observation is that asymmetry effects in depth perception can occur (Fig. 2.15), which have not been reported so far. We consider two patches, one in front of the other, each with a luminance texture that we refer to as the *support*, or say that it *supports* the disparity. Limiting the deeper patch to a lower-frequency support, makes the step between the patches less visible and, finally, disappear. When swapping the luminance patterns, the depth difference becomes visible again.

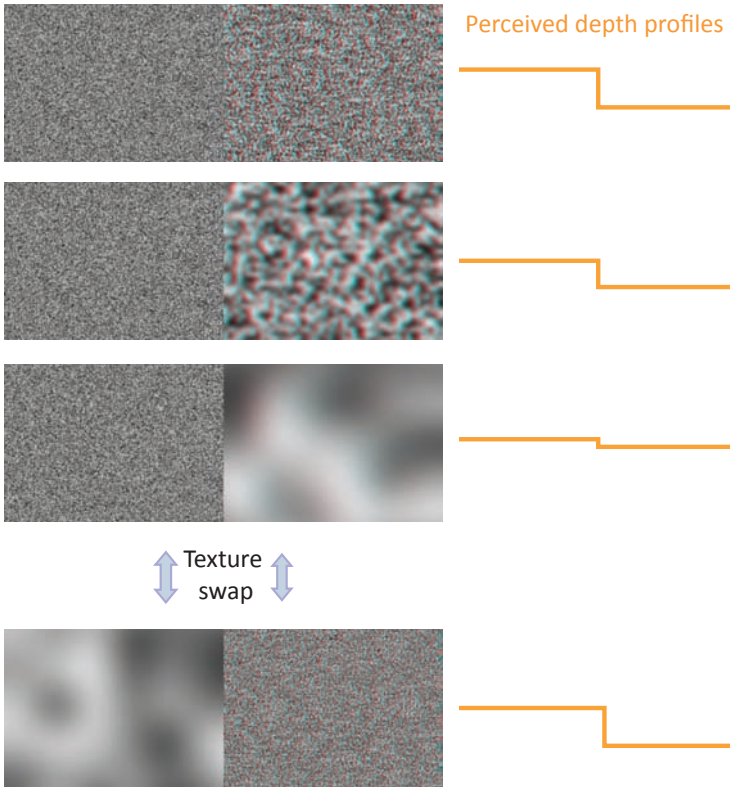


Figure 2.15: Influence of spatial luminance patterns on the depth perception (here shown in anglyph colors). The physical depth of all stimuli is equal, yet the perceived depth (orange profiles, the viewing direction is from top to bottom) varies depending on the applied texture patterns. High-frequency removal from the texture on the right/deeper patch leads to a perceived depth reduction (second and third stimuli from the top). While the third stimulus barely exhibits any perceivable depth, just swapping textures leads to a strong depth impression for the fourth stimulus. Stimuli are shown in anaglyph colors.

One could assume that pictorial depth cues overrule the influence of binocular disparity, which was studied by Marshall et al. [1996] and Mather et al. [2002]. They tested the influence of edge blur on depth discrimination between textures of different blur scales. Their observation suggests that, for a sharp depth edge, the blurred texture should be perceived as more distant than the sharp one, which also agrees with results obtained by Rohaly et al. [1999]. However, this hypothesis disagrees with

our observation that blurring the texture of the distant patch brings it closer to the observer. A different explanation is based on relative size and texture density cues. These normally indicate that lower luminance-frequency textures are closer to the observer, which matches the observation in the first three stimuli in Fig. 2.15. Further, the last stimulus, when textures are swapped, contains agreeing pictorial and binocular cues, leading to a depth impression that matches the binocular disparity. Nonetheless, this explanation disagrees with Marshall et al. and would indicate that the observed depth difference in the last stimulus should be bigger than in the first, which cannot be observed. Recently, Held et al. [2010] have investigated how perceived distance and size are affected by introducing blur.

We develop our interpretation of the asymmetry effect in Chapter 6, based on the fact that the sensitivity to pictorial cues such as texture density or relative size is much lower than sensitivity to binocular disparity in the considered depth range [Cutting and Vishton 1995].

2.3.4 Visual Comfort

As mentioned before, besides the quality of the presented 3D stereo, also viewing comfort plays a crucial role in the production process. In 3D displays, the comfort strongly depends on interactions between eye vergence and accommodation which tends to maintain the display screen within the depth of focus (DOF) that roughly falls into the range of ± 0.3 diopters [Hoffman et al. 2008]. This interaction can be usually easily interrupted by large horizontal disparities. When such are present in the scene, vergence tends to bring the retinal disparity within Panum's fusional area, driving the fixation point away from the screen plane. This way, a conflict between the fixation and the focusing point arises that can be tolerated to some degree by the accommodation-vergence mechanisms, but at the expense of possible visual discomfort [Hoffman et al. 2008]. The most recent model for visual comfort predicting the influence of horizontal disparities was presented by Shibata et al. [2011] (Figure 2.16). Besides large horizontal disparities also vertical disparities can be a source of visual discomfort. Although such disparities cannot be experienced in the real world, they might be created by imperfections in the camera setup [Woods 1993].

Apart from large disparities there are also other sources of discomfort. Most of them are related to inconsistencies between the left and right view (e. g., differences in brightness, crosstalk, magnification, photometric asymmetries). Such binocular image imperfections were recently investigated by Kooi et al. [2004].

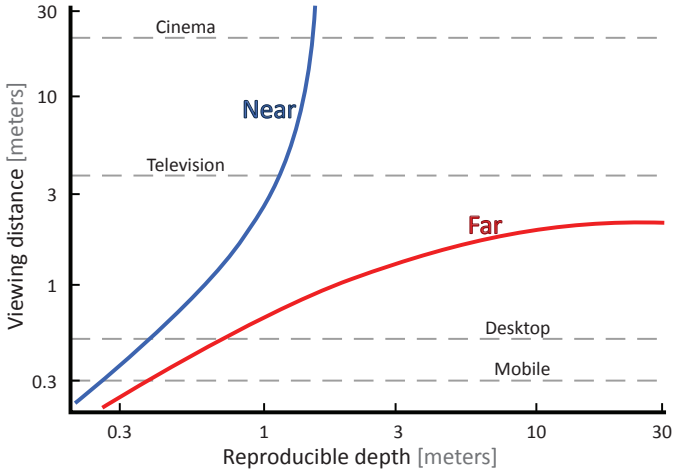


Figure 2.16: The biggest influence on visual comfort in stereo viewing has viewing distance. The smallest range for disparities that assures comfortable viewing is for display located close to the observer. While the screen is moved farther from the observer the disparity budget gets bigger. The red and blue plot represent the furthest and the nearest object that can be displayed as a function of viewing distance. Redrawn from [Shibata et al. 2011].

3

Previous Work

In this chapter, we describe previous work related to contributions presented in this dissertation. We start with an overview of temporal upsampling techniques in Section 3.1, discussing methods that are used for improving temporal resolution in computer graphics as well as in the TV industry. In the same part, we also present techniques related to the synthesis of 3D stereo images. Next, we continue with the temporal domain (Section 3.2) but this time focusing on techniques that exploit it in various ways in order to improve the perceived spatial image quality. In Section 3.3, we discuss recent work in the field of 3D stereo content creation and manipulation.

3.1 Temporal Quality

Under standard conditions, the real world is perceived by the HVS as a crisp and sharp mental image and usually, any object motion in a scene appears smooth. This is achieved by perfect and continuous stabilization of moving object on the retina which is performed by the HVS. However, these conditions are violated when we consider a display device. Moving on screen objects are no longer displayed in a continuous manner but rather discrete, i. e., a moving object is kept for extended period of time (frame duration) at the same position. At the same time human eyes act as they do in reality, i. e., they track moving objects in a continuous way. As described in Section 2.2.3, this mismatch results in hold-type blur, which can be experienced while watching animations or videos on display devices. The blur, introduced to moving objects, spoils the quality of highly detailed content that we can capture with today's cameras and destroys the effect of high-resolution displays that could such content reproduce. For display manufacturers it is a big bottleneck, as in a presence of hold-type effect it is questionable whether further increase of screen resolution to e. g., 4 or 8K is necessary before solving the problem of temporal resolution. The strength of the blur is related to the angular velocity of moving objects. Therefore, the problem gets bigger with growing screen size, which is recently desired in the context of cinemas, visualization centers or more affordable and popular home cinemas. Also, when viewers move closer to their display devices to appreciate highly detailed content, the coverage of viewing angle gets larger which amplifies the hold-type blur. Therefore, today's new designs, apart from improved contrast and brightness, need to assure that the temporal display resolution is sufficient to convey a high-quality image appearance.

A simple solution to reduce the unwanted effect of hold-type blur is to reduce the "hold time", i. e., the time for which a moving object is kept static. This normally

requires a high framerate, which is usually not provided in broadcasting applications. To solve this problem display manufacturers designed techniques that are able to reduce hold-type blur based on a low-framerate stream of consecutive frames only. Such solutions are usually implemented in a form of small computational units in TV-sets, which requires the methods to be very efficient. The problem of low framerate is also common in computer graphics, where temporal resolution is affected by expensive realistic rendering, which needs to be upsampled in order to create smooth animations. Although computer graphics solutions can produce results of higher quality than solutions implemented in TV-sets, they achieve this goal at much higher computational times, which limit possible target framerates (e. g., to 60 Hz). Also, those solutions usually use more efficient graphics units and reuse additional information available during rendering time such as depth maps, motion flow, which makes the problem easier.

In this section we discuss both groups of solutions, i. e., those embedded in TV-sets as well as those that are used for computer graphics applications. We also discuss how they extend to stereo view synthesis.

3.1.1 Industrial Solutions in TV-sets

The key idea of all methods included in new off-the shelf displays is usually to increase the framerate, e. g., to 100 or 200 Hz (respectively 120 and 240 Hz for NTSC content), by introducing intermediate frames produced internally from a low-framerate broadcasting signal. Here, we shortly summarize existing solutions while an extended survey is provided in [Feng 2006].

The simplest solution is *black data insertion* (BDI), which reduces the hold-type effect by introducing new black frames interleaved with the original. This is similar to the way old CRT displays work, where the light is emitted only for a small fraction of the frame time. This solution, however, comes at the expense of drawbacks as well as limitations. Similarly to classic CRT displays it can significantly reduce brightness or introduce temporal flickering, especially in large bright areas, where the HVS temporal sensitivity is high. Further, it can also cause color desaturation.

Instead of inserting black frames, a more efficient hardware implementation of this approach is turning on and off backlight of LCD panel. This procedure is called *backlight flashing* (BF) [Pan, Feng and Daly 2005; Feng 2006] and is possible because in many available displays, LCD panels are illuminated using hundreds of LED's, whose response is very fast. It is, therefore, easy to flash them at frequencies as high as 500 Hz. Besides hold-type blur reduction, such techniques are also useful for reducing the effect of long LC response. In modern devices, backlights are built out of hundreds of LEDs, that are flashed only after the LCD reaches its target level. Although helpful for hold-type blur and response-time problem, this approach, similarly to BDI methods is prone to visible flickering and reduces brightness due to shorter backlight duty cycles. Note that BF and BDI essentially mimic impulse-type displays, such as CRT. They, hence, reintroduce drawbacks of older displays and void the idea of power efficiency of LCD displays due to constant signal over time.

The problems of black data insertion methods can be overcome by not using black frames, but original frames that are duplicated and blurred (*blurred frame insertion* (BFI). Such solution, which was described in [Chen et al. 2005], can on the other

hand cause visible ghosting as the blurred frames are not motion compensated.

The problem of ghosting is solved by another category of methods called *frame rate doubling* (FRD). In those techniques, additional frames are obtained by interpolating pairs of original ones along their optical-flow trajectories [Kurita 2001]. Such methods are commonly used in current TV-sets, where they can easily expand standard 24 Hz content to much higher framerates e. g., 240 Hz, without reducing brightness or introducing flickering problems.

The biggest limitation of frame interpolation techniques comes from the optical-flow estimation, which is a difficult problem, prone to artifacts. All this affects the quality of in-between frames. Our experimental investigations, which we conducted on a modern TV-set using a high speed camera (1,200 frames-per-second), revealed that, although for objects that are moving slowly optical-flow-estimation methods performed by TV-sets work well, such algorithms tend to fail for occlusions, high velocity motion, and highly textured regions. This is due to the high efficiency requirement that needs to be fulfilled by such methods, which cannot deliver a good estimation at low cost. To avoid visible errors optical flow is automatically deactivated in case of doubts and original frames are simply replicated, at the expense of an increasing jaggy motion or hold-type blur. Even such precautions do not help in all situations and objectionable artifacts still can appear for some realistic scenarios. Figure 3.1 shows a pathological case, which cannot be handled by the tested TV-set. Another drawback of interpolation methods comes from the fact that interpolation of in-between frames requires knowledge of future frames. This introduces a time lag which is usually not a problem for broadcasting applications, however, in scenarios where a high interactivity is needed (e. g., video games or interactive visualizations) this lag may not be necessarily tolerated. If an input arrives between two frames (no matter the display frequency), the interaction is visualized only in the next frame. Thus, 60 Hz react with 30 Hz in the worst case. Perceptual experiments showed that subjects could detect delays in the interaction even beyond 90 Hz [Luebke et al. 2002]. Interestingly, it is sometimes stated that beyond 60 Hz, no performance increase is possible [Luebke et al. 2002]. This, however, depends strongly on the task and display. Our study presented in Section 9.2 shows that in dynamic environments higher refresh rates do have an important impact. In fact, temporal visual lags can be perceived as a strong distraction for some cross-modalities, as studied for audio [Dixon and Spitz 1980], haptics [Levitin et al. 2000] or physics [O'Sullivan and Dingliana 2001].

Instead of increasing the framerate, a software solution to reduce the hold-type blur is to apply an image filter to the content to be shown on the screen, that aims to invert hold-type blur. This technique is called *motion compensated inverse filtering* (MCIF). As hold-type blur can be modeled in image-space by a local 1D convolution kernel oriented in the direction of the optical flow, first, motion vectors are locally computed using a space-time recursive search blockmatcher. It is then assumed that eye tracking locally follows these motion which cause the blur and deconvolution technique can be used in order to inverse this process. In practice, it boils down to applying a local sharpening filtering along motion trajectory whose strength is chosen to compensate for hold-type blur as proposed by in [Klompenhouwer and Velthoven 2004]. The effectiveness of such a technique is limited by the fact that hold-type blur is a low-pass filter, which removes certain frequencies. Therefore, none of those frequencies that are completely lost can be restored. Only those that are attenuated by hold-type blur can be recovered by amplifying them beforehand using a linear filtering. What also limits

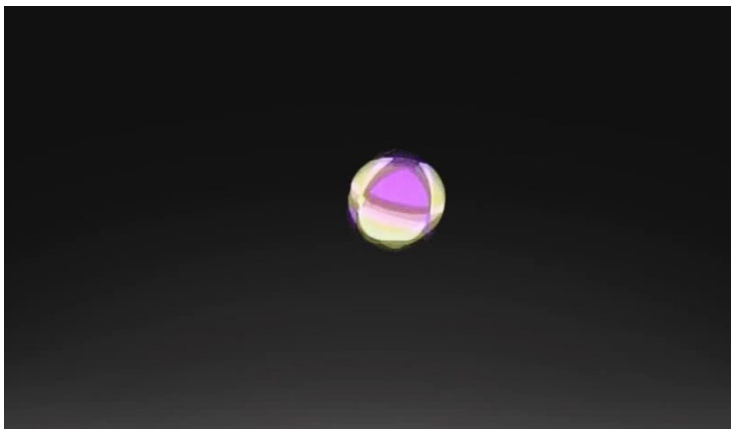


Figure 3.1: Single frame presenting a failure case of off-the-shelf display in performing image interpolation from 24 to 240 Hz signal. Frame was captured using high-speed camera at 300 fps.

this method is the fact that the amount of sharpening that is required for perfect blur compensation would lead to extreme filter band-pass properties, which is not feasible due to possible intensity clipping.

There exists a possibility to combine several of the described methods, for example the in-between frame derivation based on optical flow with the backlight flashing. However details on such custom solutions are not published. In Table 3.1 we present a comparison of different solutions used by the TV industry.

Table 3.1: Comparison of different methods for frame interpolation provided in TV-screens.

	BDI	**BF**	**BFI**	**FRD**	**MCIF**
LCD response required	High	Moderate	High	High	No
Backlight response required	No	High	No	No	No
Optical flow quality	No	No	No	High	Moderate
Ghosting artifacts	Possible	Possible	Yes	No	No
Flickering artifacts	Yes	Yes	No	No	No
Luminance reduction	Yes	Yes	No	No	No
Limitation of blur reduction	Flickering	Flickering	No	No	Freq. cut-off
Other possible artifacts	No	No	No	Fast motion	Oversaturation

3.1.2 Computer Graphics Solutions

The problem of insufficient temporal resolution is also known in computer graphics. However, in contrast to the TV community, it is not directly motivated by the hold-type blur but rather jagginess of animation. In the computer graphics community a high framerate is desired as it has a huge influence on animation smoothness and interactivity. Because high quality image generation techniques are very often time consuming, a high framerate is not always possible to achieve. Therefore, less expensive methods

Figure 3.2: Metamorphosis between two faces using image warping technique. Source: [Wolberg 1998].

for improving the temporal resolution of the content without scarifying overall quality were proposed. Those methods usually do not target such high framerates as solutions in display devices. The usual scenario is to upsample a content that provides a couple of frames per second to a stream that creates an impression of smooth animation (e. g., 30 Hz). This usually lowers efficiency requirements when compared to TV solutions, which in order to produce a 200 Hz sequence can spend only a couple of milliseconds to compute individual frames. Another facilitation in case of some methods developed for computer graphics applications is that they rely on computer-generated content. This makes additional information such as depth or motion flow available for those techniques, which enables solutions that can achieve much higher quality than TV-solutions and still have a huge range of applications to cover, e. g., games, visualizations, animated movies. Here, we present methods for temporal upsampling grouped into three categories: image warping, intermediate frames generation and interactive techniques.

Image warping

One group of methods which aim at creating additional frames are morphing techniques. They do not target directly temporal frame interpolation but rather solve a classic computer graphics problem of transforming one instance of a given object into another. This is a powerful tool for visual effects and was pioneered in 1988 in the movie "Willow". The idea behind this approach [Smythe 1990] was to morph texture as well us underlying shape between two target images, creating a sequence of interpolated images presenting a metamorphosis (Figure 3.2). For this purpose a mesh aligned with shape features was used for shape interpolation whereas textures were blended. Later, this approach was further improved. Beier et al. [1992] proposed to use line pairs as features that guide morphing. Also different kinds of features such as points, polylines or curvatures are possible using a method presented by Lee et al. [1996] where morphing between two images is determined using energy minimization approach. An extended survey discussing more of those techniques was presented by Wolberg [1998].

Recently Lie et al. [2009] used content-preserving warps targeting the problem of video stabilization. With their warping technique they can generate images as if they were taken from nearby viewpoints. This allows them to resynthesize a given video stream as if the camera path were smooth, providing a stabilized video while the original image content remains intact.

Although the here-presented morphing techniques are not designed for temporal upsampling purposes, they can be successfully used in this context.

Intermediate frames generation

Most of the image warping techniques mentioned above do not make any assumption about the similarity of two interpolated images and, therefore, can interpolate between two arbitrary images. This usually requires some user help (e. g., feature specification). However, in case of temporal interpolation for video streams, neighboring frames are usually very similar. This property is extensively explored in methods that directly target temporal upsampling, where the difference between interpolated frames comes from a very small time shift, often not bigger than a fraction of second.

An example of such a method is presented by Mahajan et al. [2009]. Their work is well-suited for a single disocclusion which allows producing high-quality results for a standard content. However, it requires a full knowledge of future frames and is computationally expensive, therefore, it is not really suitable for real-time applications.

Although the quality of additional frames is important and such methods as the one described by Mahajan et al. fulfill this requirement, in the case of temporal upsampling not all regions of the image are equally important. This fact was exploited by Stich et al. [2011] who addressed perceptual effects in temporal upsampling of image sequences. They showed that high-quality moving edges are a key feature for the HVS. Therefore, ghosting produced by temporal upsampling methods as well as ruined by the hold-type blur edge sharpness can be a strong distraction. They addressed those perceptual findings in their method which performs temporal upsampling with improved perceived quality of edges by making their movement more coherent over time via interpolation.

Interactive techniques

Temporal resolution is a crucial problem in the interactive computer graphics. Usually, very expensive global illumination techniques that provide convincing rendering results achieve framerates that are below interactive rates. In such situations, it is desired to use cheaper techniques to generate additional in-between images which can significantly improve the perceived quality. Unfortunately, due to the inefficiency of techniques described above, those cannot be used in interactive scenarios. Methods presented below take advantage of dealing with computer generated content, which can easily provide additional information such as depth or motion flow. This allows for more efficient and effective frames interpolation, which outperforms methods that have access only to interpolated sequences.

The techniques described in this part are often called “image-based rendering” methods and were pioneered by Chen et al. [1993] as well as McMillan et al. [1995] who used a simplified version of the plenoptic model (Figure 3.3) previously introduced by Adelson et al. [1991] and applied it to view interpolation. This function defines an image information visible from any point in space and captures all information which is necessary to reconstruct any view in 3D space. The plenoptic model is defined as a 5D function: position of pixel (3D) and direction of incoming light (2D). Adelson et al. proposed a 7D function where additional dimensions were time and wavelength. One can think of extending it even further to higher dimensional function by, for example, including polarization.

A similar idea was applied in one of the first attempts to increase the number of frames in the context of 3D interactive applications. A method, which was presented

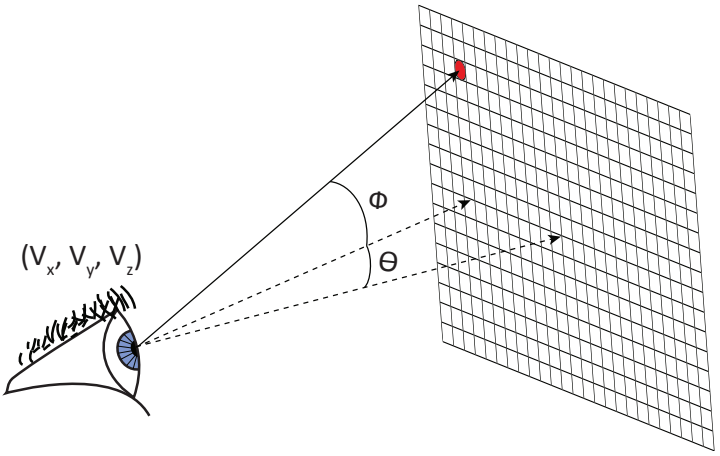


Figure 3.3: Plenoptic model used by McMillan et al. is defined as five-dimentional function. Position of the observation is defined in 3D space by (V_x, V_y, V_z) and the direction where the signal comes from by two angles (θ, ϕ).

by Mark et al. [1997], relies on depth information which allows for easy reprojection of shaded pixels from one frame into another. In this way, their method is able to create many frames out of a single one which was originally rendered. In order to avoid the problem of disocclusions the authors proposed to use two originally rendered views to compute in-between frames. This drastically decreases the amount of unknown information in the interpolated frame and the remaining part is only due to regions that are not visible in both interpolated views. Similar ideas were exploited in later solutions, where a re-use of very expensive shaded samples was used to speed up the image generation process. First such methods, as the one presented in Render Cache by Walter et al. [1999], scattered information from previously rendered frames into new ones by means of forward reprojections. This approach is very effective, e. g, in global illumination where samples are very expensive. Due to problems with occlusions and gaps that needed to be fixed explicitly, Nehab et al. [2007] proposed a new caching scheme where forward reprojection is replaced by reverse reprojection, which also better fits common GPU architectures. Instead of directly using pixel colors (the final result of rendering) from originally rendered frames it is possible to reuse intermediate results. Sitthi-Amorn et al. [2008] presented a method which can automatically choose computationally expensive intermediate values and reuse them in the next frame rendering, speeding up the whole image generation process. Another interesting idea was presented by Herzog et al. [2010]. They explicitly made use of temporal coherence of computer generated animations. Instead of rendering high-quality individual frames, they proposed to render lower-quality images and take advantage of the temporal coherence during the upsampling process. Exploiting time-varying phenomena is interesting particularly in the context of remote rendering, where not only rendering time and quality is important but also bandwidth [Pajak et al. 2011]. More extensive survey on those techniques can be found in [Scherzer et al. 2011].

3.1.3 Stereo-view synthesis

Nowadays, the problem of interpolating, extrapolating and creating low-cost additional frames becomes even a bigger issue as stereo 3D is becoming used in all computer graphics applications (e. g., movies, visualizations, video games). Therefore, the ultimate goal is to provide a content that has not only a high temporal resolution but also allows for stereo viewing. This, however, requires two views (one for each eye) at the same time. It was noted early, that 3D computer graphics is an excellent means to generate stereo images [Morland 1976], simply by rendering two individual views. The main drawback of such a stereo-view creation is that rendering time is doubled. A surprisingly simple form of stereo view synthesis was proposed as early as 1974 by Ross [1974]. Assuming a horizontal moving camera, previous frames look similar to the one eye's view and future frames look similar to the other eye's view. Therefore, playing a video stream with different delays for left and right eyes gives a stereo impression. This approach, however, is limited to horizontal movements and requires knowledge of the future or introduces delay, which are both unwanted for interaction in virtual worlds. Still, the observation that a previous rendering for one eye can serve as a source for the other eye's view encourage to exploit this coherence and lower the cost of second image generation. This makes image-based techniques an interesting approach for creating the second view. Such techniques can be also of high importance in scenarios, where depth modifications are desired either to ensure viewing comfort or for artistic purpose (Section 2.3). In those cases, resulting from the manipulations depth maps do not necessary match real depth and corresponding stereo images cannot directly be obtained in a rendering process.

Many methods described in the previous subsection that target temporal upsampling can be successfully applied to stereo view synthesis. The main difference can be found in the correspondence that is required for creating new views. In the case of temporal upsampling a temporal relation between consecutive frames must be known. Such data can usually be obtained using motion-flow techniques or can be computed from depth and camera settings for rendering approaches. In the case of stereo-view synthesis, this correspondence is pixel disparity which can be directly computed from depth. When such correspondence is known, image warping or reprojection techniques can be applied. Because the relation between left and right view is defined as a horizontal shift, those techniques become faster when compared to temporal upsampling as the samples are reprojected only in the horizontal direction.

Although many ideas from temporal upsampling methods can be brought to stereo-view synthesis, techniques that directly target temporal view interpolation such as [Stich et al. 2011; Mahajan et al. 2009] cannot be easily exploited in the context of stereo as they would required "more-left" and "more-right" images to synthesize new views. Instead, stereo-view synthesis requires a creation of completely new views, which is related to view extrapolation rather then the interpolation.

In many cases when the stereo content needs to be created, only one view without depth information is available. This makes the problem of stereo-view synthesis more challenging and led to techniques that try to compute a stereo image from a single view. One possible solution is to first recover depth information using techniques such as "structure from motion" [Hartley and Zisserman 2000] and, based on this information, do the stereo-view reconstruction. Such a technique was recently proposed by Knorr et al. [2008]. A strength of their method is, that it can work for images without

depth information if they exhibit sufficient features. Zhang and co-workers [2007] combined those two steps, i. e., depth recover and stereo-view synthesis, and developed a technique that avoids direct depth map computation but instead synthesizes the parallax between left and right views directly from the motion parallax present in monocular video.

Instead of synthesizing a stereo image based on one view only, there are techniques that rely on partial information of the synthesized frame. This is similar to temporal upsampling methods where not necessarily all frames are rendered with highest quality [Herzog et al. 2010]. Sawhney et al. [2001] considered a pipeline where one view is rendered in a full resolution and the other view, which is supposed to be synthesized, is available in a low resolution version. They demonstrated how such information can be efficiently used for stereo-view synthesis. Methods like this are particularly interesting if subsets of pixels are used to render subsets of views, which fits well to ray-tracing and volume-rendering [Domonkos et al. 2007].

One of the biggest challenges in synthesizing stereo image pairs are possible artifacts that can usually be noticed at disocclusion regions. Recently, Lang et al. [2010] presented mesh-based warping techniques for stereo-content manipulation, which by solving an energy minimization problem tries to hide possible problems in non-salient regions. The second biggest challenge in stereo view creation is efficiency. Some of the techniques presented here, work offline therefore interactive rates are out of the reach. Other, although achieve interactive speed, still add a significant cost to the rendering pipeline, hence, stereo viewing cannot be achieved at low computational time.

3.2 Spatial Quality Exploiting Temporal Domain

While higher framerates can improve image sharpness and smoothness of animation leading to a visible improvement of perceived image quality, it is worth exploring the temporal domain also in different contexts. Interestingly, it can be used to enhance other image qualities, which are not necessarily directly related to the temporal domain. The key idea for such enhancements is to rely on temporal averaging performed by the HVS and show images which after integration on the human retina convey an impression of higher quality. In this part, we show how such enhancement can be achieved for color and resolution.

3.2.1 Color

The main goal of computer graphics as well as the display industry is to faithfully reproduce the real world. A huge impact on realistic image reproduction has color. Therefore, people for decades have been trying to reproduce real world colors on display devices in a way that they create a believable illusion of not looking on a synthetic content. There has been a huge development in capture devices and acquisition systems which led to a point where we are able to precisely capture color with its high-dynamic-range variations. However, still current display devices are only able to reproduce a subspace of all colors that can be captured.

One of the problem comes directly from the discrete nature of every display device. It is commonly known that most of current displays are able to display 24-bit colors

(i. e., 8 bits per channel). It turns out that this is not enough. When a smooth gradient is shown on a standard device, banding artifacts can be visible (Figure 3.4 a). It is, however, not commonly known that many display devices are not even able to physically display 8 bits, although even standard synthetically generated images on graphics cards usually provide 8-bit color information. Limiting colors to only 6 bits per color channel makes the quantization errors even bigger (Figure 3.4 b) and not acceptable, even if natural images are presented. To display the different nuances (2 bits) which would be normally lost, a simple technique called *Frame Rate Control* exploits the fact that the human eye integrates information over time. Whenever a color is not representable by 6 bits, the screen displays its immediate color neighbors in quick succession over time [Artamonov 2004]. Hereby, the apparent bit depth is effectively increased because the eye integrates the information and reports an average value to the brain. It turns out that if we are equipped with a 120 Hz display, popular now for 3D stereo applications, we can expand dynamic range even further.

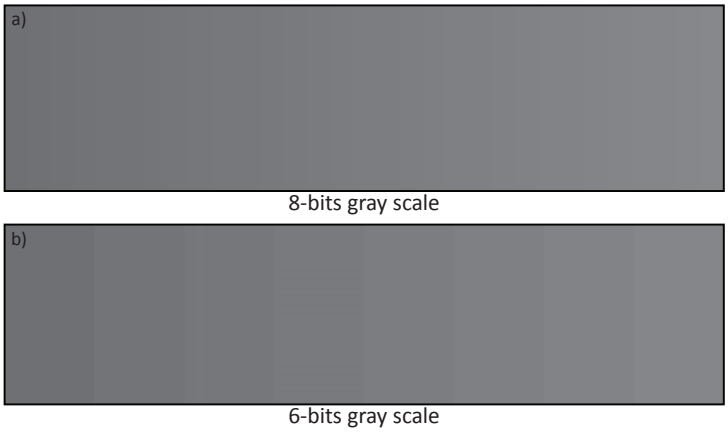


Figure 3.4: Small quantization artifacts can be visible when 8-bit grayscale gradient is presented. The artifacts get stronger when the number of bits is reduced to 6. The effect might be weak in a print due to a poor color reproduction. For better reproduction please see the electronic version of this document.

Such averaging properties of the HVS were explored much earlier in video games where different effects were limited by a small color palette. When for example, in a video game one wanted to add a shadow to the scene, one would usually rely on a darkened version of the affected pixels, yet, the necessary darker tints were not always available. To overcome this limitation, one could draw the shadow only in one out of two frames, resulting in a flickering shadow. If the refresh rate is high enough to exceed the critical flickering frequency (CFF) [Kalloniatis and Luu 2009], the affected colors start to mix with the black of the shadow, hereby, leading to an apparent extension of the available colors. Similar techniques were also used to create transparent objects.

Temporal integration in the HVS is also exploited for color fusion in digital light processing (DLP) video projectors. While many devices rely on a spatial *RGB*-subpixel integration, DLPs display the *RGB* color components sequentially with a temporal frequency over the perceivable flickering limit of the HVS. It is done using a color wheel with three corresponding filters located in front of the light source. The wheel rotates accordingly to displayed channels producing red, green and blue channels of

an image independently, but in rapid succession. These mono-colored images are integrated in the eye and then lead to the impression of a full-color image. Time averaging in the context of color is also used in plasma displays, where the brightness is manipulated by quickly turning primary colors on and off.

3.2.2 Resolution

Besides color, resolution is another quality that is considered as a very important factor, which influences the appearance of the perceived image. There is a continuous development in capture as well as display devices, which enables better reproduction of the unlimited details present in the real world. However, one observation is that although all devices get better in terms of resolution, there is still a huge gap between resolution of screens and currently available content. A straightforward example is scale-preserving rendering, where the resolution mismatch is a big issue. In the real world, we can clearly distinguish individual hair strands, while such details are usually rendered much thicker, hence affecting realism. Metallic paint, as often applied to cars, can have sub-pixel size sparkling effects where a higher resolution increases faithfulness. Fidelity sensitive applications (e.g., product design, virtual hair styling, makeup design, even surgical simulations) suffer from such shortcomings, although todays' rendering techniques allow for rendering content at very high resolutions. Also standard capturing devices such as digital cameras offer today a resolution much higher than the standard resolution of display devices. Besides that, there are computer graphics techniques that allow for creating even higher resolution content. Techniques such as panorama stitching or gigapixel photography [Kopf et al. 2007] can combine multiple images creating bigger images that can easily exceed thousands of megapixels. Also subpixel information acquired via subtle camera motion has proven useful in many applications, such as super-resolution reconstruction [Park, Park and Kang 2003] or video restoration [Tekalp 1995]. In these schemes, subpixel samples from subsequent frames are merged into explicitly reconstructed images, which exhibit significantly higher resolution than original footage.

As mentioned before, this development in capturing techniques outperforms possibilities of current display technologies. The commonly used full-HD standard, which offers roughly 2 megapixels images, is far from what we are able to capture. Therefore, there is a common problem of image resampling [Mitchell and Netravali 1988] which enables displaying continuous or high resolution input images on a finite, lower resolution display. The display image is reconstructed by convolving the input image with a *reconstruction filter* for every output pixel. This way the information that cannot be displayed on the screen (i. e., high frequencies) is removed, and remaining information can be shown on the device. Popular reconstruction filters are Lanczos' and Mitchell's filter [1988]. The latter allows simple tuning of the filter depending on the content and can mimic most of other filters used in image resampling. Although people consider different settings and different filters such a downsampling procedure filters out high-frequency spatial information and leads to the loss of crucial image details. Therefore, people try to come up with ideas to improve perceived resolution by rising the level beyond which all frequencies are lost while displayed on a screen.

One of the simplest idea is to use tiled multi-projector displays [Majumder and Brown 2007], such as PowerWalls, which instead of showing for example one 2 megapixels image, show twelve of them increasing the total resolution to 24 megapixels.

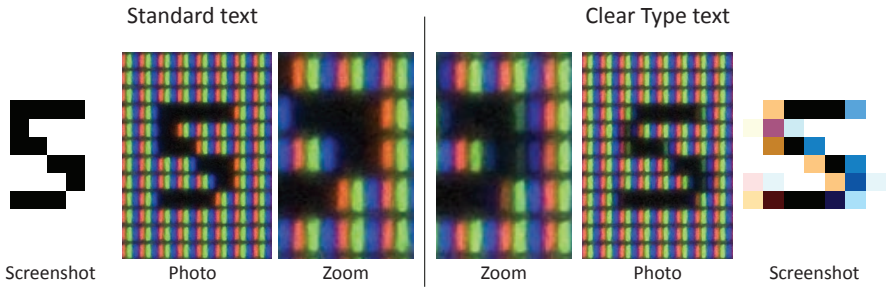


Figure 3.5: A photograph of Clear Type text reveals how subpixels are addressed individually in contrast to the standard scenario where all three RGB component are either turned on or off.

Such solutions, however, are only practical in case of huge screens where the required space for building such a system is not an issue. Those are also solutions that will probably never be available on a consumer-level. Because of this, a much better idea is to use existing displays. One of such concept is to use the subpixel structure of LCD displays, where the color of the whole pixel is created by fusing R, G, and B components (so-called color blending). This idea, called *subpixel rendering*, increases the image resolution by breaking the assumption that R, G, or B channels are unified in a single pixel by controlling intensities of the subpixels independently. If their arrangement is known, one can use channels from neighboring pixels to increase spatial resolution. Platt [2000] showed an optimal filtering for liquid crystal displays (LCD), which was used in the ClearType font technology (Figure 3.5). The resulting resolution enhancement is limited to the horizontal direction only and works best for black-and-white text. Subpixel rendering is advantageous for complex images as well, but saturated colors or naïve compensations of spatial color-plane misalignments may lead to color fringes at sharp edges, as well as color moiré patterns for high frequency textures [Messing and Kerofsky 2006]. This is underlined by Klompenhouwer and de Haan [2003] who found that subpixel rendering shifts luminance aliasing caused by frequencies over the display's Nyquist limit into the chrominance signal. For this reason, optimization frameworks are often used that involve the precise physical layout of subpixels including inactive or defective subpixels [Messing and Kerofsky 2006].

Interestingly, it is possible to enhance the perceived resolution exploiting the temporal domain and properties of the HVS. Instead of registering a content with a subtle camera motion and improving resolution via supersampling techniques, one can rely directly on temporal processing of the HVS. Krapels et al. [2005] reported better object discrimination for subpixel camera panning than for corresponding static frames (independently confirmed in [Bijl, Schutte and Hogervorst 2006]). In their experiments where relatively poor-quality images, captured by an undersampled thermal imager, were considered, object discrimination was improved regardless of the subpixel sensor motion rate, except for *critical velocities* [Tekalp 1995, C. 13] such as a one-pixel shift. A similar observation applies to rendering with supersampling where several images, rendered with slightly differing camera positions, are integrated in order to gain information.

In other techniques, the advantage of HVS temporal processing is used even further. In *wobulated* projectors, multiple unique slightly-shifted subimages are projected

on the screen using an opto-mechanical image shifter [Allen and Ulichney 2005], which is synchronized with the rapid subimage projection to avoid flickering. Due to temporal averaging of the HVS the perceived image resolution and active pixel areas (otherwise limited by the door grid between physical pixels) are enhanced. A similar effect is achieved by *display supersampling* using multiple carefully-aligned standard projectors [Damera-Venkata and Chang 2009], where also an optimization for arbitrary (not raster aligned) subpixel configurations is performed.

3.3 Stereo 3D Quality

Due to big movie productions, such as "Avatar", stereo technology is gaining bigger attention. However, it is not the first time when 3D productions are so popular. Since the first stereoscopic movies in 1922 were released, there was one more big "3D boom" in 50's (Figure 3.6). This one, however, turned out to be not very successful as a couple of years later the number of 3D productions drastically dropped. There were probably many reasons for it. One of them might be that there was no interesting content to show in 3D, therefore, the depth impression could not significantly improve the experience of a standard viewer. Also, there was a lack of necessary techniques for proper stereo content preparation. Such methods are, however, very important as stereo impression that we can experience, e. g., in cinema, is only an illusion and many problems with its faithful reproduction can be observed. Currently, due to the huge development of 3D display techniques as well as new methods for stereoscopic content manipulations, the number of 3D productions does not seem to decrease.

In this section, we describe software techniques that enable stereo content adjustments either for artistic purpose or to assure viewing comfort. We discuss also techniques that were developed for a better understanding of visual discomfort, perceived geometry deformation and the overall quality of depth reproduction. Those techniques require taking into account binocular disparity perception which, as shown in Section 2.3, is similar to luminance perception. Therefore, we also give an overview of some techniques for luminance processing that in many cases served as an inspiration for techniques used for 3D content manipulation. We discuss here also techniques that directly aim for modifying depth. Although they are usually not motivated by 3D stereo they are related to it.

3.3.1 Luminance processing and models

Luminance/contrast is routinely processed automatically by a camera firmware, while advanced users apply specialized solutions using specialized software packages that enable interactive enhancements. Standard techniques include gamma manipulation, histogram equalization and unsharp masking. More advanced methods rely on multi-scale detail control, such as multi-resolution edge-preserving decompositions [Farbman et al. 2008]. Gradient-domain frameworks enable direct contrast manipulation [Fattal, Lischinski and Werman 2002], but require solving the Poisson equation for image reconstruction. Perceptual models of contrast can enable perceptually-linear contrast manipulation [Mantiuk, Myszkowski and Seidel 2006; Mantiuk, Daly and Kerofsky 2008].

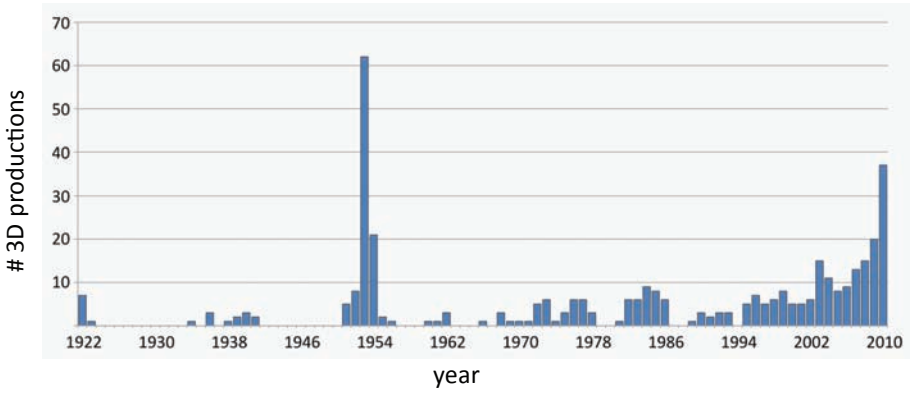


Figure 3.6: The plot visualizes number of 3D productions over last 90 years.

Many different techniques for luminance manipulation were proposed in the context of High Dynamic Range Imaging (HDRI). The most prominent problem in this area is tone mapping, which deals with images of higher dynamic range than the range that can be displayed on a regular screen. Tone-mapping is a range-mapping problem where the signal is luminance in a digital image and the target range is the limited screen luminance [Reinhard et al. 2010]. Many existing local operators [Farbman et al. 2008; Li, Sharan and Adelson 2005] are tuned for the best use of dynamic range and enable multi-resolution manipulation of detail visibility. Notably, it is possible to use certain illusions in order to enhance contrasts that were compressed. Krawczyk et al. [2007] suggest a local operator based on the Cornsweet illusion. In a perceptual framework, they analyze the distortion (i. e., loss) in contrast caused by an arbitrary operator in various bands and re-introduce contrast when possible via Cornsweet profiles. As the reintroduced contrast consists of only higher frequency the resulting physical dynamic range is not changed, but apparent dynamic range of the image can be significantly improved.

Also several techniques exist to enhance luminance based on depth information [Luft, Colditz and Deussen 2006; Bruckner and Gröller 2007; Ritschel et al. 2008; Bezerra et al. 2008]. Those techniques try to make a spatial layout more apparent using similar techniques as those presented before.

Besides extensive luminance manipulation images undergo different kinds of distortions due to for example compression which is necessary to enable content transfer or adjustment for a certain display device. At this point it is important to analyze the introduced differences to prevent unwanted changes. This led to a number of techniques that allow image comparison and are known as 2D image quality metrics. They usually focus on near-threshold [Daly 1993] and supra-threshold [Lubin 1995] difference discrimination as well as on functional [Ramanarayanan et al. 2007] and structural [Wang et al. 2004] differences. For example the Visual Difference Metric relies on the contrast *transducer*, which represents a hypothetical response of the HVS to a given contrast [Wilson 1980; Lubin 1995; Mantiuk, Myszkowski and Seidel 2006]. For a more complete survey on 2D image quality metrics we refer to [Wang et al. 2004].

3.3.2 Geometry

A related method to disparity manipulations is work presented by Weyrich et al. [2007]. They showed how arbitrary three-dimensional geometry can be compressed into the limited range of an almost flat object like a coin (i. e., bas-relief). To this end, they exploited a non-linear global operator and a local gradient-domain decomposition into frequency bands similar to tonemapping presented by Fattal et al. [2002]. In principle, their manual artistic controls enable the addition of a Cornsweet profile into the compressed depth, which enhances small depth differences as discussed in Section 2.3.2.

3.3.3 3D Disparity

Disparity as luminance undergoes different modifications during capturing as well as post-processing. This, as mentioned in Section 2.3, is mostly due to the fact that naïvly captured stereo content usually does not produce a desired stereo impression and often may lead to discomfort caused by large disparities. Also fulfilling an artists' design makes disparity manipulation needed in a post-processing step. Additional modification to the disparity signal can be also introduced in a compression step that is performed in order to reduce bandwidth for later transfer purposes, e. g., broadcasting. Other modifications include disparity smoothing that is applied in 3DTV applications to depth maps derived using computer vision methods to improve the quality of warped images (e.g., better fill disocclusion holes) [Tam and Zhang 2004]. All those manipulations, although necessary, may create also disturbing artifacts. Therefore, care has to be taken while preparing such content. Besides disparity manipulations, the stereo impression can be directly affected by viewing conditions for which the content was not prepared. In this section, we present techniques for both, disparity manipulations, as well as techniques dealing with different kinds of distortions and artifacts of stereo content.

Disparity Manipulations

Disparity manipulation enables fitting the scene's entire disparity range into a limited depth range (called comfort zone) where the conflict between accommodation and vergence is reduced [Lambooij et al. 2009; Shibata et al. 2011]. Adjustments can usually be performed by changing camera parameters during capturing process. Jones et al. [2001] presented a mathematical framework to manipulate interaxial (i. e., distance between cameras) and convergence (i. e., angle between optical axes). Recently, similar approach was proposed for real-time applications by Oskam et al. [2011]. They optimize camera parameters according to control points that assign scene depth to a desirable depth on a display device. This allows not only for keeping the scene in desired range but also for optimization according to an artists' design. Heinzle et al. [2011] presented a complete camera rig that provides a intuitive and easy to use interface for controlling 3D impression. Their setup consists of a computational stereo camera, which can alter its interaxial and convergence during stereo acquisition. An extreme example of stereo content manipulations is *microstereopsis* [Siegel and Nagata 2000], where the camera distance is reduced to a minimum, meaning that a stereo image pair has just enough disparity to create a 3D impression.

The above techniques give a powerful tool to capture stereo content. Due to the view dependence of the stereo impression, also post-processing techniques for disparity adjustment were proposed. They work directly on pixel disparity maps to either compress or expand a depth range to respect limitations of a display device and convey the stereo impression desired by artists. An example of such technique was presented by Lang et al. [2010]. By the analogy to tone-mapping operators they proposed similar techniques to those used in luminance processing, e. g., a non-linear or gradient disparity mapping. For improving stereo impression of important objects in the scene, they also suggest to use saliency prediction. Later, the problem of computing adjusted stereo images pairs is formulated as an optimization process that guides the warping of stereo image pairs while respecting constraints imposed on the resulting disparity, its temporal changes, as well as saliency-driven image distortions.

3D Image Quality

Since 3D content usually needs to be post-processed, it is necessary to provide an automatic check of the resulting quality similarly as it is done by luminance metrics. While it has been recognized that image quality metrics for conventional 2D images should be extended to meaningfully predict the perceived quality of stereoscopic 3D images, relatively little research addresses this issue. Meesters et al. [2004] postulate a multidimensional 3D-image-quality model that incorporates perceptual factors related to disparity distortions, visual comfort, and 3D image impairments induced by the camera configurations, compression, and display technology. In practice, all these factors are considered in isolation and existing quality metrics are mostly driven by 3D image compression applications. A comprehensible 3D-image-quality metric seems to be a distant goal. Here, we give an overview of work that considers three kinds of distortions in stereo content: compression artifacts, 3D image impairments and misperception.

Compression Artifacts MPEG and JPEG compression artifacts in the color information, affect image quality, but have little influence on perceived depth [Seuntiens, Meesters and Ijsselsteijn 2006]. Sazzad et al. [2009] developed a non-reference stereoscopic image quality metric which combines the estimate of blockiness and blur with a disparity measure based on the difference of a zero-crossing rate between corresponding blocks in the left and right eye images. The correlation coefficient between the original and compressed disparity is considered in conjunction with the structural similarity index (SSIM) [Wang et al. 2004] outcome. In another scenario, the per-pixel Euclidean distance between the disparities is directly incorporated into the SSIM along with the existing contrast, brightness, and structure distortion measures. Benoit et al. [2008] report significant correlation with subjective mean-opinion-score data for stereo images, when the disparity error is incorporated into standard 2D image metrics in particular when the SSIM is incorporated.

3D Image Impairments Typical 3D image impairments can have different sources such as camera configuration, compression or imperfections in display technology. Most of them are attributed to higher level (cognitive) aspects of the HVS and affect mostly the 3D appreciation and visual discomfort and are less related to depth perception. Examples of such impairments include distortion of depth plane curvature

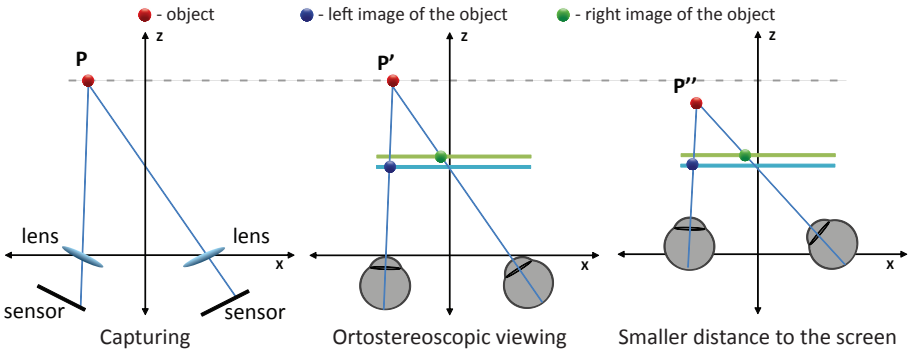


Figure 3.7: Misperception: When viewing conditions correspond to capturing setup (ortostereoscopic viewing) perceived object P' appears at the same depth as the one in the real scene. However, when viewing condition changes (e. g., distance to the screen), the object is perceived on different depth (P''). This leads to distortions of perceived geometry.

(incorrect vertical and horizontal parallax for non-coplanar camera/image plane configurations), puppet theater effect (visual size distortions due to an object angular size and perceived distance mismatch), cross-talk (ghosting due to imperfect eye-image separation), cardboard effect (flat object appearance due to the underestimation of the distance to the 3D display), shear distortion for non head-tracked displays (correct 3D perception limited to one viewpoint), picket fence effect and image flipping (vertical banding and noticeable angular zones in autostereoscopic displays). Meesters et al. [2004] provide a detailed survey of techniques dealing with 3D-image impairments.

Misperceptions A special case of 3D image impairments is misperception of stereo content shown on stereoscopic displays (Figure 3.7). It is often caused by a wrong viewing distance or position. This is because the correct stereo view (ortostereoscopic view) is achieved only when camera parameters used in the acquisition step match perfectly the observer position (i. e., viewing distance, position, interaxial). Whenever those requirements are not fulfilled perceived shapes do not corresponds to the captured scene and the perceived shape is distorted. This is very common in real world scenarios where viewers do not stay in one position while looking at display devices. Recent work by Held et al. [2008] presents a mathematical model for predicting these distortions.

4

Temporal Upsampling

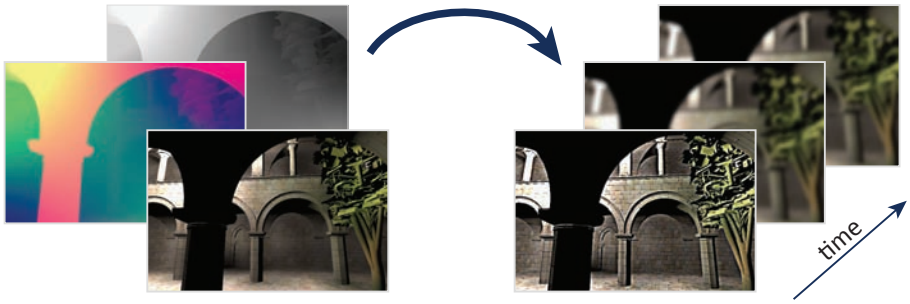


Figure 4.1: Our method performs an efficient perceptually-inspired temporal upsampling taking advantage of additional data provided during rendering.

The continuous quest for better image quality forces display manufacturers to enhance contrast, brightness, display size and pixel resolution. At the same time viewers tend to move closer to the display to enjoy image details due to higher resolution and contrast, as well as to improve immersion in the visual experience, which arises from a wider field of view that is covered by the display. This has profound consequences in terms of the HVS which creates new challenges for display technology as well. Hold-type blur described in Section 2.2.3 can significantly lower image quality as the wide field of view increases the angular velocities of moving objects in the image. The wider field of view increases also the role of peripheral vision, which is tuned through a specialized visual channel to low spatial frequencies and high temporal frequencies as required to detect motion (and timely react for the presence of predators) [Burr 1981]. This increases the viewer's sensitivity to image flickering, which becomes readily visible, in particular for bright displays. Therefore, while designing new display devices and algorithms accompanying them, those issues need to be taken into account.

To overcome those limitations new low-cost high-refresh-rate (100+ Hz) display devices have recently become available on the consumer market and quickly gain on popularity. Also new desktop displays such as the Samsung 2233RZ and Viewsonic VX2265wm FuHzion (120 Hz), besides their primary application in 3D stereo, aim to reduce the perceived blur created by moving objects that are tracked by the human eye. In this case the improvement is only achieved if the video stream is produced at the same high framerate (i. e., 120 Hz). As described in Section 3.1 such content is rarely available. Therefore, many different strategies for increasing temporal resolution

were developed. As the problem is difficult, current solutions always offer a trade-off between efficiency and quality. Many of those techniques can provide superior quality but their efficiency is too low for producing 120 Hz content when only 8 ms can be spent on a single frame computation. This is expected to be an even more important issue for the newly appearing Super-HD displays (4096×2160 resolution). Other techniques, as those included in TV-sets, are very efficient; however quality of resulting sequences often reveal some artifacts, including possible flickering. Another problem is a lag introduced by most of those methods, which is caused by the fact that the knowledge of future frames is usually required to produce additional frames. This is a big issue in interactive applications such as video games.

Although recent desktop displays can be fed externally with 120 frames per second, and do not rely on an internal frame replication described in Section 3.1. The question then arises how to efficiently synthesize frames specifically for displays of this type so they lead to a sharp and convincing image when viewed by a human observer. In this chapter we present an approach that finds a good trade-off between existing solutions.

The chapter is structured as follows. First, in Section 4.1, we give a short overview of our approach. Then, we describe our technique in Section 4.2 and give implementation details in Section 4.3. To check how our upsampling technique performs we conducted an experiment which we describe in Section 4.4. This is followed by conclusions in Section 4.5.

4.1 Overview

In order to reduce hold-type blur we propose to upsample the stream of rendered images using the pipeline depicted in Figure 4.2. Our solution is fast and produces additional frames at a scene-independent cost due to an efficient frame warping which takes advantage of 3D information (e. g., depth, motion flow, occlusions) generated as a by-product of GPU rendering. The additional information allows us to improve the quality of in-between frames and accelerate their computation. Our method avoids artifacts produced usually by solutions embedded in TV-sets, where motion flow needs to be estimated based on input video stream (Section 3.1.1). We also perform extrapolation instead of interpolation, which solves the problem of possible lag.

In order to extrapolate one (or multiple) in-between frames, we use motion flow to warp the previously shaded result. As such extrapolation is fast, it may lead to visible artifacts. We remove them by locally blurring extrapolated frames where required. This step is inspired by the idea proposed by Chen et al. [2005] as well as the motion sharpening effect used in compression (Section 2.2.4). The local blur hides artifacts if warping fails and makes extrapolation sufficiently accurate. The lost high frequencies are compensated relying on Bloch's law (Section 2.2.1). In the context of high-refresh-rate displays, where subsequent images are fused by the HVS, the law suggests that a feature displayed with enhanced intensity in a single frame is perceived in the same way as the same feature present in two frames, each with a halved intensity. We exploit this observation to maintain local average power in each frequency component and amplify high frequencies in fully rendered frames to compensate for the blurred (warped) frames that follow. Exploiting in an inexpensive way those perceptual findings provides naturalness and efficiency.

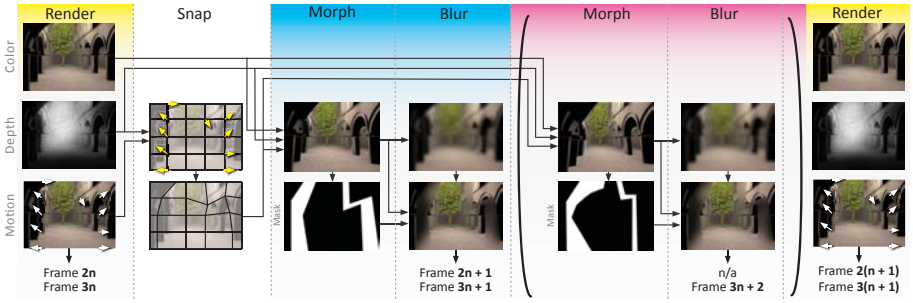


Figure 4.2: Our pipeline, from left to right: To extrapolate one (or multiple) in-between frames, we use motion flow to warp an originally rendered image into an in-between frame, that is then locally blurred to hide artifacts caused by morphing failures. Finally, we compensate for the lost high-frequencies by enhancing them in originally rendered frames where necessary.

4.2 Temporal Upsampling Pipeline

In this part we present individual components of our temporal upsampling technique. First we provide a description of fast warping to later concentrate on the artifacts removal.

4.2.1 Motion flow

Contrary to motion flow from videos, we can extract motion flow during rendering. The graphics card has knowledge about object displacements, which is different from special displays to combat hold-type blur because they need to reverse engineer imperfect 3D motion via optical flow. By taking the difference in position for every vertex, we can compute perfect motion flow and rasterize it into a buffer. While higher-order motion models are possible, a linear assumption proved sufficient in our tests.

4.2.2 Morphing

Morphing takes the original frame and maps every pixel into its new predicted position, but this can be costly. In our implementation, we make this mapping piecewise linear by mapping a subset of pixels – a *grid* – and extrapolating the deformation over this grid.

Morphing might map multiple source pixels to a single destination pixel. We can resolve such ambiguities, by relying on depth, extracted just like the motion flow. Note that such information is not available to image-based approaches such as those used by TV manufacturers (Section 3.1.1).

We will show that blur can remove inconsistencies to a large extent, but morphing a fixed resolution regular grid can lead to significant artifacts that are not easily fixable. E. g., diagonal edges, or discontinuities can fall between grid vertices, such as depicted in Figure 4.3 (left). Our solution to this problem is to snap the grid to deformation

discontinuities (i. e. optical flow) in the *original* domain Figure 4.3 (middle), before morphing them to their new location in the *morphed* domain as seen in Figure 4.3 (right).

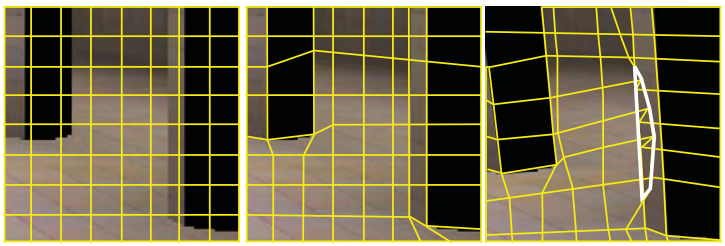


Figure 4.3: To warp the original (Left) into the in-between frame (Right), we proceed in two steps. First, a uniform grid (Left) is snapped to discontinuities in motion flow (Middle). Second, those vertices are warped into a new location (Right). By doing so, discontinuities in motion are preserved, which is an important perceptual cue. Further, conventional depth buffering resolves overlaps (White area) by comparing depth values from the original frame to a depth buffer in the target frame before writing.

Preferably the snapping is done to nearby edges, hereby trading of regularity against adaptivity. Vertices on the image border, are kept at their location in order to prevent undefined regions, e. g., black borders. Note, that this warping does not lead to disocclusion holes, as it is usually the case for reprojection. This makes special hole-filling strategies unnecessary.

4.2.3 Blur

Morphing can result in artifacts, because, even though we handle new occlusions using depth values, disocclusions remain a challenging problem. Disoccluded surfaces are not present in the original frame and selective re-rendering is expensive for rasterization, even when using masking techniques. Especially, the entire geometry would need to be processed again.

Fortunately, the possibility to interleave high-frequency and low-frequency content allows us to improve upon these problems. As small features that appear due to disocclusion would result in high-frequency content, consequently, by blurring the image with a Gaussian kernel, we hide potentially introduced artifacts caused by missing information.

The downside is that such an operation changes the frequency content of the image and we need to compensate by adding back high frequencies. This is difficult for in-between frames due to the lack of information, but it can be done for the original image by subtracting a blurred version. Exploiting the HVS incapability to detect interleaved high- and low-frequency content at high refresh-rates allows us to produce a visually equivalent output by adding increased high frequencies to the original frame only.

4.2.4 Gamma Correction

We need to ensure that the increased frequencies lead to the correct appearance when integrated over time by the HVS. In theory, if we use $N-1$ in-between frames, it is sufficient to scale the high-frequency layer by a factor of N before adding it back. In practice, the process is slightly more involved because we need to also counteract the display's gamma curve. For this derivation, we will assume that the image is stationary, and we denote the high- and low-frequency layer H, L, respectively, γ the gamma exponent of the display, and $\hat{H}$ the modified detail layer needed to ensure a correct compensation for the blurred in-between frames. Over N frames the result should, energy-wise, be equivalent to $N(H+L)^\gamma$. For our $N-1$ in-between frames, we have $(N-1)L^\gamma$.

$$\begin{aligned} N(H+L)^\gamma = &\quad (L+\hat{H})^\gamma + (N-1)L^\gamma \qquad \leftrightarrow \\ \hat{H} = &\quad (N(H+L)^\gamma - (N-1)L^\gamma)^{\frac{1}{\gamma}} - L \end{aligned}$$

Only for $\gamma = 1$, we get $\hat{H} = NH$ (a simple scaling).

4.2.5 Selective Blur

Although it is in theory desirable to apply the blur to the entire image [Chen et al. 2005], some problems can make it preferable to apply the blur selectively.

Introducing blur to in-between frames poses two problems. First, the modified original frame can saturate and exceed the display's dynamic range. Second, we might not always be able to reproduce perfect black levels when relying on blurred frames. Because the blur makes neighboring pixels bleed into black areas, these black pixels can contain grey values in the blurred frames that make the black pixels appear slightly brighter. Although these two problems might at first glance not be related, both are a consequence of physical limitations. For saturation, we exceed the upper bound of displayable brightness, and to compensate for the brightening of black pixels, we would need to be able to display negative values in the enhanced original frame.

To address these issues, we perform a simple analysis after having split the frame in its low- and high-frequency content, as shown in Figure 4.4. We verify whether we cross the boundaries of the displayable dynamic range and, if this is the case, we will reintroduce some of the high-frequency content in the low-frequency layer. If the original pixel is darker than the low-frequency counterpart, we keep the original. If the enhanced original exceeds the limits of the dynamic range, we subtract the exceeded content and shift it to the low-frequency layer. Energy-wise and, thus, integrated over time, these operations deliver the correct result. It might seem necessary to propagate the locations of such modifications to the following in-between frames in order to correctly compensate for these changes, but it is unnecessary because decisions are based on luminance values only.

For disocclusions and strong deformations, the warped grid content cannot be reliable and any high-frequency information increases the likelihood of artifacts. Thus, we maintain these regions blurred and use a smooth-step function that maps grid distortion to blur strength, to blend between the true low-frequency content and our enhanced version.

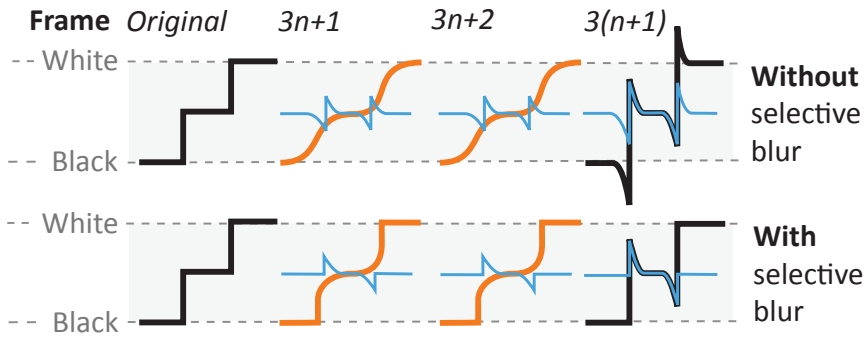


Figure 4.4: Selective blur: The first row presents the situation, where blur is introduced every frame and full compensation is not possible due to limited dynamic range. The second row depicts, how the problem is solved by shifting some content from the original frame into the blurred frames.

4.2.6 Limitations

Some limitations for this approach have to be kept in mind.

Motion Flow

Pixels affected by transparency (e. g., transparent materials, simulated motion blur or depth of field), do not have a simple motion flow. The mapping of such pixels to multiple motion flows and the introduction of strategies for specular materials (e. g., glass), or meshes with changing topology are left for future work.

Morphing

We assume a certain predictability and linearity of motion flow over the in-between frame. Consequently, discontinuities in velocity, might not be well represented, leading to motion that smears over the in-between frame. Our blur somewhat counteracts this phenomenon and the potential problem was not observed in practice, even when the actual motion is highly non-linear (e. g., rotating fan). For more irregular motion this problem is further reduced because of limited tracking capabilities of human observers in such scenarios.

4.3 Implementation

Our upsampling is implemented in vertex and fragment shaders. While current GPUs are very fast, it is still challenging to perform frame extrapolation in a few milliseconds. Therefore we describe implementation details in this section.

4.3.1 Morphing

We morph frames by warping a two-dimensional distorted (snapped) grid of $N \times M$ vertices. To respect discontinuities we want to translate each vertex from its regular grid position to a nearby discontinuity (maximal gradient) in the motion flow. For this, each vertex examines a small neighborhood around its original position (typically 8×8). To avoid snapping two vertices to the same location, we choose the original grid such that no two neighborhoods overlap, but the entire image is covered.

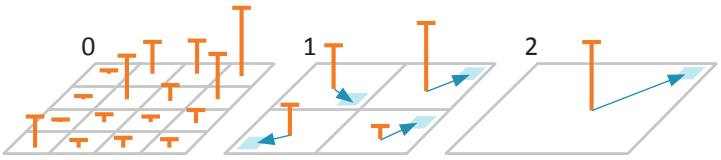


Figure 4.5: Finding the maximum gradient in a neighborhood: At level 0 of a 4×4 grid, different gradients are denoted as vertical bars. Going to level 1, the maximum gradient (vertical bar), as well as a pointer (blue arrow) to the location of the maximum (blue square) is stored on a 2×2 grid. Level 2 stores the maximum and its location.

We find the maximum value and its location by relying on a special form of MIP map (Figure 4.5). For level 0, each pixel stores the gradient magnitude and its coordinates. For successive levels i, we recursively combine four pixels from level $i-1$. We find the maximum gradient value and copy the entire entry to level $i-1$. The result is a traditional max MIP map that additionally stores *where* the maximum occurred. In practice, we encode a relative position with respect to the vertex that will search the corresponding neighborhood. This allows us to quantize the information, 2×5 bits for position, and 6 bits for gradient magnitude, in a total of 16 bits per pixels. This fine-grained parallel strategy leads to a speedup of a factor of two over a sequential loop in the vertex program to find the maximum.

After finding the maximum, we snap the vertex to this location and adjust its texture coordinates to reflect the new position. In this way, the grid is warped and respects discontinuities, but the texture is still undistorted. The distortion only comes from the motion flow, which is then additionally applied to the vertex position.

While we allow (and intend) fold-overs, we still draw a closed, connected grid of $N \times M$ vertices from an OpenGL vertex buffer (2×16 bit per vertex) to achieve a $N-1 \times M-1$ tile grid. Further, we enable the depth test and pass depth from the original frame to resolve occlusions at fold-overs.

4.3.2 Blur

Instead of employing a full gaussian blur, we use a MIP map with a recursive 3-tap binomial weight filter. We then read the MIP map at a higher level using tri-cubic reconstruction. As the blur occurs after tone mapping, it is done using 8 bit RGB values.

4.3.3 Motion Flow

To compute high-quality per-pixel motion flow, each vertex' position is transformed into homogeneous clip space at time t and time $t + \Delta t$. During rasterization, the two homogeneous vertex positions are projected into Euclidean space and their difference produces the optical flow for that pixel which avoids problems at the clipping planes.

Extrapolating frames using previous frame motion flow for high velocities and complex motion is difficult. Fortunately, the tracking performance of human observers is limited and allows us to bound the deformations. Based on the findings in [Daly 1998] tracking is possible up to 80 deg/s.

According to these findings, we simulate the loss of tracking accuracy by a simple function f. For 70 deg/s we assume perfect tracking ($f(70) := 1$), for 90 deg/s no tracking ($f(90) := 0$). Using a cubic smooth-step curve ($f'(70) := f'(90) := 0$) gives good overall results. We extend beyond 80 deg/s because Dali et al. measured random motion, whereas 3D scenes usually exhibit more coherence. In fact, velocity damping is usually preferred, even over 120 Hz (see next Section), as it tends to reduce blur (in this case the motion blur due to imperfect tracking).

4.3.4 Performance

Our implementation allowed making our temporal upsampling very efficient. Time needed for generating two additional frames is significantly lower than 8 ms, which is time budget for one frame at 120 Hz. Table 4.1 presents performance numbers for our technique on an 3.0 GHz Core 2 Duo CPU with an NVIDIA GTX 260.

Scene	Motion Flow	Morph	Blur	Total
Sponza	0.40 ms	1.92 ms	3.34 ms	5.66 ms
Tower	1.64 ms	1.95 ms	3.36 ms	6.95 ms
Fan	0.33 ms	1.86 ms	3.38 ms	5.57 ms
Trees	1.00 ms	1.93 ms	3.38 ms	6.31 ms
Camel	0.49 ms	1.75 ms	3.37 ms	5.61 ms

Table 4.1: Performance breakdown for various scenes (Figure 4.6) when upsampling 40 Hz to 120 Hz (resolution is 1024×1024). If rendering takes more than half of the total upsampling time to produce one frame, our operator is useful as it produces two frames at the same time.

4.4 Experimental Validation

We conducted a series of psychophysical experiments to understand how our temporal upscaling compares to standard rendering methods. In total 5 experiments were conducted. First, we investigated blur-reduction (“Rating” experiment) and possibly introduced artifacts (“Artifacts"). We also checked the impact of our upsampling game-related task performance (“Game”). In “Stereo vision" experiment we showed that our technique could be used for generating additional views for example for stereo-view

synthesis. In this section, we present the study design with statistical data analysis as well as details on participants and apparatus.

4.4.1 Participants

14 participants with normal or corrected-to-normal vision took part in the "Rating" and "Artifacts" experiments. Only a subset of 10 participants took part in the "Game" and "Stereo Vision" experiments, as well as additional study over the "Camel" scene. Subjects were compensated for their efforts with a small fee (14 $). Participants were recruited from the university campus and were mostly students of computer science. Subjects were naïve regarding the goal of the experiment and inexperienced in the field of computer graphics.

4.4.2 Materials and Apparatus

All stimuli were presented on a 22-inch (diagonal) Samsung 2233RZ 120 Hz display of resolution 1680×1050 that was connected to a personal computer with an NVIDIA GTX 260 running in the synchronization mode. The monitor was viewed by the subjects orthogonally at a distance of $60 - 80$ cm. The video sequences and images of resolution 512×512 have been used in all studies except "Game", where the full display resolution has been used. Experiments "Rating" and "Artifacts" required that three sequences are simultaneously shown next to each other in a horizontal arrangement.

4.4.3 Procedures

The participants were seated in front of a monitor running the experimental software in a room with controlled artificial lighting. They received standardized written instructions regarding the procedure of the experiment. In all experiments (except "Game") the time for each trial has been unlimited. In case of the "Game" experiment a unlimited-time practice session has been offered until the subject felt comfortable with the game.

In our study we did not have any restrictions concerning the experience of participants. They all played video games before but of course the level of experience varied. Although Green et al. [2003] have shown that video games can modify visual selective attention, in our case, there seemed to be no direct correlation between detection of artifacts and the level of video game experience. It is important to mention that all subjects who noticed problems with artifacts, reported only slight differences between our upsampling and the original 120Hz rendering. This can be explained by the fact that our method, which upsamples 40 Hz signal, has significantly less information over time than the original 120Hz rendering.

Rating

The goal of the first experiment was to judge the amount of perceived blur by rating three rendering methods: our temporal upsampling from 40 Hz to 120 Hz, and native rendering with low (40 Hz) as well as high (120 Hz) framerate. All three pre-rendered sequences have been simultaneously shown on the screen next to each other in a

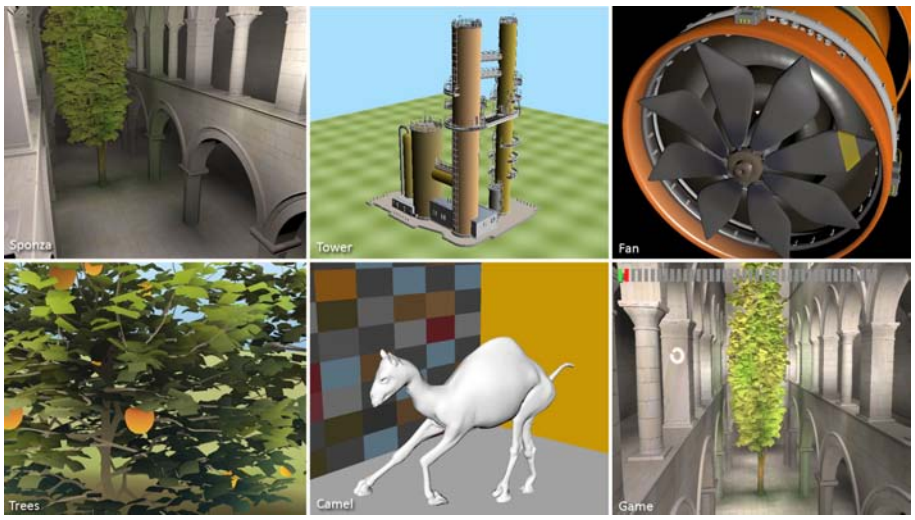


Figure 4.6: The stimuli used. The "Sponza" scene has moderate geometric and texture detail. "Tower" has many occlusions and disocclusions are difficult to extrapolate in image space. "Fan" shows rotational movement that is difficult to extrapolate. Even more heterogeneous movement is found in the "Camel" mesh animation. Many occlusions and disocclusions occur in the "Tree" scene. The "Game" scene was used to measure task performance.

randomized order. Subjects had unlimited time, during which ≈ 20 s long sequences were looped, to rate the perceived amount of blur in the scale from 1 to 9 for each rendering method. The stimuli depicted in Figure 4.6, covering a range of possible applications, such as computer games or medical and technical visualization have been used. We diversified also stimuli in terms motion complexity, which decides upon the eye tracking efficiency.

Figure 4.7 as well as Tables 4.2 and 4.3 summarize the obtained results. Independent ANOVA tests computed for each stimuli revealed statistically meaningful differences in the perceived amount of blur between rendering methods. Adjusted pair wise contrasts (the paired sample t-test with the Bonferroni correction) indicate that for all scenes (except "Camel", which we included to the study as the special case) our method performed significantly better with respect to native 40 Hz rendering and comparably to 120 Hz rendering (in the latter case a statistically significant difference has only been found for the "Sponza" scene).

We included one special scene ("Camel") where it is virtually impossible for an observer to track the fast and complicated leg motion. In this case no hold-type blur is present and all three tested methods performed similarly. Our goal was to show that our method is failsafe for untrackable motion, as it locally reduces morphing based on a prediction of poor eye tracking (refer to Section 4.3). Our rendering outcome is perceived comparable to native 40 Hz rendering, whereas 120 Hz, due to the lack of tracking, results in perceived distinct copies of legs at discrete positions (like a strobing effect in undersampled motion blur rendering [Sung, Pearce and Wang 2002]), which even reduces the overall contrast. To investigate this case further, we informally asked 10 subjects to report on similarity in the appearance of our and 120 Hz sequences with

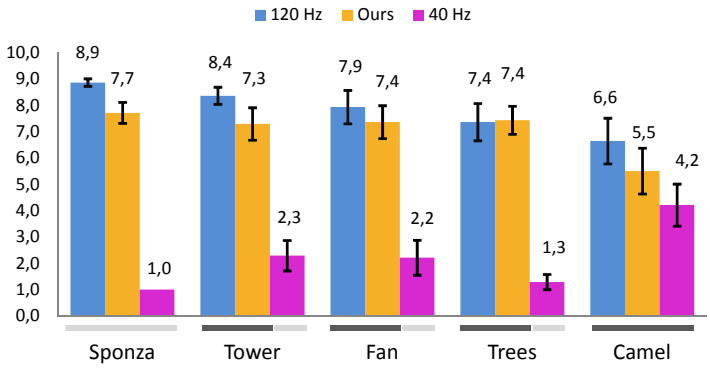


Figure 4.7: Quality rating for 5 scenes. The dark horizontal lines under each scene indicate *no significant statistical difference* (series of t-test). Error bars represent ± 1 SEM (standard error of the mean).

	ANOVA		
Scene	$F(2,26)$	R^2	p
Sponza	302.07895	.93936	<.00001
Tower	38.50120	.66380	<.00001
Fan	24.08909	.55264	<.00001
Trees	43.10340	.68852	<.00001
Camel	2.06101	.09559	.14110

Table 4.2: "Rating" experiment: The table contains F-, p- and $R-squared$ values for ANOVA applied for each scene independently to rating data for our temporal upsampling from 40 Hz to 120 Hz *vs*. 120 Hz and 40 Hz native rendering. $R-squared$ values are computed as a ratio of the explained sum of squares to the total sum of squares.

	Our vs. 120Hz			Our vs. 40Hz		
Scene	$t(13)$	p	Cohen's d	$t(13)$	p	Cohen's d
Sponza	-2.70153	.011988	-1.021081	16.86200	<.00001	6.373233
Tower	-1.54022	.135592	-.582147	5.92260	<.00001	2.238529
Fan	-.64210	.526431	2.130014	5.63549	<.00001	2.130014
Trees	.08069	.936312	.030495	10.17879	<.00001	3.847221

Table 4.3: "Rating" experiment: The table contains t- and p- values as well as the effect size (Cohen's d) for pairwise comparison of our method with respect to 120 Hz and 40 Hz native rendering are given. Note that for the "Camel" scene already ANOVA (Table 4.2) shows that there is no main effect.

respect to a selected static frame. The subjects reported better match in similarity for our method. This observation may suggest that brute-force increasing of the framerate may not always improve the animation appearance, and local frame processing that anticipates the eye-tracking ability is required.

Artifacts

The next important question is whether our method does introduce artifacts as a side effect of blur reduction. In a second experiment that immediately followed the first one, the subjects were presented the same animation sequences again, but this time our method was singled out by a red frame. The subjects were asked whether they see any artifacts in our sequence which they cannot see or are much weaker in the other two sequences. By asking this specific question and giving unlimited time for the answer we wanted to ensure that the subjects carefully analyze the presence of possible artifacts. The side-by-side comparison eases the detection of differences significantly. Further, we did not specify any kind of possible artifacts to not bias the subjects in their observations. The vast majority did not report any observations for any of the sequences (over 82 % responses). Apart from isolated remarks on the differences in shadows (justified), contrast and color changes (the latter two, mentioned in 3 % of the cases, seem to be less grounded), all other comments addressed various aspects of temporal aliasing. The subjects reported that such artifacts, due to undersampling, are slightly more pronounced in our rendering with respect to 120 Hz sequences. Temporal aliasing has been mostly reported (in all but one cases) for "Sponza", "Tower", and "Trees" scenes, where the camera is panning and natural supersampling of pixels fused by the eye is achieved for 120 Hz rendering. Perhaps, this effect can explain the slightly lower rating of our method with respect to 120 Hz as can be seen in Figure 4.7, although aliasing was not directly related to hold-blur rating in this experiment. Similar observations were not made in the context of 40 Hz rendering, probably due to the excessive hold-type blur.

We conclude from those findings that our temporal upsampling is comparable to 120 Hz rendering in terms hold-type blur reduction and overall animation appearance, but with significantly less computational effort. Our technique does not cause additional aliasing with respect to 40 Hz rendering and the slight difference to 120 Hz was only seen by a few subjects.

Game

We finally demonstrate that our approach can lead to a better task performance by a simple game (refer to the "Game" scene in Figure 4.6), in which the participant is asked to tell apart two classes of moving targets. We use a three-dimensional Landolt circles as target classes, which we show in a randomized fashion and ask the participant to press one button when a target is a closed circle or a different button if it is an open circle. Not pressing a button with an object in sight is counted as failure. Pressing a button without an object in sight is ignored. We investigated four rendering scenarios: native rendering with refresh rate of 40 Hz, 60 Hz, and 120 Hz, as well as our temporal upsampling from 40 Hz to 120 Hz. 10 subjects took part in the experiment. On average the scores obtained by the subjects playing using our method were 45 % better than those for original 40 Hz, 12.7 % better than for 60 Hz and 3.3 % worse than for 120 Hz (Figure 4.8). The statistical analysis with ANOVA over the scores for each method reveals the main effect ($F(3,27) = 17.07, p < 0.00001$). Adjusted pair wise contrasts (the paired sample t-test with the Bonferroni adjustment) indicate statistically significant differences between our approach and 40 Hz ($t(9) = 7.71, p < 0.001$) as well as 60 Hz ($t(9) = 4.25, p < 0.01$). No effect has been found when our technique has

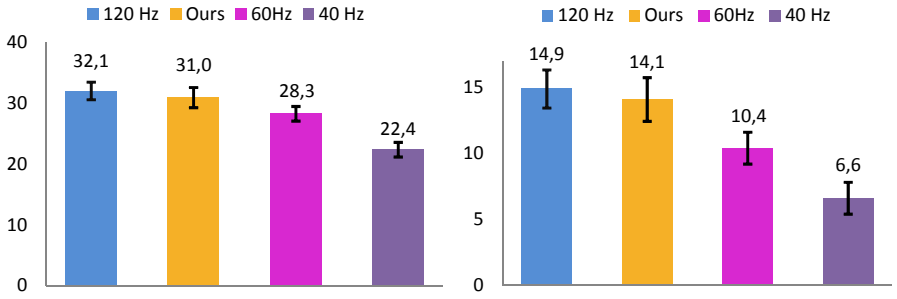


Figure 4.8: Game statistics: The left plot presents the mean value of scores for all methods with ± 1 standard error of the mean (SEM). Such scores include correctly identified Landolt shapes as well as the full circles (toruses). The plot on the right shows only the mean value of the successes of Landolt shape detection with ± 1 SEM.

been compared to 120 Hz ($t(9) = -2.18, p > 0.05$). Details can be found in Table 4.4. We conclude, that the hold-type blur effect can decrease task performance while our approach can restore it to a quantifiable extend.

Refresh rate	$t(9)$	p	Cohen's d
40 Hz	7.71	.00005	2.29894
60 Hz	4.25	.01000	1.02826
120 Hz	-2.18	.12062	-.54211

Table 4.4: "Game" experiment: The table presents outcome of t-tests performed over the subject scores obtained for our temporal upsampling from 40 Hz to 120 Hz *vs.* native rendering with refresh rates 40 Hz, 60 Hz, and 120 Hz, respectively.

Stereo vision

Another application of this work is synthesis of stereo image pairs out of a single image, by warping from central view into the view of each eye. This follows the idea described in Section 3.1.3 that many temporal upsampling methods can be adopted for new view generation. Using our technique, as illustrated in Figure 4.9, we can perform the stereo view synthesis within a few milliseconds. In Chapter 8, we describe a more advanced technique for stereo image synthesis, which is also inspired by our temporal upsampling method. Here, we test whether applying our upsampling directly to new view creation gives satisfactory results.

For simplicity we used anaglyph stereo, but other passive and active stereo techniques would work. 10 subjects in our study were shown a video and were then allowed to freely navigate in a virtual environment. They were asked to compare the synthesized stereo image with the ground truth rendered for two eyes. All subjects have difficulties in perceiving differences between the two approaches.

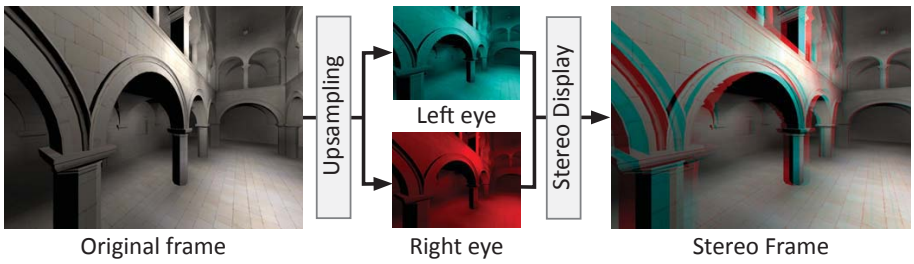


Figure 4.9: Generating stereo frames using our upsampling.

Optical Flow Experiment

We did not include any practical solution relying on optical flow computation used in modern TV sets. The problem is that such algorithms are not revealed and it is currently not possible to send our output into any available TV because it cannot be externally fed with 120 Hz sequences. Therefore, we decided to experiment with one of the state-of-the-art optical flow techniques proposed in [Zach, Pock and Bischof 2007]. The technique is of significantly lower performance than ours (about 30 Hz on a modern GPU with a 512^2 resolution). In a precomputation, we interpolated in-between frames based on the two nearest keyframes, to obtain 120 Hz sequences. We did not include such obtained sequences in our study as visual artifacts have been readily visible (Figure 3.1). Further, such interpolation always implies a one-frame lag.

Comparison to other methods

In our experiment, we did not compare our method to all those described in Section 3.1. Comparing to black data insertion or backlight flashing is currently impossible due to technical reasons. Refresh-rates above 200 Hz are needed and, even though such TV sets are available on the market, they cannot work with more than a 60 Hz input signal. Further, some general drawbacks of these solutions exist (e. g., brightness and contrast reduction). Nevertheless, it could be interesting to combine our solution with those strategies in the future. We also did not compare to motion-compensated inverse filtering because such solutions cannot recover frequencies that are lost by the hold-type effect. Only high frequencies can be enhanced (unsharp masking) to slightly improve the perceived sharpness. We found that, nowadays, the best methods are those based on frame interpolation. For this reason, we compared our method to state-of-the-art implementations of optical flow, which are more accurate than those available in TV-sets. To make our study more challenging we compared our method to the ground truth and showed that our upsampling from 40 Hz to 120 Hz and the original 120 Hz rendering are almost indistinguishable in terms of blur.

4.5 Conclusions

In this chapter, we presented an efficient GPU-based upsampling approach that can reduce hold-type blur significantly for 3D content such as video games or animations.

An interesting avenue of future work is to implement our solution in a small hardware device. Alternatively, our technique could optionally use a cheaper secondary GPU to perform the upsampling task. Combinations of temporal with spatial upsampling or spatial superresolution are worth investigating. In Chapter 8, we show how such a combination can be used for the purpose of stereo images synthesis.

After publishing our technique new solutions for temporal upsampling were proposed. Yang et al. [2011] presented a fast interpolation technique, which uses a fixed-point iteration to find correct mapping of originally computed pixels to interpolated frames. This technique, however, introduces lag and aims for lower framerates, e. g., expanding a signal from 15 Hz signal to 60 Hz. Our method could be used as a complementary step for further framerate expansion. Recently, Bowles et al. [2012] have combined our mesh-based approach with the fixed-point iteration idea. They also presented new applications of those techniques, e. g., motion blur or depth of field rendering.

5

Apparent Resolution Enhancement

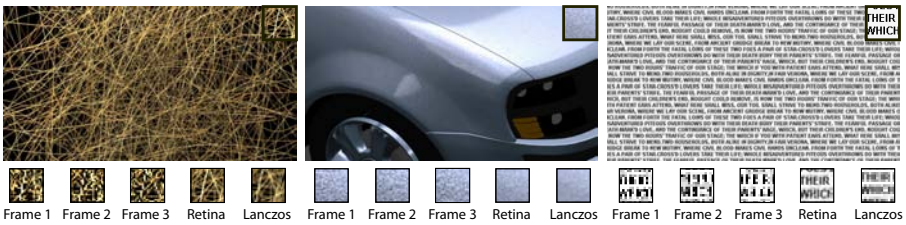


Figure 5.1: Depicting fine details such as hair (left), sparkling car paint (middle) or small text (right) on a typical display is challenging and often fails if the display resolution is insufficient. In this work, we show that smooth and continuous subpixel image motion can be used to increase the perceived resolution. By sequentially displaying varying intermediate images at the display resolution (as depicted in the bottom insets), subpixel details can be resolved at the retina in the region of interest due to fixational eye tracking of this region.

Spatial resolution, although often affected by the temporal resolution, is considered as one of the most important image qualities. This is probably due to details that we can observe in the real world and wish to reproduce on a display. As described in Section 3.2.2, today's capturing as well as computer graphics rendering techniques can provide highly detailed content; however, the image-display stage usually ruins the visual effect. This is particularly striking for smaller devices, which have recently gain popularity, and where resolution is often very limited. Although in many cases, scrolling or zooming may allow for details exploring of larger images, seeing the whole image or larger parts in full detail is often more appealing.

In the previous chapter we showed that in order to achieve high quality content it is not necessary for each frame to be of the highest quality (i. e., the selective blur helped to solve the problem of hold-type blur and achieve the quality of ground truth). Therefore, it is interesting whether quality of images cannot be improved further by careful and individual computation of each frame. In this chapter, we show that it is possible, and skillful frame preparation can improve quality of presented content even beyond the capabilities of a display device.

In this chapter, we present a novel technique for apparent resolution enhancement which makes use of high-framerate displays and is based on temporal integration

properties of the HVS. Achieving higher resolution on lower resolution devices is interesting not only in the context of portable devices. The problem of resolution is bigger when large displays and small viewing distances are considered. Such situations become more common nowadays. Also, the problem of energy consumption, which is associated with increasing pixel resolution, makes energy-efficient high-refresh rates together with additional software solutions a good alternative for achieving enhanced resolution. In particular, the upcoming OLED technology will make another step in this direction enabling refresh rates of thousands of hertz.

The chapter is structured as follows. First, we give an overview of our method in Section 5.1. In Section 5.2, we introduce a model of temporal integration which is later used in the algorithm described in Section 5.3. Our apparent resolution enhancement technique addresses also the problem of flickering as detailed in Section 5.4. In Section 5.5, we evaluate our method in a perceptual experiment before discussion and conclusions in Sections 5.6 and 5.7 respectively.

5.1 Overview

In this part of the dissertation, the problem of several input image pixels that map to the same display pixel is addressed. We want to present them to the observer without applying detail-destructive resolution adjustments. Our main idea is to transform a high-resolution input image into N images of the target display resolution (Figure 5.2), that we call *subimages*. This is done assuming a certain motion of the original image on the screen. We then render the subimages sequentially on a high-refresh-rate display (120 Hz). At the end of each rendering cycle we apply a shift, which corresponds to a displacement of the original image made during the time of N frames, and restart the process from the new position. The result gives the impression of a smooth motion. When an observer focuses on an interesting image detail, the eye will track the feature and SPEM (Section 2.2.2) of matching velocity is established. This is critical for our approach to work because then the subimage details are consistently projected to predictable locations of the retina. By exploiting the integration in the human eye (both temporal, via quickly displayed subimages, and spatial, via rigid alignment of moving and retinal images), the effect of apparent resolution enhancement is achieved. As we rely on a sequential display of subimages, which potentially causes temporal flickering, we analyze the flickering perceptibility specifically for viewing conditions that are relevant for our technique, and apply a flickering reduction step to avoid possible problems.

Our problem is in some sense an inverse problem with respect to super-resolution image reconstruction (Section 3.2.2), where low resolution images are given and the goal is to reconstruct missing high spacial frequency content. We consider decomposition of high resolution image, where the high frequency content is available, into subimages in order to improve the quality of finally perceived images. On the other hand, display supersampling methods described in Section 3.2.2 are similar to our approach. In both cases, high resolution images are needed and the problem is how to decompose those images into subimages, which, when combined either on the screen or on the retina appear as close as possible to their high resolution counterpart. However, instead of using customized setup, we aim at a single desktop display or projector with a limited resolution and fixed pixel layout. Our method works best for displays

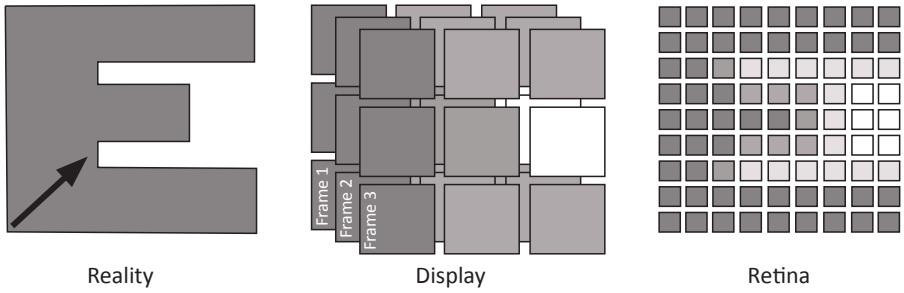


Figure 5.2: Fixational eye tracking over an region of interest in combination with a low-resolution image sequence leads to an apparent high-resolution image via integration in the retina.

with high framerate (e. g., 120 Hz), which are becoming affordable due to rapid gain in popularity of stereoscopic 3D display technology.

5.2 Model

In this section we describe our model of the HVS temporal integration, which is later used for subimages optimization. We do it in a couple of steps. First, we concentrate on a simple case of a single receptor and a constant velocity. Then, we use this model to predict the perceived image for arbitrary subpixel images to later show how an optimization process can be used to transform a high-resolution input into an optimal sequence of subimages.

5.2.1 Photoreceptors

The light response of the human photoreceptor is a well-studied issue by neurobiophysicists. Recently, van Hateren [2005] presented a complete model describing the response characteristics of a single cone stimulated by a light signal. Later, this model was used for encoding high dynamic range video [van Hateren, 2006]. In our work, we need to rely mostly on psychophysical findings, which take into account the interaction between photoreceptors as well as higher-level vision processing.

One particular element is the CFF, introduced in Section 2.2.1. The eye has a certain latency and rapidly changing information is integrated over a small period of time which depends on the CFF. In most cases, we used a 120 Hz screen and displayed three subimages, before advancing by one pixel and displaying the same three subimages again. Hence, each subimage sequence takes $1/40$ of a second. Although this frequency is generally below the CFF and a special processing is needed (Section 5.4), 40 Hz is usually close to the CFF in our context. Higher framerates would allow us to add even further subimages. We detail these points in Section 5.4 and assume for the moment that the subimage sequence is integrated by the eye.

5.2.2 Receptor vs. Changing Image

Human visual system is very sensitive to any motion present in the real word. Therefore SPEM has no problem with compensating for the moving picture on the screen enforcing on eyes proper speed (Section 2.2.2). First, we investigate the retinal response for a standard moving picture observed under such conditions. In contrast to the real world where, during tracking, the signal arriving on a photoreceptor is basically constant, the situation is different on today's displays. A single frame is usually displayed over an extended period of time (hold-type displays) or multiple times (general high-refresh rate screens) instead of flashing the information only once (CRT displays). Thus, for an insufficient frame rate, the eye movement over the screen mixes neighboring pixel information. As a consequence, tracking of screen elements leads to the undesirable *hold-type blur*. In our case, we will make use of this observation to increase the perceived resolution.

To understand the effect, let's derive a simple mathematical formulation. We first consider a static photoreceptor with an integration time of T that observes a pixel position p of an image I. If I changes over time and is thus a function of time and space, the response is given by $\int_0^T I(p,t)\,dt$. If the receptor moves over the image during this duration T on a path $p(t)$, the integrated result is:

$$\int_0^T I(p(t),t)\,dt. \tag{5.1}$$

5.2.3 Retina

In order to predict a perceived image, we need to make simplifying assumptions about the layout of photoreceptors on the retina. While the real arrangement is complex and non-uniform [Curcio et al. 1990, Figure 2], we assume a uniform grid-aligned positioning with a higher density than the image resolution. The latter assumption reflects that in the dense region of the fovea several receptors observe a pixel (Section 2.1).

5.3 Resolution Enhancement

Our goal is to use the temporal domain to increase spatial information and, hence, to enhance the apparent resolution. Unfortunately, as indicated by Eq. 5.1, it is not possible to increase the resolution of a *static* image without eye movement. In such a case, neighboring receptors that observe the same display pixel also share the same integrated information (Figure 5.3, cases A,B).

Precisely, this observation implies that for a given time t_0, $I(p(t),t_0)$ is constant for all $p(t)$ in the same pixel and $I(p(t_0),t)$ is constant during the time that we display the same pixel intensities. Therefore Eq. 5.1 becomes a weighted finite sum of pixel values:

$$\sum_{t=0}^{T} w_t I(p(t),t). \tag{5.2}$$

This equation reveals two crucial elements. First, the simulation can be achieved via a simple summation which will allow us to apply a discrete optimization strategy.

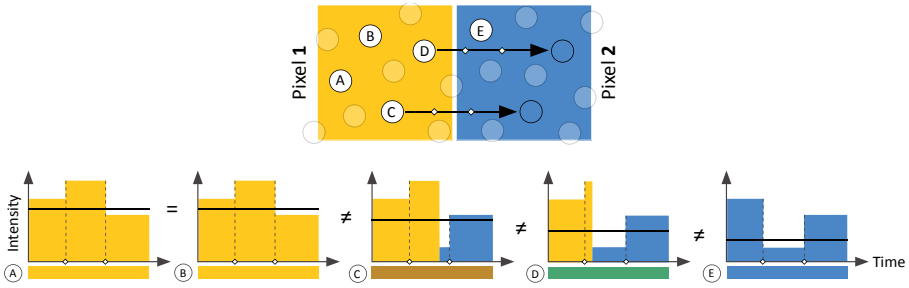


Figure 5.3: Spatio-temporal signal integration. Left: Two pixels (yellow and blue square) covered by receptors (circles). Right: Intensity response of receptors A–E over time for changing pixel intensity in three frames (dotted lines). For static receptors (A, B and E) the resolution cannot be improved because the same signal is integrated over time. Due to motion (arrow), receptors C and D, although beginning their integration in the same pixel, observe different signals which we exploit for resolution enhancement.

Second, for differing paths $p(t)$ (even if only the starting points differ) the outcome of the integration generally differs. This will be key in increasing the apparent resolution. Due to the changing correspondence between pixels and receptors during SPEM, as well as the temporally varying pixel information, differing receptors usually receive differing information (Figure 5.3, cases C,D). Consequently, we can control smaller regions in the retina than the projection of a single pixel.

5.3.1 Simple Case

Before generalizing our approach, we will first illustrate the simple case of a static high-resolution 1D image I_H. For each high-resolution pixel we assume a single receptor r_i, while our 1D display can only render a low-resolution image I_L. Let us assume for now that the resolution of I_H is twice as high as the resolution of I_L and that the image is moved with a velocity of half a display pixel per frame. In theory, we could change the value of each display pixel on a per-frame basis. Nevertheless, we assumed that all receptors track the high-resolution image perfectly. Hence, after two frames, all receptors have moved exactly one screen pixel. We find ourselves again in the initial configuration and the same two-frame subimage sequence can be repeated.

For this particular case, each receptor will, while tracking the image, either see the color of exactly one pixel during the duration of two frames or of two consecutive pixels. More precisely, following Eq. 5.2, receptor i captures:

$$r_i = \begin{cases} (I_L(i,0)+I_L(i,1))/2 & : i\%2 == 0 \\ (I_L(i,0)+I_L(i+1,1))/2 & : i\%2 == 1 \end{cases} \tag{5.3}$$

In order to make the retinal response best match I_H, r_i should be close to $I_H(i)$. This can be formulated as a linear system:

$$W \begin{pmatrix} I_L^1 \\ I_L^2 \end{pmatrix} = I_H, \tag{5.4}$$

where I_L^t is the subimage displayed at time t and W a matrix that encodes the transfer on the receptors according to Eq. 5.2. In the scenarios we considered, the matrix W is

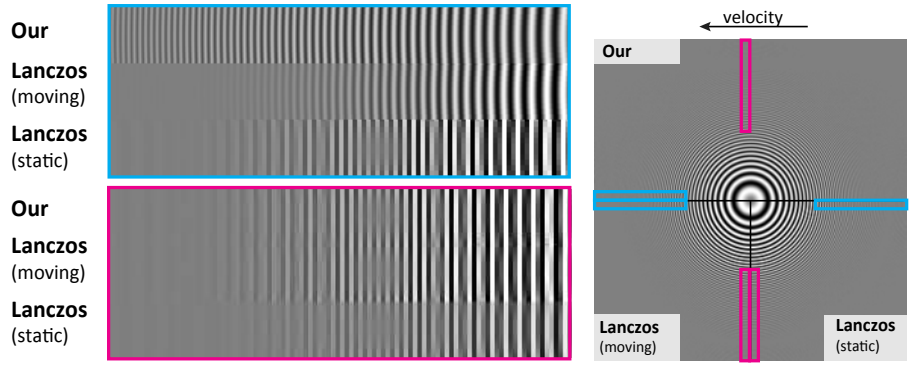


Figure 5.4: Our method vs. Lanczos. Our method uses image motion to improve the perceived resolution along this movement direction by showing 3 subimages on a rapid display. While we rely on a frame optimization, moving Lanczos resampling derives subimages by filtering the translated original image. The eye integration is computed by blending the subimages, assuming perfect tracking.

overdetermined, meaning that there are more independent equations in our system than variables (unknown pixels in subimages). We usually assume that there are fewer pixels displayed over time than the total resolution of the original image and the resolution of the retina is considered to be at least as high as the resolution of the original image. We find the final solution using a constrained quadratic solver [Coleman and Li 1996]. While a standard solver would also provide us with a solution that is coherent with respect to our model, a constrained solver respects the physical display limitations with respect to brightness. Therefore, this approach guarantees that the final subimages can be displayed within the range of zero (black) to one (white). Our problem is convex and so convergence can be guaranteed.

It is natural that subimages contain aliasing. The receptors will integrate the image along the motion path and therefore filter the values. On the other hand, our optimization minimizes the residual of the perceived final image with respect to the original high resolution version. Therefore, as long as the original frame does not exhibit aliasing problems, the optimization should avoid aliasing in the perceived image as well. Although it is difficult to formally prove this cancellation, no aliasing problems were observed during our experiments.

Figure 5.4 shows an example of a horizontal movement. The resulting spatial apparent resolution is much higher horizontally (blue) than for a standard bandwidth-filtered image, while vertical resolution (red) is similar to the case of a moving Lanczos resampling.

5.3.2 General Case

An important property is that an integer movement allows us to reuse the subimage sequence after a few iterations. This is interesting for static images where one can choose a displacement direction and enhance resolution using only a small amount of texture memory.

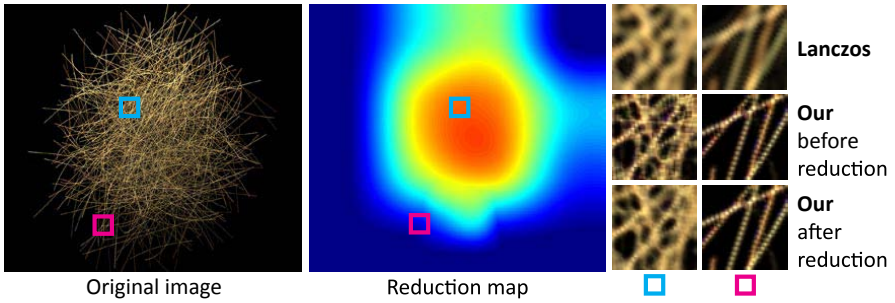


Figure 5.5: Flickering reduction. Left: Original high resolution image. Center: Reduction map. Right: Outcome of Lanczos filtering, as well as our approach before and after flickering reduction for the marked regions. Note that in the regions of strong temporal contrast reduction an improvement over Lanczos filtering is visible. Similar to Figure 5.4, images for our approach are simulations of perceived images assuming motion and perfect eye tracking.

It is possible to treat more general movements by adapting the integration weights w_t from Eq. 5.2. Basically, the weights should be proportional to the time, that a pixel's color is seen by a photoreceptor. To formally compute these weights, we introduce one weight $w_{x,y,t}$ for each pixel value $I^t_{x,y}$ where x, y is a discrete pixel position and t the discrete time interval during which the pixel's color is constant, such that: $\int_0^T I(p(t),t)\,dt = \sum w_{x,y,t}\, I^t_{x,y}$. It follows:

$$w_{x,y,t} := \frac{1}{|p|} \int \chi_{(i,j)}(p(t))\, \chi_k(t)\, dt, \tag{5.5}$$

where χ describes a characteristic function and $|p|$ is the total length of path p. Precisely, $\chi_{(i,j)}(p(t))$ equals one if $p(t)$ lies in pixel (i, j), else it is zero, $\chi_k(t)$ is a similar function to test the time intervals. One underlying assumption is that the receptor reaction is immediate with respect to a changing signal. Consequently, temporal integration corresponds to a box filter in the temporal domain.

5.4 Flicker Reduction

The previous result respects the limits of the display device, but it does not necessarily respect the limits of the HVS. We made the crucial assumption that the HVS integrates a fixed number of subimages and our method only works if their pixel information is fused without producing objectional flickering. To arrive at a flicker-free solution, we proceed as follows: First, a perceptual flicker model computes the amount of flicker for every pixel in the optimized sequence (the *reduction map*). Second, we use this map to define a pixel-wise blending between our potentially flickering, but optimized sequence and a never-flickering, non-optimized standard filtering sequence (Figure 5.5).

5.4.1 Flicker Detection Model

The flicker detection model, used in our solution, is multi-scale, conforming to the scale-dependence of the CFF. It derives per-scale reductions that are pushed to the pixel level where the final contrast reduction happens. In detail, we first compute the maximal intensity fluctuation in each pixel of our subimages. Because flickering is strongly scale-dependent [Mäkelä, Rovamo and Whitaker 1994], we cannot just rely on these values. We use a gaussian pyramid to add a scale component. For each level, this results in a fluctuation measure of the corresponding area in the original image. We can then rely on the perceptual findings in [Mäkelä, Rovamo and Whitaker 1994, Figure 1], to predict the maximally-allowed temporal variation that will not lead to perceived flickering for such an area (measured as an angular extent). If we find that these thresholds are exceeded, we compute by how much the temporal fluctuation needs to be reduced. We then propagate these values to the lowest-pixel level by taking the maximum reduction that was attributed to it on any of the higher levels (refer to the flickering map in Figure 5.5). The maximum ensures that the final flickering will be imperceptible on all scales.

5.4.2 Flicker Sensitivity vs. Pattern Spatial Extent

Related experiments with flickering visibility of thin line stimuli (with the angular length up to 2°) indicate a low flickering sensitivity, both in the fovea and periphery [McKee and Taylor 1984, Figure 5]. Further evidence exists that the sensitivity generally drops rapidly for small patterns [Mäkelä, Rovamo and Whitaker 1994]. This is of advantage to our method as it hints at flickering being mostly visible in large uniform regions. As these uniform regions are those lacking detail and, consequently, our subimages will strongly resemble the original input, any value fluctuation is eliminated.

Hecht and Smith [Kalloniatis and Luu 2009, Figure 10] found that for a stimuli of 0.3° angular extent and adaptation luminance below 1000 cd/m^2, the CFF does not exceed 40 Hz. Similar observations can be made in the Ferry-Porter law that indicates a roughly linear CFF increase with respect to the logarithm of time-averaged background intensity up to 40 Hz where the CFF starts to stagnate and the law no longer holds. This seems to indicate that the choice of three intermediate images for a 120 Hz display is very appropriate. In practice, we encountered very few flickering artifacts when displaying a three-subimage solution unprocessed. Consequently, our postprocess leaves most of the original solution unaltered. Nevertheless, when longer integration times are needed, either because more subimages are added or the display's refresh rate is reduced, the processing improves the result significantly. On a 120 Hz display, four subimages became possible without visible flickering. Such a case is illustrated in Figure 5.5. Four subimages lead to more details than the three subframe solution and we can work with lower velocities.

5.4.3 Discussion

Our approach keeps the highest amount of detail possible while ensuring that the outcome does not result in a perceivable flickering as detected by our flickering model. The blur in Figure 5.5 (bottom-right) is a natural consequence of this trade-off between

detail/flickering and low-resolution/no-flickering. Since our optimization guarantees that the resulting image fits to the display range, which is also the case for energy-preserving Lanczos filter, any interpolation between such a pair of images cannot cause intensity clipping. Artifacts, e. g., ringing cannot occur, because the reduction map, used for blending, needs only to be a conservative bound in order to detect perceived flickering. Hence, it is possible to find a conservative band-limited image (in practice, a dilation followed by smoothing).

One alternative flicker suppression would be to incorporate the constrains on the maximal temporal fluctuations of signal into the optimization, but this has disadvantages. The process would no longer be quadratic, endangering convergence. It would increase computation times and put pressure on the hard constraints needed to match the display's dynamic range.

A second alternative would be to suppress flickering via temporal smoothing, but such attempts prove inadequate. Temporal smoothing combines information that should be kept separate to achieve the resolution enhancement according to our model. To illustrate this, consider the receptor C in Figure 5.3 moving from one pixel to the next at time t. Filtering over time, would introduce information in the first pixel that occurs after time t, this information was not supposed to be seen by C which at time t is already in the second pixel. We exploit this combination of time and space in our model.

Our flicker reduction, is general and is executed in milliseconds on the GPU. It could be used in other contexts, e. g., to detect and then remedy temporal aliasing for real-time rendering.

5.5 Experimental Validation

To illustrate the versatility of our approach, we present several application scenarios and tested them in a user study in order to illustrate their effectiveness. In this section we describe procedures of our experiments as well as details such as participants, materials and apparatus.

5.5.1 Participants

14 participants with normal or corrected-to-normal vision took part in the main part of experiments. In an additional 3D rendering part five participants were considered. Subjects were compensated for their efforts with a small fee ($14). Participants were recruited from the university campus and were mostly students of computer science. Subjects were naïve regarding the goal of the experiment and inexperienced in the field of computer graphics.

5.5.2 Materials and Apparatus

All stimuli were presented on a 22 inch (diagonal) Samsung 2233RZ 120 Hz display at its native resolution 1680×1050 that was connected to a personal computer with an NVIDIA GTX 260 running in the synchronization mode. We investigated also lower

Figure 5.6: Images used as stimuli in our experiment with high resolution images.

resolutions to address the fact that displays constantly grow, often already exceeding 100 inches, but keeping their resolution on the level of full HD. On a 100-inch screen, pixels would approximately be four times bigger than in our experiments. The monitor was viewed by the subjects orthogonally at a distance of 50–70 cm. Because some experiments required that two images are simultaneously shown next to each other in a horizontal arrangement, the video sequences and images of resolution 600×600 have been used in all studies. We considered a 120 Hz refresh rate, decomposing the original images into three subimages to illustrate that the details are also visible for the faster-moving variant (compared to four subimages).

5.5.3 Procedures

We conducted a couple of different studies whose procedures are described here. In all of them the participants were seated in front of a monitor running the experimental software in a room with controlled artificial lighting. They received standardized written instructions regarding the procedure of the experiment. In all experiments the time for each trial has been unlimited.

High-resolution Images

In our first study we considered five stimuli (Figure 5.6), including detailed rendering and text as well as natural images (photographs of a cat and a car). The hair and car images have been rendered with a high level of detail and include subpixel information from elongated hair strands and tiny sparkles in the metallic paint. Text was used to evaluate readability as an indicator of detail visibility. Finally, we used photographs to check the performance of our method for real images, which often exhibit slightly blurry edges with respect to synthetic images. In case of the car photograph we were interested in the perceived appearance of regular structures with details in all directions. Our aim was to show that our method outperforms standard image-downsampling techniques. We tested various velocities and compared our method to Lanczos resampling as well as Mitchell and Netravali [1988], asking people to compare the detail visibility.

The images have been moved in different directions and the subimages have been obtained as a result of the optimization procedure described in this chapter. For each moving direction the velocity has been chosen so that precomputed three subimages can be sequentially repeated.

In the first part of the study, subjects compared the static reference image of high-resolution that was placed on the right to a moving image on the left. The left image was per-frame Lanczos-filtered or our solution, initialized randomly and not labeled.

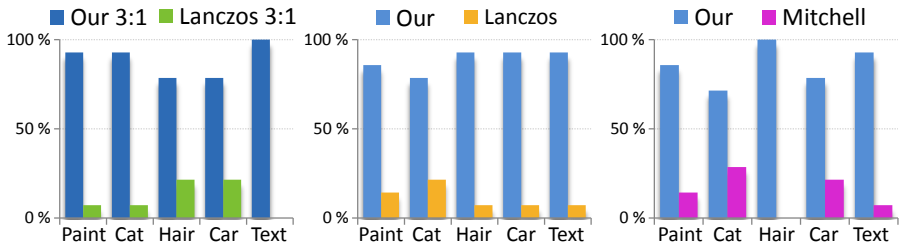


Figure 5.7: Our method against Lanczos in 3:1 scale (left) and original scale (middle) as well as Mitchell-Netravali (right).

We did not consider more naïve solutions like nearest-neighbor filtering, as their lower quality and objectionable flickering are readily visible. Subjects could toggle between two methods via the keyboard without any time limit. Subjects were asked to choose the method for which the reproduction of details is closest to the reference version. The results of this part of experiment are shown in Figure 5.7 (left). The second part of the study was similar to the first (Lanczos), but moving full-HD resolution images have been compared without any reference (Figure 5.7, middle). For this first experiment, the pixel size in the moving image was three times enlarged to match the scale of the reference image. All other experiments used the native resolution.

Next, we tested our method against Mitchell-Netravali [1988] filtering. The filter can be balanced between sharpening and smoothing using two parameters which makes it adequate for a large variety of images. We asked subjects to adjust the parameters to match their preferences with respect to the high-resolution image. Later, they were asked to compare their result with our technique, again by toggling between the methods (Figure 5.7, right).

Our technique performed better in terms of detail reconstruction, even when allowing filter parameter adjustments. During all experiments no flickering or temporal artifacts have been observed. A series of t-tests (Table 5.1) showed statistical difference in all cases with a significance level of .05.

Minimal Text

Encouraged by the outcome of the first experiment with the text stimuli we wanted to check what are the readibility limits in terms of font size. To this end we investigated horizontally moving text often used for TV news channels, as well as hand-held devices. To push our technique to the limits, we attempted to produce a 2×3 pixel sized font containing English capital letters. We created it by hand at a 6×9 resolution (Figure 5.8, but did not invest much time in optimizing the characters. We showed all the letters in random order to subjects asking for identification and compared our method to Lanczos filtering. The characters have been placed in chunks of five characters rather than isolated fonts to mimic a text document. We did not compare to static text because any attempts to produce such a small font were futile.

Although not perfect (Figure 5.9), the results indicate the quality-increase due to our apparent resolution enhancement. Performed series of t-tests showed significant difference between our font and standard downsampling for 13 out of 26 (Figure 5.9)

	Paint	Cat	Hair	Car	Text
Our vs. Lanczos (3:1 scale)					
$t(26)$	8.485	8.485	3.551	3.551	∞
p	$< .001$	$< .001$	.002	.002	$< .001$
Cohen's d	3.207	3.207	1.342	1.342	∞
Our vs. Lanczos					
$t(26)$	5.204	3.551	8.485	8.485	8.485
p	$< .001$	.002	$< .001$	$< .001$	$< .001$
Cohen's d	1.967	1.342	3.207	3.207	3.207
Our vs. Mitchell					
$t(26)$	5.204	2.419	∞	3.551	8.485
p	$< .001$	.023	$< .001$	.001	$< .001$
Cohen's d	1.967	0.914	∞	1.342	3.207

Table 5.1: High-resolution images experiment: The table contains t- and p-values as well as effect size (Cohen's d) for pairwise comparison of our method with respect to Lanczos and Mitchel filtering.

Figure 5.8: The 6×9 font used by our optimization procedure to be displayed in 2×3 raster.

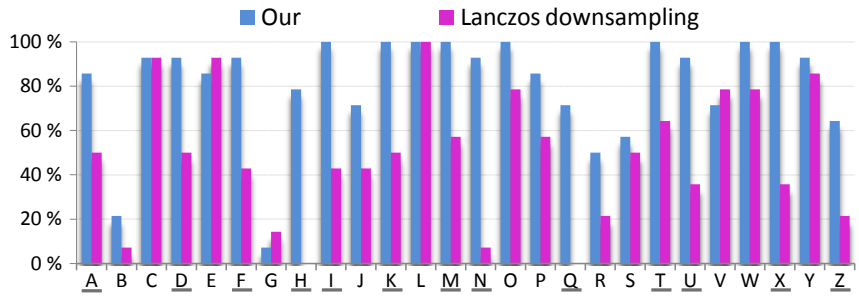


Figure 5.9: Character recognition: Standard filtering vs. Ours. Dark horizontal lines indicate significant statistical difference.

for a significance level of .05. The biggest improvement was, as expected, in the horizontal direction that coincides with the movement. *H, K, M, N, R, X* contain much horizontally oriented information, making them easy to read. On the other hand, the lack of improvement in vertical direction, affects letters such as: *B, G*.

3D Rendering

We also conducted a smaller study for 3D rendering applications. We estimated the eye tracking based on a derived motion flow. We assumed that the motion is piecewise linear for different image regions, and, thus, we can apply our technique locally. We

used a scene showing highly detailed hair and a 3D terrain in a fly-over view similar to Google Earth. Similar to the high-resolution image experiment, subjects could toggle between moving images for our method and respectively Lanczos and Mitchel filtering. All five subjects chose our solution over Lanczos and Mitchel for both scenes.

5.6 Discussion

Slightly moving images have become common practice of web-page designers that present scrolling photos, or scrolling text (e.g., news), and small animations. Besides guiding attention and looking more natural and lively (the Ken Burns effect), improved detail perception, as shown in our experiments, might explain this trend. Our experiments suggest that the strongest enhancement can be obtained using our technique, but even frame-wise downsampling (taking into account the current mapping to physical pixels) is a better strategy than naïve resampling of a downsampled image.

Proper filtering becomes even more important for large displays, as illustrated by our study, but big velocities imply the need for higher refresh rates to counteract the hold-type effect.

The optimization scheme delivers a high-quality result, but is computationally costly (e.g., double full-HD image 3840×2160 using three subimages is processed in approx. 18 minutes, standard full HD needed 5 min). However, our CPU-based optimization could be improved, especially using a GPU implementations. Our first experiments with a gradient-descent GPU solver (enforcing constraints in each iteration via clamping), showed that the computation time can be reduced to below 1 s. An efficient GPU-solver for the subimages generation was later described in [Templin et al., 2011].

Our model does not rely on any profound hardware-specific assumptions (e.g., not on the RGB subpixel layout) which makes our technique relatively immune to technological and perceptual differences. In our experiments with an 120 Hz CRT display as well as 60 Hz DLP and LCD projectors we have obtained a clear resolution enhancement. Therefore, we also expect that our technique works for OLED displays where very high frame rates should lead to an even stronger resolution enhancement. Essentially, our model conforms with the major goals of display manufacturers to reduce the visibility of RGB subpixel layout and screen door effect, which otherwise could ruin the impression of image integrity and continuity.

Motion is key to our approach because it ensures that the pixel grid projects to different locations on the retina, which we exploit in our approach. Consequently, there is a link between the motion direction and the apparent resolution increase, e.g., horizontal / vertical motion only enables horizontal / vertical improvements.

Our method relies on the efficiency of SPEM, which was shown to perform well even for more complex motions (Section 2.2.2) than those used here. Also switching between two objects, while tracking, is not a problem as saccadic movements are very fast. This makes our resolution enhancement feasible for more complex than panning motion.

5.7 Conclusions

Due to the limited spatial resolution of current displays, the depiction of very fine details is difficult. In this chapter, we proposed a novel reconstruction for moving images, that takes human perception into account to improve the detail reproduction. To this end, we vary pixel intensities rapidly over time and rely on temporal integration properties of the HVS. Our work is general in the sense that it extends to arbitrarily high frame rates. We discussed how finding the optimal temporal variation for a specific eye-movement, retinal temporal integration time, image resolution and display resolution allows us to virtually "address" apparent super-resolution pixels on a conventional-resolution display. Finally, we evaluated the improvement in terms of apparent details in a perceptual study. We presented various applications including improved photo details, panorama pop-up views, online rendering, and scrolling texts (where we pushed the limits by showing that a 2×3 pixel font can still be legible). In many cases, significant improvements can be achieved using our method. In other cases, no new artifacts – such as flickering – are introduced.

We further improved our techniques in Templin et al. [2011], showing that the apparent resolution enhancement can be applied to regular video sequences where SPEM is estimated using optical flow techniques. This allowed us to enhance resolution by exploiting the motion present in the scene. We also proposed there a fast GPU-based method for computing subimages. Recently, the concept of resolution improvement relying on temporal integration properties of the HVS was extended by Berthouzoz and Fatal [2012]. They achieved resolution enhancement by moving a display in a periodical manner (i. e., vibrating), instead of introducing motion. This allowed them to achieve a similar resolution gain but for a static content.

In the future, we want to exploit the display mosaics similarly to the ClearType fonts. Initial attempts have not led to a clear quality improvement and this issue requires further investigation. We demonstrated the applicability of our approach to offline rendering. In the future, we would like to opt for an online context, eventually combining our solution with an eye trackers to only locally perform the optimization computation. In a first attempt, we also tried to construct directional filters from the results of the optimization process, but (because our optimization is not a filtering) the values often exceed the dynamic range of the display. We can reduce the contrast via blending, but then the results are clearly inferior to the full optimization and show color aberrations. This remains an exciting avenue for future work. It would be also interesting to investigate newer display devices such as new Sony PlayStation 3D Display which offers 240 Hz framerate. Using such a screen would potentially improve results even further and reduce possible problems with flickering.

6

A Perceptual Disparity Model and Metric

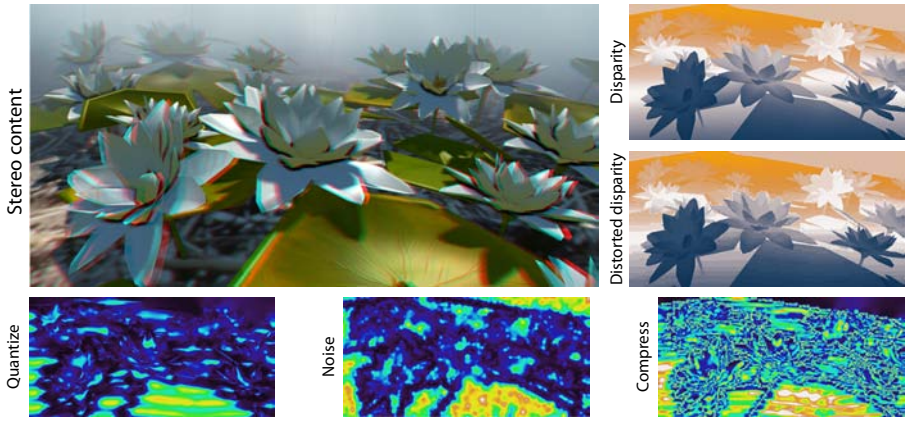


Figure 6.1: A metric derived from our model, that predicts the perceived difference (bottom) between original and distorted disparity (top right).

In the previous chapters we considered resolution enhancement as well as improving animation smoothness which were both achieved by means of skillful manipulation applied in temporal domain. Such techniques allow better adjustment of available footage to a display device characteristic and leads to a perceived quality improvement. Similar content manipulations have recently become crucial in the context of 3D stereo where quality of perceived footage does not only depend on the footage itself and available 3D equipment, but also on viewing conditions (Chapter 2). The complexity of depth perception makes the process of preparing good stereo content difficult. This comes mostly from the fact that the HVS uses many different cues [Palmer 1999; Howard and Rogers 2002] to estimate spatial objects configuration, which is essential for the scene understanding.

The interplay among different depth cues has challenged artist for many centuries, who tried to convey a believable depth impression on 2D surfaces [Livingstone 2002]. Also today, when 3D stereo gains significant attention, it has been identified as an important problem in contemporary computer graphics [Wanger, Ferwerda and Greenberg 1992; Matusik and Pfister 2004; Lang et al. 2010]. Current 3D display technology allows us to make use of binocular disparity, but often, a good stereo impression is

obtained through trials and errors. This, however, is usually a very tedious and time consuming task for artists. The biggest challenge in the stereo footage creation is limitation of depth range that can be reproduced using current 3D technology (Section 2.3.4). When an object presented in the scene violates depth limits (comfort zone [Shibata et al. 2011; Lambooij et al. 2009]), viewing discomfort can be experienced (Section 2.3). Therefore, in order to ensure comfort, minimize perceived distortions, but also to improve the 3D content creation process and allow artist to modify depth impression, recent years have produced many successful methods for disparity manipulation [Jones et al. 2001; Lang et al. 2010]. Whenever such modifications are applied, it is important to analyze their impact on disparity perception. Such prediction would lead to better control of the changes and could provide useful guidelines for better disparity/depth manipulation techniques.

In this part of the dissertation, we aim at providing computational models for better understanding and modeling the HVS disparity processing. There are many known and unknown high-level processes involved in stereo perception and although there have been many attempts to better understand how the HVS interprets depth and how sensitive it is to different cues [Cutting and Vishton 1995; Howard and Rogers 2002], no method for modeling the depth perception has been proposed so far. In our work, we will exclusively consider binocular disparity, a low-level, pre-attentive cue, attributed to the primary visual cortical areas [Howard and Rogers 2002, Chapter 6] as it is one of the most important and appealing stereo cues for short distances (up to 30 meters) [Cutting and Vishton 1995].

This chapter is organized as follows. In section Section 6.1, we give an overview of the techniques presented in this chapter . In Section 6.2, we introduce a perceptual disparity model together with its derivation. Later, in section Section 6.3, we show how such a model can be used to construct a disparity metric that is able to predict perceived differences between stereo images. In Section 6.4, we improve the disparity model by proposing a new one which accounts for influence of underlying luminance information on disparity perception, and show how this enhances performance of the disparity metric. This is followed by an additional discussion (Section 6.5) and conclusions (Section 6.6).

6.1 Overview

Inspired by the existing similarities between brightness and depth perception described in Section 2.3 as well as work on luminance perception (Section 3.3.1), in this chapter, we show how a perceptual model for disparity can be developed. Our technique allows for better understanding of binocular disparity perception and prediction of the HVS response to various disparity stimuli. The key information needed to build such a model are measurements of disparity detection thresholds. Although there were a couple of psychophysical measurements performed to acquire the sensitivity of the HVS to disparities [Tyler 1975; Bradshaw and Rogers 1999; Howard and Rogers 2002], none of them allowed for building a model as complete as those available for luminance [Daly 1993; Lubin 1995]. Therefore, in our work, we conducted a number of psychovisual experiments to obtain the thresholds for sinusoidal patterns in depth with different frequencies and amplitudes. This allowed us to fit an analytic function to the obtained data, which describes detection thresholds for a whole range of possible sinusoidal

depth corrugations. Such a function can be later used for constructing a model based on so-called *transducer functions*, which map disparity values to a perceptually linearized space. The transducer functions are invertible, therefore changing from physical disparity values into a perceptual space and back becomes possible.

In existing disparity processing algorithms the influence of RGB image content on depth perception has been usually ignored. Intuitively, a certain magnitude of luminance contrast is required to make disparity visible, while stereopsis is likely to be weaker for low-contrast and blurry patterns. In this chapter, we also show that luminance contrast does have a significant impact on depth perception and should be taken into account for a more faithful computational model. Therefore, we further improve our disparity model including luminance, which allows for more accurate prediction of the HVS response. One key challenge of a combined luminance-disparity model is the growing dimensionality, which we limit to 4D by considering frequency and magnitude of disparity, as well as frequency and contrast of luminance. Similarly as for the previous model, we conducted a psycho-visual experiment that provides necessary data to construct such a model.

Since the here-presented disparity model can predict a response of the HVS to disparity patterns, it can be used for computing perceived differences between stereo images. In this chapter, we also show how such a disparity metric can be constructed using our model and how taking into account luminance pattern influences the prediction of differences.

6.2 Disparity Model

In this section we describe a model that predicts the HVS response to a disparity signal. This model accounts only for disparity signals so its prediction is an upper-bound on the factual response of the HVS, as a perfect luminance pattern is assumed.

Our disparity model is based on transducer functions which were introduced earlier for luminance [Wilson 1980; Mantiuk, Myszkowski and Seidel 2006]. They allow for mapping disparity values to a perceptually linearized space as well as mapping back to original disparity space. To derive those functions, we need precise detection and discrimination thresholds that cover the full range of magnitudes and spatial frequencies of corrugated patterns that can be seen without causing diplopia. Therefore, we first describe an experiment for obtaining such data and then show how it can be used to build the disparity model.

6.2.1 Measurements

While some disparity detection data is readily available [Bradshaw and Rogers 1999; Tyler 1975] (see [Howard and Rogers 2002, Chapter 19.6.3] for a survey), we are not aware of any set of densely measured discrimination thresholds. The closest experiment to ours has been performed by Ioannou et al. [1993] where observers matched corrugations of various spatial frequencies to a variable amplitude-reference corrugation of fixed intermediate frequency. Only three suprathreshold amplitudes (up to 8 arcmin) have been investigated [Howard and Rogers 2002, Fig. 19.24 d], and we are more interested in the disparity-difference discrimination within the same

frequency to account for intra-channel masking. Also, for the disparity detection task, different stereoacuity has been reported ranging from 2–6 arcsec [Bradshaw and Rogers 1999, Figure 1] up to 30 arcsec, which accordingly to Coutant et al. [1993] is a more representative value for most individuals. Furthermore, existing measurements are often performed with sophisticated optical setups (e. g., [Blakemore 1970]), they require participants to fixate on points or bars, sometimes for only a short time, whereas we want to acquire data for modern, inexpensive 3D displays, which are also used in our applications (Chapter 7).

In our experiments, we allow for free eye motion , making multiple fixations on different scene regions possible, which approaches real 3D-image observations. In particular, we want to account for a better performance in relative depth estimation for objects that are widely spread in the image plane (see [Howard and Rogers 2002, Chapter 19.9.1] for a survey on possible explanations of this observation for free eye movements). The latter is important to comprehend complex 3D images. In our experiments, we assume that depth corrugated stimuli lie at the zero disparity plane (i. e., observers fixate corrugation) because free eye fixation can mostly compensate for any pedestal disparity within the range of comfortable binocular vision [Lambooij et al. 2009; Hoffman et al. 2008]. Such zero-pedestal disparity assumption guarantees that we conservatively measure the maximum disparity sensitivity [Blakemore 1970], which in such conditions is similar for uncrossed (positive, i. e., $\omega - \theta > 0$ as in Figure 2.9) and crossed (negative) disparities [Howard and Rogers 2002, Fig. 19.24 c]. For this reason in what follows we assume that only disparity magnitude matters in the transducer derivation.

Parameters

Our experiments measure the dependence of perceived disparity on two stereo image parameters: disparity magnitude and disparity frequency. We do not account for variations in accommodation, viewing distance, screen size, luminance, or color and all images are static.

Disparity Frequency specifies the spatial disparity change per unit visual degree. Note, that it is different from the frequencies of the underlying luminance, which we will call luminance frequencies. We considered the following disparity frequencies: 0.05, 0.1, 0.3, 1.0, 2.0, 3.0 cpd. In the pilot study, we experimented with more extreme frequencies, but the findings proved less reliable (consistent with [Bradshaw and Rogers 1999]).

Disparity Magnitude corresponds to the corrugation pattern amplitude. The range of disparity magnitude for the detection thresholds to suprathreshold values that do not cause diplopia have been considered, which we determined in the pilot study for all considered disparity frequencies. While disparity differences over the diplopia limit can still be perceived up to the maximum disparity [Tyler 1975], the disparity discrimination even slightly below the diplopia limit is too uncomfortable to be pursued with naïve subjects. Therefore, the maximum disparity magnitude that we consider in our experiment is decreased explicitly, in some cases, significantly below this boundary. After all, we assume that our data will be mostly used in applications within the disparity

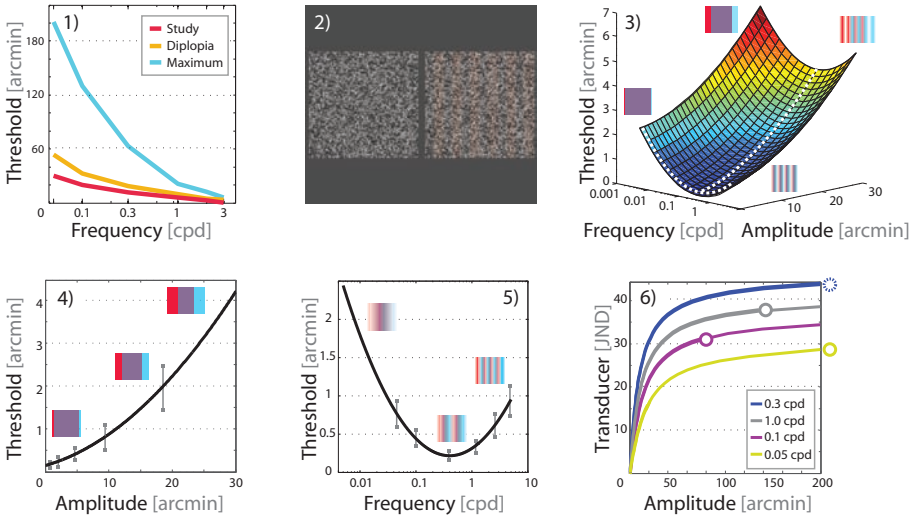


Figure 6.2: (1) Disparity magnitude ranges: (red) maximum disparity used in our experiments, (yellow) diplopia and (blue) maximum disparity limits. (2) The experimental setup where subjects select the sinusoidal gratings which exhibits more depth. (3) Our fit to the disparity discrimination threshold function $\Delta d(\mathbf{s})$. (4) The cross section of our fit at the most sensitive disparity frequency 0.3 cpd (the error bars denote the standard error of the mean (SEM) at measurement locations). (5) Analogous cross section along frequency axis showing the detection thresholds. Both cross sections are marked with white dashed lines in (3). (6) The transducer functions for selected frequencies. Empty circles denote the maximum disparity limits.

range that is comfortable for viewing. Figure 6.2.1 shows our measured diplopia and maximum disparity limits, as well as the effective range disparity magnitudes that we consider in our experiments.

Stimuli

All stimuli are horizontal sinusoidal gratings with a certain amplitude and frequency with a random phase. Similarly to existing experiments, the disparity is applied to a luminance pattern consisting of a high number of random dots, minimizing the effect of most external cues (e. g., shading). A cue that could influence our measurements is texture density. However in our case, as we seek to measure 1 JND, subjects always compare patterns with very similar amplitudes. Therefore the difference in texture density between two stimuli is always imperceivable and does not influence detection thresholds as confirmed by Bradshaw et al. [1999]. Formally, we parameterize a stimulus $\mathbf{s} \in \mathbb{R}^2$ in two dimensions (amplitude and frequency). The measured *discrimination threshold function* $\Delta d(\mathbf{s}) : \mathcal{S} \rightarrow \mathbb{R}$ maps every stimulus within the considered parameter range to the smallest perceivable disparity change.

In order to generate stimuli, an image-based *warping* is used to produce both views of the stimulus independently. First, the stimulus' disparity map D is converted into a pixel disparity map D_p, by taking into account the equipment, viewer distance, and screen size. We assumed standard intra-ocular distance of 65 mm, which is needed for

conversion to a normalized pixel disparity over subjects. Next, the luminance image is traversed and every pixel $L(\mathbf{x})$ from location $\mathbf{x} \in \mathbb{R}^2$ is warped to a new location $\mathbf{x} \pm (D_\mathrm{p}(\mathbf{x}),0)^T$ for the left, respectively right eye. As occlusions cannot occur for these stimuli, warping produces artifact-free valid stimuli. To ensure sufficient quality, super-sampling is used: Views are produced at 4000^2 pixels, but shown as 1000^2-pixel patches, down-sampled using a 4^2 Lanczos filter.

Equipment

We use three representative forms of stereo equipment (refer to [Onural et al. 2006] for a 3D display technology survey): active shutter glasses, anaglyph glasses and an auto-stereoscopic display. We used Nvidia 3D Vision active shutter glasses ($\sim$ \$100) in combination with a 120 Hz, 58 cm diagonal Samsung SyncMaster 2233RZ display ($\sim$ \$300, 1680×1050 pixels), observed from 60 cm. As a low-end solution, we also used this setup with anaglyph glasses. Further, a 62 cm Alioscopy 3DHD24 auto-stereoscopic screen ($\sim$ \$6000, 1920×1080 pixels total, distributed on eight views of which we used two) was employed. It is designed for an observation distance of 140 cm. Unless otherwise stated, the results are reported for active shutter glasses.

Subjects

All subjects in our experiment are naïve, paid, and have normal or corrected-to-normal vision. We verified that no subject was color [Ishihara 1987] or stereo-blind [Richards 1971].

Task

In this experiment, we sample Δd at locations $S = \{\mathbf{s}_i | \mathbf{s}_i \in \mathcal{S}\}$ by running a discrimination threshold procedure on each to evaluate $\Delta d(\mathbf{s}_i)$. A two-alternative forced-choice (2AFC) staircase procedure is performed for every $\mathbf{s}_i$. This technique is called the PEST (Parameter Estimation by Sequential Testing) procedure [Taylor and Creelman 1967] and was also used for luminance by Bradley et al. [1986]. Each staircase step presents two stimuli: one defined by $\mathbf{s}_i$, the other as $\mathbf{s}_i + (\varepsilon,0)^T$, which corresponds to a change of disparity magnitude. Both stimuli are placed either right or left on the screen (Figure 6.2.2), always randomized. The subject is then asked which stimulus exhibits more depth amplitude and to press the "left" cursor key if this property applies to the left otherwise the "right" cursor key. After three correct answers ε is decremented and after a single incorrect answer it is incremented by the step-size determined via PEST (Parameter Estimation by Sequential Testing) [Taylor and Creelman 1967].

In total 27 PEST procedures have been performed per subject. Twelve subjects participated in the study with the shutter glasses and four subjects with each other setup of stereo equipment (anaglyph and auto-stereoscopy). Each subject completed the experiment in 3–4 sessions of 20–40 minutes. Four subjects repeated the experiment twice for different stereo equipment.

6.2.2 Model

We use the data from the above procedure to determine a model of perceived disparity by fitting an analytic function to the recorded samples. It is used to derive a transducer to predict perceived disparity in JND (just noticeable difference) units for a given stimulus which is the basis of our stereo difference metric (Section 6.3).

Fitting

To model the thresholds from the experiment, we fit a two-dimensional function of amplitude a and frequency f to the data (Figure 6.2.3–5). We use quadratic polynomials with a log-space frequency axis to well fit (the goodness of fit $R^2 = 0.9718$) the almost quadratic "u"-shape measured previously [Bradshaw and Rogers 1999, Fig. 1]. Figure 6.3 and 6.4 summarize the obtained data for each type of the equipment in our discrimination threshold experiments. For each set of data we fit the discrimination threshold function, which is denoted as d_s, d_{ag}, d_{as} for shutter glasses, anaglyph and auto-stereoscopic display respectively:

$$\Delta d_s(f,a) = 0.2978 + 0.0508\,a + 0.5047\,\log_{10}(f) + 0.002987\,a^2 + 0.002588\,a\log_{10}(f) + 0.6456\,\log_{10}^2(f).$$

$$\Delta d_{ag}(f,a) = 0.3304 + 0.01961\,a + 0.315\,\log_{10}(f) + 0.004217\,a^2 - 0.008761\,a\log_{10}(f) + 0.6319\,\log_{10}^2(f).$$

$$\Delta d_{as}(f,a) = 0.4223 + 0.007576\,a + 0.5593\,\log_{10}(f) + 0.0005623\,a^2 - 0.03742\,a\log_{10}(f) + 0.7114\,\log_{10}^2(f).$$

For all devices the minimum disparity sensitivity was found for ~0.4 cpd, which agrees with previous studies [Bradshaw and Rogers 1999]. Our results indicate that the disparity sensitivity near the detection threshold and for low disparity magnitudes is the highest for the shutter glasses (Figure 6.4) . For larger disparity magnitudes the differences in the sensitivity are less pronounced between different stereo technologies and overall the shape of discrimination threshold functions is similar regardless the equipment.

Measurements for auto-stereoscopic display revealed large differences with respect to shutter and anaglyph glasses. This, we think, is due to much bigger discomfort, which was reported by our subjects. Also measurements for such displays are more challenging due to difficulties in low spacial frequency reproduction, which is caused by relatively big viewing distance (140 cm) that needs to be kept by a observer. The disparity sensitivity drops significantly when less than two corrugations cycles are observed due to lack of spatial integration [Howard and Rogers 2002], which might be a problem in this case. We observed that measurements for disparity corrugations of low spacial frequencies are not as consistent as for higher frequencies and they differer among subjects. Surprisingly, our experiments seem to indicate that for larger disparity magnitudes the disparity sensitivity is higher for the auto-stereoscopic display than

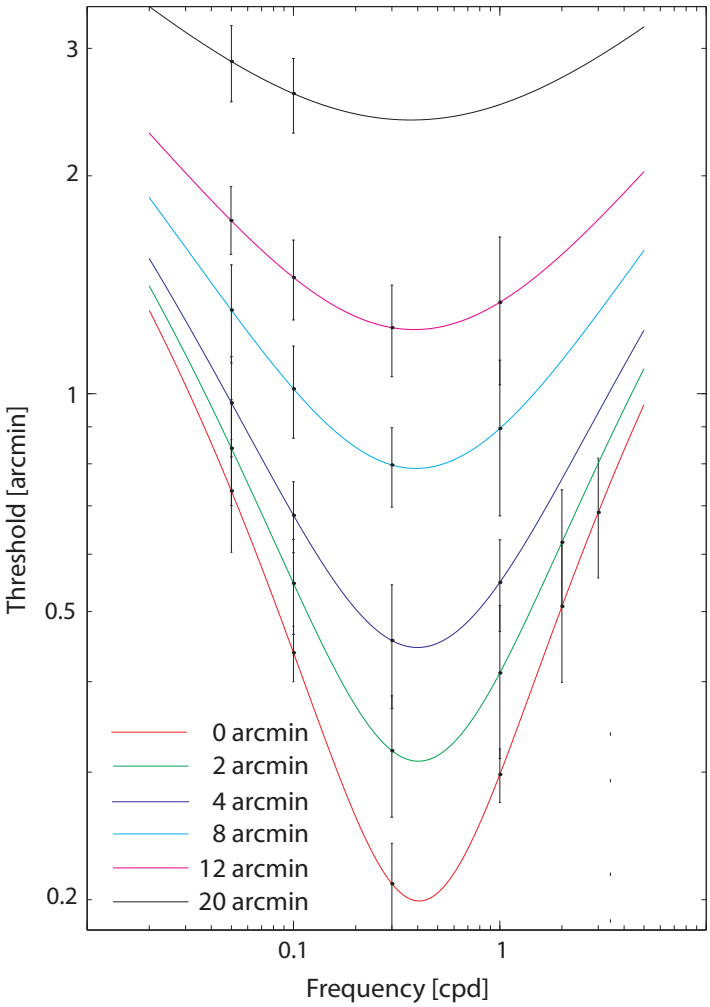


Figure 6.3: Shutter glasses: Disparity detection and discrimination thresholds as a function of the spatial frequency of disparity corrugations for different corrugation amplitudes as specified in the legend. Points drawn on curves indicate the measurement samples. The error bars denote the standard error of the mean (SEM).

for other stereo technologies we investigated. In future work it would be interesting to consider bigger auto-stereoscopic displays, which would cover larger field of view, allowing for discrimination thresholds measurements for lower frequencies which are still important for the disparity perception. In this work, due to financial limitations we did not used bigger screen.

Transducers and Inverse Transducers

Based on the obtained threshold functions, we derive a set of transducer functions which map a physical quantity x (here disparity) into the sensory response r in JND

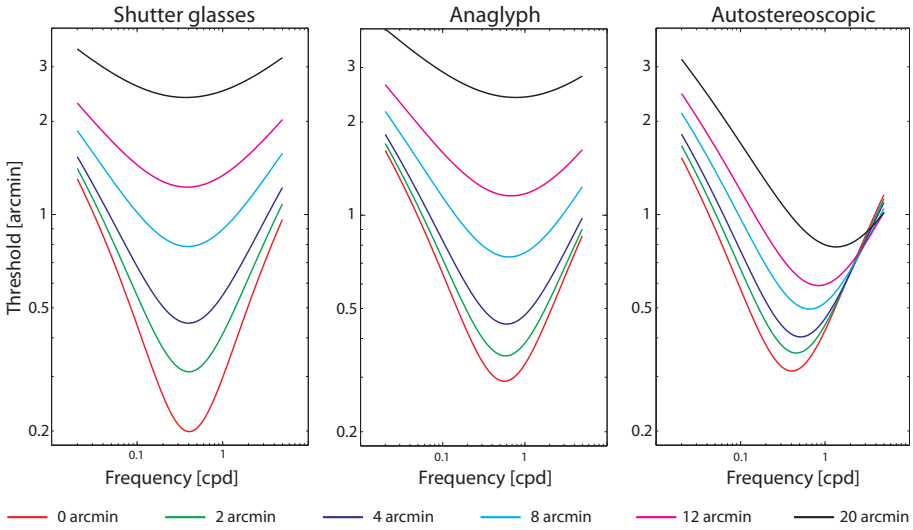


Figure 6.4: Comparison of disparity detection and discrimination thresholds for three different stereo devices.

units. Each transducer $t_f(x) : \mathbb{R}^+ \to \mathbb{R}^+$ corresponds to a single frequency f and is computed as $t_f(x) = \int_0^x (\Delta d(a,f))^{-1} \mathrm{d}a$, where Δd is one of the fitted functions. As all of them are positive, $t_f(x)$ is monotonic and can be inverted, leading to an *inverse* transducer $t_f^{-1}(r)$, that maps a number of JNDs back to a disparity. Again, an inverse transducer depends on the frequency f. For more details on transducer derivation refer to Wilson [1980] or Mantiuk et al. [2006].

One should notice that limiting disparity magnitudes below the diplopia limits in our experiments (Section 6.2.1) has consequences. Our $\Delta d(\mathbf{s})$ fit is, strictly seen, only valid for this measured range. Consequently, transducers (Figure 6.2.6) have to rely on extrapolated information beyond this range. While the transducer functions look plausible, they should actually remain flat beyond the maximum disparity limits, which are denoted as empty circles in Figure 6.2.6. In those regions we enforce that the overall increase of the transducers remains below a one-JND fraction, reflecting that depth perception becomes impossible, but securing the inconvertibility of the function.

In practice, we rely on a family of transducers T_f discretized using numerical integration and inverse transducers T_f^{-1} found by inversion via searching. All transducers are pre-computed (Figure 6.2.6) and stored as look-up tables.

Pipeline

The transducers of the previous section can be integrated in a pipeline to compute perceived disparity of a stimulus (Figure 6.5). This pipeline takes a stereo image, defined by luminance and pixel disparity, as input and outputs the perceived disparity decomposed into a spatial-frequency hierarchy that models disparity channels in the HVS. Such spatial-frequency selectivity is usually modeled using a hiearchal filter bank with band-pass properties such as wavelets, Gabor filters, Cortex Transform [Watson 1987; Daly 1993], or Laplacian decomposition [Burt and Adelson 1983].

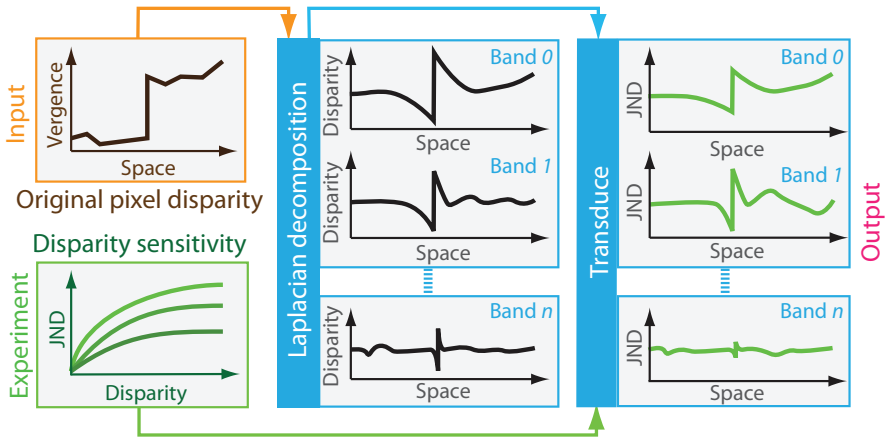


Figure 6.5: Our perceived disparity model pipeline: Starting from angular vergence derived for pixel disparity (top left, orange), a Laplacian decomposition separates disparity in different frequency bands. The transducers acquired from our experiments (bottom left, green) are used to transform disparity into perceptual units (JND).

The latter is our choice, mostly for efficiency reasons and the fact that the particular choice of commonly used filter banks should not affect qualitatively the quality metric outcome [Winkler 2005, p. 90].

First, the pixel disparity is transformed into corresponding angular vergence, taking the 3D image observation conditions into account. Next, a Gaussian pyramid is computed from the vergence image. Finally, the differences of every two neighboring pyramid levels are computed, which results in the actual disparity frequency band decomposition.

In practice, we use a standard Laplacian pyramid with 1-octave spacing between frequency bands. Finally, for every pixel value in every band, the transducer of this band maps the corresponding disparity to JND units by a simple lookup. In this way, we linearize the perceived disparity.

To convert perceived disparity e. g., after a manipulation (see applications - Chapter 7), back into a stereo image, an inverse pipeline is required. Given a pyramid of perceived disparity in JND, the inverse pipeline produces again a disparity image by combining all bands similarly to previous work on luminance [Mantiuk, Myszkowski and Seidel 2006] and applying inverse transducers.

6.3 Metric

Based on our model, we can define a perceptual stereo image metric. Given two stereo images, one original D^{o} and one with distorted pixel disparities D^{d}, it predicts the spatially varying magnitude of perceived disparity differences.

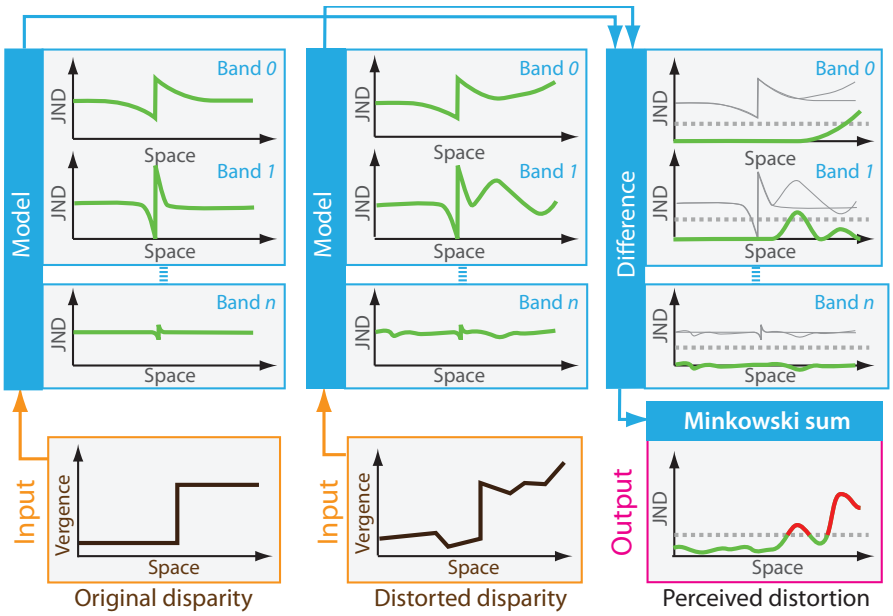


Figure 6.6: Perceptual disparity image difference metric: First, an original and a distorted pixel disparity map (bottom left) are transformed to vergence. Next, we convert them into JND using our pipeline (top left). Subtracting both JND results we obtain a per-band spatially varying perceived disparity difference (top right). Finally, Minkowski summation combines all bands into a single distortion map scaled in JNDs (bottom right).

6.3.1 Perceived Disparity Difference

In order to compute perceived disparity difference between two disparity maps D^{o} and D^{d}, we insert both of them into our pipeline (Figure 6.6). First, we compute the perceived disparity R^{o}, respectively R^{d}. This is achieved using our original pipeline from Figure 6.5 with an additional phase uncertainty step (also called the *phase independence operation* in [Lubin 1995]) before applying per-band transducers. This eliminates zero crossings at the signal's edges and thus prevents incorrect predictions of zero disparity differences at such locations. In practice, we use a 5×5 Gaussian low-pass filter at every level of our Laplacian pyramid and compensate for the resulting amplitude loss, which is a part of the calibration procedure (below). Than every pixel i, j and each band k the difference $R^{\mathrm{o,d}}_{i,j,k} = R^{\mathrm{o}}_{i,j,k} - R^{\mathrm{d}}_{i,j,k}$ is computed and finally combined using a Minkowski summation [Lubin 1995]: $d_{i,j} = \left(\sum_k \left| R^{\mathrm{o,d}}_{i,j,k} \right|^{\beta} \right)^{\frac{1}{\beta}}$, where β, found in the calibration step, controls how different bands contribute to the final result. The result is a spatially-varying map depicting the magnitude of perceived disparity differences, which can be visualized, e. g., in false colors, as in Figure 6.1 (right).

In our metric, we consider all frequency bands up to 4 cpd, which cover the full range of visible disparity corrugation frequencies and we ignore higher-frequency bands. Note that the intra-channel disparity masking is modeled because of the compressive

nature of the transducers for increasing disparity magnitudes. The band decomposition enables us to model frequency selectivity of such masking, where disparity signals from different channels are ignored.

6.3.2 Calibration

We performed the metric calibration to compensate for accumulated inaccuracies of our model. The most serious problem is signal leaking between bands during the Laplacian decomposition, which offers also clear advantages. Such leaking effectively causes inter-channel masking, which conforms with the observation that the disparity channel bandwidth of 2–3 octaves might be a viable option [Howard and Rogers 2002, Chapter 19.6.3d]. This justifies relaxing frequency separation between 1-octave channels such as we do. While decompositions with better frequency separation between bands exist such as the Cortex Transform, they preclude an interactive metric response. Since signal leaking between bands as well as the previously-described phase uncertainty step lead to an effective reduction of amplitude, a corrective multiplier K is applied to the result of the Laplacian decomposition.

In order to find good K and calibrate our metric, we used the data obtained in our experiment (Section 6.2.1) and all the experiment stimuli used in measurements. As distorted images, we considered the corresponding patterns with 1, 3, 5, and 10 JNDs distortions. The magnitude of 1 JND distortion directly resulted from the experiment outcome and the magnitudes of larger distortions are obtained using our transducer functions. The correction coefficient $K = 3.9$ lead to the best fit and an average metric error of 11%. Similarly, we found the power term $\beta = 4$ in the Minkowski summation.

6.3.3 Validation

In order to show that the response predicted by our model correlates and agrees with what can be observed, we evaluated our metric in a few of steps. First, we tested for the need of having different transducers for different bands. This is best seen when considering the difference between two *Campbell-Robson* disparity patterns of different amplitude (Figure 6.7). Comparing our metric and a metric, where the same transducer for all bands is used, shows that ours correctly takes into account how the disparity sensitivity depends on the pattern frequency. Our method correctly reports the biggest difference in terms of JNDs for frequencies to which the HVS is most sensitive to (i. e., $\sim$0.4 cpd). Using only one transducer is still beneficial comparing to not using it, which in such a case would result in an uniform distortion reported by the metric.

Next, we checked whether subthreshold distortions as predicted by our metric cannot be seen, and conversely whether over threshold distortions identified by our metric are visible. We prepared three versions of each stimulus (Figure 6.8): a reference, and two copies with a linearly scaled disparity which our metric identifies as 0.5 JND and 2 JND distortions.

In a 2AFC experiment, the reference and distorted stereo images were shown and subjects were asked to indicate the image with larger perceived depth. Five subjects took part in the experiment where stimuli have been displayed 10 times each in a randomized order. For the 0.5 JND distortion the percentage of correct answers falls

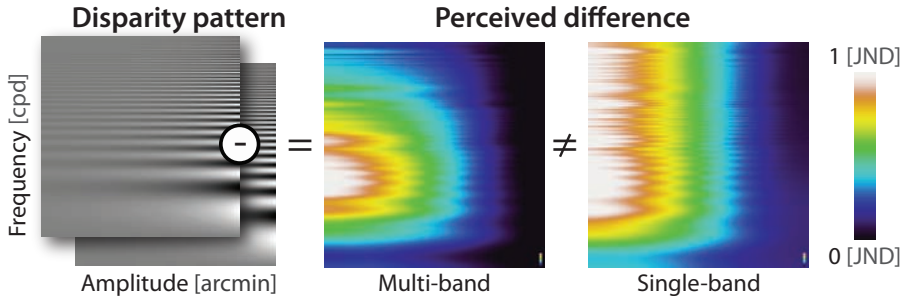


Figure 6.7: A comparison of perceived difference between the Campbell-Robson disparity pattern and the same pattern after adding a constant increment of amplitude (left), once using one transducer per band (multi-band, center) vs. the same transducer for all bands (single-band, right).

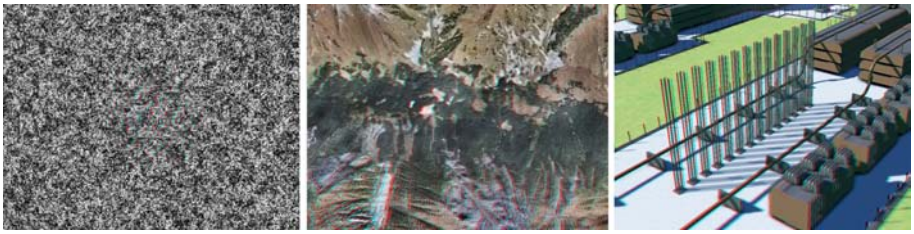


Figure 6.8: Left to right: Stimuli of increasing complexity and increasing amount of external cues shown in red-cyan anaglyph: a Gabor patch, a 3D Terrain, and a Factory.

into the range 47–54%, which in practice means a random choice and indicates that the distorted image cannot be distinguished from the reference. For the 2 JND distortion the outcome of correct answers was as follows: 89%, 90%, and 66% for the scenes Gabor, Terrain, and Factory, respectively. The two first results fall in the typical probability range expected for 2 JND [Lubin 1995] (the PEST procedure asymptotes are set at the level 79%, equivalent to 1 JND [Taylor and Creelman 1967]). On the other hand, for Factory the metric overestimates distortions, reporting 2 JND, while they are hardly perceivable. The repeated experiment for this scene with 5 JND distortion lead to an acceptable 95% of correct detection. The results indicate that our metric correctly scales disparity distortions when disparity is one of the most dominating depth cues. For scenes with greater variety of depth cues (e. g., occlusions, perspective, shading), perceived disparity is suppressed and our metric can be too sensitive. The t-test analysis indicates that the distinction between 0.5 and 2 JND stimuli is statistically significant with p-value below 0.001 for the Gabor and Terrain scenes. For Factory such statistically significant distinction is obtained only between 2 and 5 JND stimuli.

6.4 Disparity/Luminance Model

The model presented above accounts only for variations in disparity pattern ignoring underlying luminance signal, which, as shown earlier in section Section 2.3.3, can significantly reduce sensitivity of the HVS to disparity and lead to degradation of depth

impression. Here, we address this issue and present an improved disparity model for predicting the HVS response to a disparity signal in presence of a supporting luminance pattern. Some parts of this model as well as its derivation are similar to those presented before. Therefore, in our description, we concentrate only on parts where those two models differ.

6.4.1 Threshold Function

Similarly to the previous disparity model discussed in Section 6.4, the first step in deriving the current model is to acquire a threshold function. Here, however, as we seek covering the influence of both disparity and luminance signal on depth perception, instead of two-parameters function, i. e., disparity frequency and amplitude, we need to consider four-parameters function $th(f_\mathrm{d}, m_\mathrm{d}, f_\mathrm{l}, c_\mathrm{l})$. This function for each combination of its parameter values (disparity frequency f_d and disparity magnitude m_d, luminance frequency f_l, and luminance contrast c_l,) defines the smallest perceivable change (i.e., equivalent to 1 JND) in disparity magnitude (expressed in arcmins units).

The huge problem in acquiring such data necessary to model the function is its dimensionality. However, as indicated by Legge and Gu [1989], only low-level luminance contrast affects stereoacuity, while otherwise having little to no influence. Further, Cormack [1991], presented a corresponding disparity-threshold function for luminance contrast that is expressed in units of threshold multiples. Consequently, we decided to factor out the luminance contrast dimension in order to reduce dimensionality problem, leading to the following model:

$$th(f_\mathrm{d}, m_\mathrm{d}, f_\mathrm{l}, c_\mathrm{l}) = s(f_\mathrm{d}, m_\mathrm{d}, f_\mathrm{l}) / Q(f_\mathrm{l}, c_\mathrm{l}), \tag{6.1}$$

where s is a discrimination-threshold function assuming maximal contrast and Q is a function that compensates for the increase of the threshold due to a smaller luminance contrast c_l.

Similarly to our previous model, we express s using a general quadratic polynomial function:

$$\begin{aligned} s(f_\mathrm{d}, m_\mathrm{d}, f_\mathrm{l}) &= p_1\, log_{10}^2(f_\mathrm{d}) + p_2\, m_\mathrm{d}^2 + p_3\, log_{10}^2(f_\mathrm{l}) \\ &+ p_4\, log_{10}(f_\mathrm{d})\, m_\mathrm{d} + p_5\, log_{10}(f_\mathrm{d})\, log_{10}(f_\mathrm{l}) + p_6\, m_\mathrm{d}\, log_{10}(f_\mathrm{l}) \\ &+ p_7\, log_{10}(f_\mathrm{d}) + p_8\, m_\mathrm{d} + p_9\, log_{10}(f_\mathrm{l}) + p_{10}, \end{aligned} \tag{6.2}$$

where $\mathbf{p}$ is a parameter vector obtained by minimizing the following error:

$$\arg\min_{p \in \mathbb{R}^{10}} \sum_{i=1}^{n} \left((s(o_i) - \Delta m_i) / (\Delta m_i) \right)^2,$$

where o_i are stimuli with their corresponding thresholds Δm_i, as determined in our psychophysical experiment (Section 6.4.5). Hereby, we obtain $\mathbf{p} = [\,0.3655, 0.0024, 0.2571, 0.0416, -0.0694, -0.0126, 0.0764, 0.0669, -0.3325, 0.2826]$ (Figure 6.9). The use of the log domain is motivated from our previous model and lead to better results. The range of magnitudes of disparity detection thresholds specified by our model is in a good agreement with data in [Lee and Rogers 1997] for measured there mid-range of disparity and luminance spatial frequencies. For more extreme ranges, similar to [Hess, Kingdom and Ziegler 1999], we observe that, for low-frequency

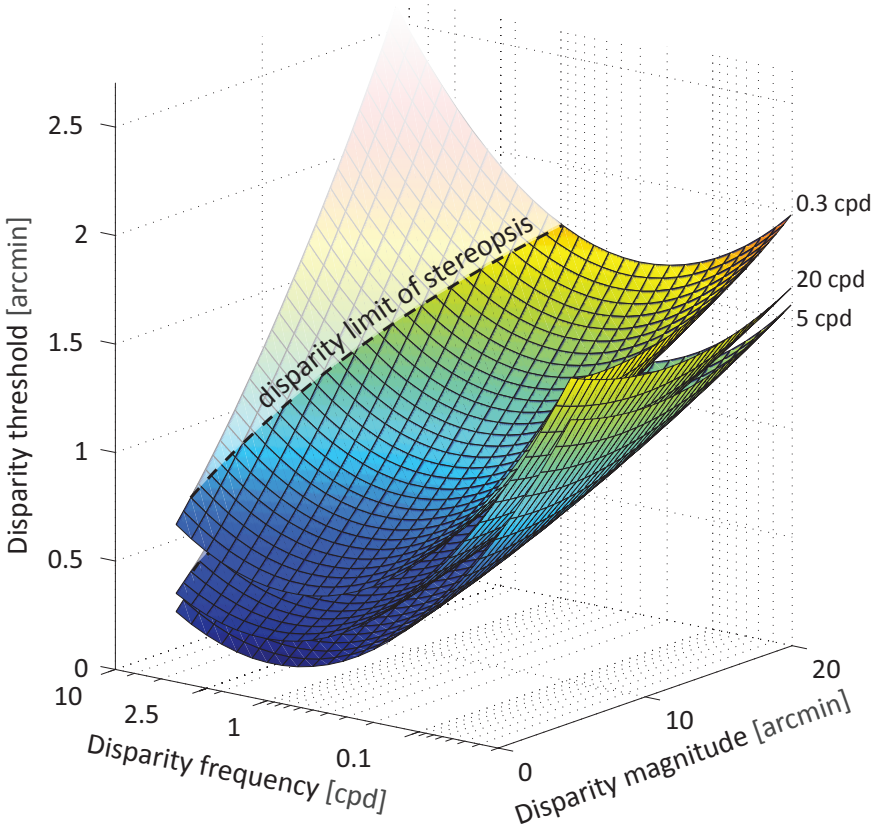


Figure 6.9: Plot visualizing slices of our model of the disparity discrimination function for sinusoidal corrugations. We illustrate three surfaces corresponding to different luminance frequencies (0.3 cpd, 5 cpd and 20 cpd) and a well visible contrast (above 10 JNDs). The model is limited by the disparity limit of stereopsis [Tyler 1975].

disparity corrugations, a wide range of luminance frequencies lead to good stereoacuity, while for higher-frequency disparity corrugations stereoacuity is weak for low luminance frequencies.

To determine the scaling function Q, we use the data by Cormack [1991], expressed in units of threshold multiples c_m, to which we fit a cubic polynomial in the logarithmic domain:

$$T(c_m) = exp(r_1\, log_{10}^3(c_m) + r_2\, log_{10}^2(c_m) + r_3\, log_{10}(c_m) + r_4), \tag{6.3}$$

where $\mathbf{r} = [\,-0.9468, 4.4094, -6.9054, 4.7294]$ is a parameter vector obtained from fitting above model to the experimental data of Cormack [1991]. Q is then expressed as:

$$Q(f_1, c_1) = \begin{cases} T(c_1 \cdot csf(f_1))/T(u) & \text{if } c_1 \cdot csf(f_1) \le u \\ 1 & \text{if } c_1 \cdot csf(f_1) > u \end{cases}, \tag{6.4}$$

where csf is the contrast sensitivity function proposed by Barten [1989] and u specifies the level at which luminance contrast has no further influence on the disparity threshold [Legge and Gu 1989], hence, $T'(u) = 0$ implying $u = 35.6769$. Our fit is illustrated in Figure 6.10.

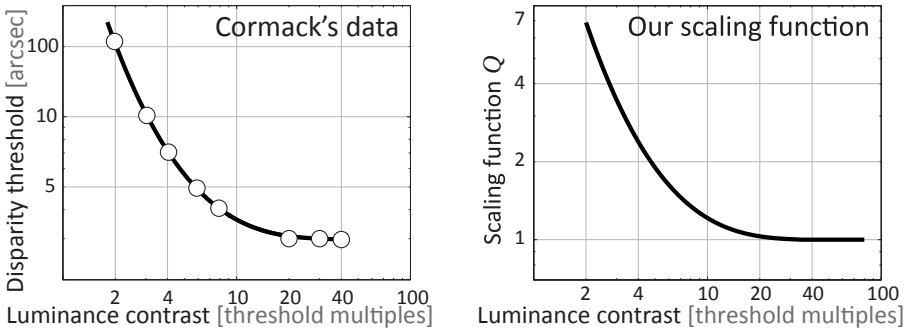


Figure 6.10: Our function fitting to Cormack's data (marked by empty circles), as well as our scaling function Q.

6.4.2 Transducer

Analogously to the previous model, the one presented here is also based on a transducer function. We build this function, relating physically-measurable quantities to the HVS response (in JND units), directly from the threshold function th. Previously, we assumed a perfectly-visible luminance pattern and proposed a two-dimensional transducer of disparity frequency and magnitude, which leads to a conservative prediction. Consequently, perceived disparity is generally overestimated. Here, we extend this solution to a four-dimensional transducer $t(f_\mathrm{d}, m_\mathrm{d}, f_\mathrm{l}, c_\mathrm{l})$, where the additional parameters f_l and c_l stand for luminance frequency and luminance contrast, respectively:

$$t(f_\mathrm{d}, m_\mathrm{d}, f_\mathrm{l}, c_\mathrm{l}) = \int_0^{m_\mathrm{d}} th(f_\mathrm{d}, x, f_\mathrm{l}, c_\mathrm{l})^{-1} \mathrm{d}x \tag{6.5}$$

The function $t(f_\mathrm{d}, \cdot, f_\mathrm{l}, c_\mathrm{l}) : \mathbb{R} \to \mathbb{R}$ (a partial application of t to $f_\mathrm{d}, f_\mathrm{l}, c_\mathrm{l}$) is monotonic, hence, there usually[1] exists an inverse transducer $(t(f_\mathrm{d}, \cdot, f_\mathrm{l}, c_\mathrm{l}))^{-1}$. t maps disparity-luminance stimuli to a perceptually linear space of disparity and t^{-1} can be used to reconstruct the stimuli. E.g., for disparity compression, mapping via t makes removing imperceptible disparities easy and t^{-1} can be used to reconstruct the modified disparity map. Similarly, we can build a transducer to convert luminance contrast to a uniform space.For more details on constructing transducer functions please refer to work by Wilson [1980] and Mantiuk et al. [2006].

In practice, a transducer function t can be evaluated by numerical integration and stored in a table and t^{-1} can be implicitly defined via a binary search. Nonetheless, in four dimensions, the memory and performance cost can be significant. A better solution makes use of the factorization: $t(f_\mathrm{d}, m_\mathrm{d}, f_\mathrm{l}, c_\mathrm{l}) = t'(f_\mathrm{d}, m_\mathrm{d}, f_\mathrm{l}) / Q(f_\mathrm{l}, c_\mathrm{l})$, where $t'(f_\mathrm{d}, m_\mathrm{d}, f_\mathrm{l}) = \int_0^{m_\mathrm{d}} s(f_\mathrm{d}, x, f_\mathrm{l})^{-1} \mathrm{d}x$. Functions t' (and t'^{-1} if wanted) can be discretized, precomputed, and conveniently stored as 3D arrays. The inverse transducer for a given $f_\mathrm{d}, f_\mathrm{l}, c_\mathrm{l}$ is then: $m_\mathrm{d} = t'^{-1}(f_\mathrm{d}, Q(f_\mathrm{l}, c_\mathrm{l}) \cdot R, f_\mathrm{l})$, where R is a JND-unit response to disparity.

Similarly as it was done for our previous disparity model, in order to account for the HVS limits of perceivable stereopsis, we use our threshold function only within

[1] only for constant luminance patterns, the function cannot be inversed

the limits measured by Tyler et al. [1975] (Figure 6.9). Beyond this range, transducer functions should remain flat. We enforce this by keeping the overall increase of transducer function below one-JND fraction. This accounts for the loss of the disparity perception maintaining at the same time the invertibility of the model.

6.4.3 Computational Model

The above transducer is valid for abstract stimuli. For real content, we decompose the input's luminance and disparity into corresponding Laplacian pyramids, such as it has been done independently for luminance [Mantiuk, Myszkowski and Seidel 2006] and disparity in Section 6.2.2.

For luminance, we compute a Laplacian pyramid C of the luminance pattern, which contains Michelson contrast values (which are required for Q in Eq. (6.4)). For disparity, we first compute vergence values from pixel disparity maps and then build a Laplacian pyramid D [Burt and Adelson 1983]. The value $D_i(\mathbf{x})$ corresponds to the disparity value at location $\mathbf{x} \in \mathbb{R}^2$ in the i-th level frequency of the pyramid i. e., $\alpha/2^i$ cpd (where $\alpha \approx 20$ for our setup). To convert disparities in JND units, we apply the transducer function (Eq. 6.5) to the values of the Laplacian pyramid. $f_\mathrm{d} = \alpha/2^i$ and $m_\mathrm{d} = D_i(\mathbf{x})$. To evaluate the transducer, we also need to know the frequency f_l and contrast c_l of the supporting luminance pattern.

To combine luminance and disparity, we follow the independent-channels hypothesis for disparity (Section 2.3.3); stereoacuity is determined by the most sensitive channel and remains uninfluenced by other channels. Consequently, given a disparity frequency f_d, we assume that the response is the **maximum** of all responses for all higher-luminance frequencies f_l in the image region corresponding to half a cycle of f_d.

Formally, the response is then:

$$D'_i(\mathbf{x}) = \max_{j \in (0,\ldots,i-1)} t(\alpha/2^i, D_i(\mathbf{x}), \alpha/2^j, S_j(\mathbf{x})), \tag{6.6}$$

where $S_j(\mathbf{x})$ evaluates the luminance support, defined as the average of all contrast values C_j of the j-th level of the luminance decomposition that fall into a rectangular region $\sigma_i(\mathbf{x}) = (\mathbf{x} - (w,w)^\mathsf{T}, \mathbf{x} + (w,w)^\mathsf{T})$ of size $w = 2^i$ around $\mathbf{x}$ (Figure 6.11). The values of S can be pre-computed from C and later accessed in constant time. The resulting structure is a Laplacian pyramid with a MIP map defined on each of its levels, as visualized in Figure 6.11, right. Note, that computing the maximum of S_j over all levels and, then, applying a transducer independently is *not* equivalent.

6.4.4 Asymmetries

So far, our computational model does not account for the asymmetries described in Section 2.3.3, as it would be necessary to study an even higher-dimensional space including neighborhood configurations. Nonetheless, we can exploit a few observations to derive a perceptually-motivated model that we verify practically (Section 7.7).

Although, in Section 2.3.3, we provide a couple of possible explanations for the asymmetry effect in disparity, all of them are based on pictorial depth cues and none

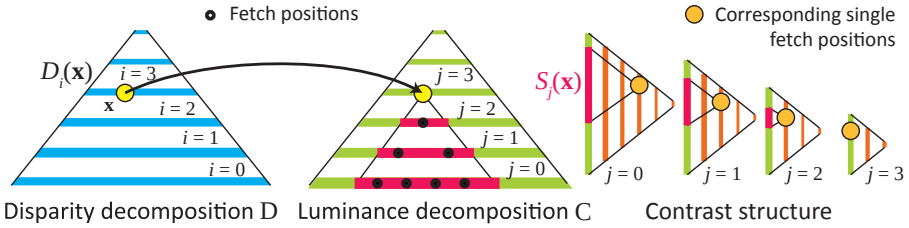


Figure 6.11: For a disparity $D_i(\mathbf{x})$ at location $\mathbf{x}$, the model needs to involve levels $j < i$ in the luminance Laplacian pyramid C. In each level j, an average contrast $S_j(\mathbf{x})$ of a region $\sigma_i(\mathbf{x})$ (marked in red) around $\mathbf{x}$ is computed and its impact evaluated. For acceleration, averages can be pre-computed in MIP maps for each level *(right)*.

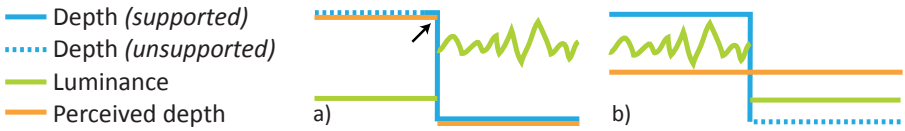


Figure 6.12: Luminance patterns *(green)* influence the depth perception *(orange)* of the same depth profile *(blue)*. Some allow us to well discriminate depth (solid blue), while others do not (dotted blue). The frontal-patch edge can benefit from the luminance contrast between patches (a, *arrow*). If the luminance pattern of the deeper patch renders localization difficult the depth step disappears (b).

of them is convincing as the arguments contradict each other. For the purpose of our model, we develop new interpretation of the asymmetry effect in Figure 6.12, which is based on the fact that the sensitivity to pictorial cues such as texture density or relative size is much lower than sensitivity to binocular disparity in the considered depth range [Cutting and Vishton 1995]. Through this, we would like to argue that this effect comes mostly from the binocular disparity cue. This can be observed using anaglyph glasses in Figure 2.15. Based on our interpretation of this effect we explain here how it can be handled in our disparity model.

In fact, in order to perceive a sinusoidal depth corrugation, peaks as well as valleys of the sinusoid need to be well supported by luminance contrast. However, as illustrated in Figure 6.13, this is not always the case. To account for the full wave, a 3 $\times$ 3 neighborhood at the given level of the Laplacian decomposition is evaluated and the minimum response chosen. Hereby, we ensure that a full cycle is well supported and visible.

While this extension already explains several cases in Figure 2.15, it is insufficient to explain the entire asymmetry. The texture swap would not yet be detected to influence depth perception. In order to better model the response, we need to take disocclusion into account. In fact, the occluding patch's edge introduces a high-contrast luminance edge in superposition with the patch beneath. If they are present in both views (left and right eye), these high frequencies allow us to localize the edge in space - we disregard pathological cases where disparity and luminance frequency perfectly agree. Consequently, we propose to evaluate the luminance contrast for both views and use the maximum response. Hereby, a point on the deeper patch will be disoccluded

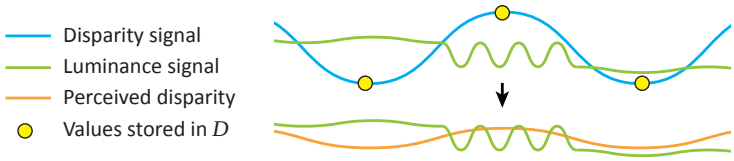


Figure 6.13: Disparity response attenuation due to weak support in perceived luminance contrast. The physical contrast in the valleys of sinusoidal depth function is of low spatial frequency, for which the HVS is less sensitive, and of too small magnitude to make the depth corrugation visible.

in one view and reveal its high-frequency luminance neighborhood, and points on the edge will maintain a high-contrast edge in both views. Also note, that this effect affects not only the points directly on the edge, but also in a small neighborhood near the edge. This relates to findings on backward-compatible stereo Section 2.3.2. Similarly to the Cornsweet effect for luminance, the HVS extrapolates depth information to neighboring locations. Although heuristic, this solution performs well in practice (Section 7.7).

6.4.5 Measurements

To derive the parameters of s (Eq. 6.2), our experiment explores: disparity frequency f_d (measured in cpd), disparity magnitude m_d (measured in arcmins), and luminance frequency f_l (measured in cpd). The construction of the experiment is similar to the one performed for the disparity model in Section 6.2.1. Viewing conditions as well as procedure of selecting subjects were the same. The biggest difference is the fact that we needed to consider different luminance patterns as we want to measure the influence of luminance on disparity perception. We also restricted current experiment to shutter-glasses-based screen that was used earlier. Here, we describe only stimuli and task that was used, and for other details (i. e., equipment and participants) please refer to Section 6.2.1.

Stimuli

All stimuli are horizontal sinusoidal disparity corrugations with luminance noise of a certain frequency. First, we create a luminance pattern by producing a noise of frequency f_l and scale it to match the maximal reproducible contrast on our display. Using such a texture excludes any external depth cues, such as shading. Next, we create a disparity pattern; a sinusoidal grating with frequency f_d and magnitude m_d. Such disparity gratings do not produce occlusions. Finally, the luminance pattern is warped according to the disparity pattern to produce an image pair for the observer's left and right eye similarly as it was done in previous experiment. All steps are adjusted to the viewing conditions, i. e., the screen size and viewing distance. We assume standard intra-ocular distance of 65 mm.

Task

In this experiment, we seek measuring a disparity-discrimination threshold for stimuli defined in three-dimensional parameter space. For a given stimulus $o = (f_\mathrm{d}, m_\mathrm{d}, f_\mathrm{l})$, we run a threshold estimation procedure. In each step, we show two stimuli o and $o + [0, \Delta m_\mathrm{d}, 0]$. One located on the left-hand side of the screen and the other on the right. The position is randomized. The task of the participant is to judge which stimulus exhibits larger depth magnitude and choose via the "left" or "right" arrow keys. Depending on the answer, Δa is adjusted in the next step using the QUEST procedure [Watson and Pelli 1983]. When the standard deviation of the estimated value is lower than 0.05, the process stops. We decided to use QUEST instead of PEST, which was used in previous model, as the first one turns out to converge much faster which was crucial in this case as the dimensionality of the problem is bigger then previously.

In total there were 24 participants who took part in the experiment (12 women and 12 men). They were all between 22 and 30 years old. One participant was discarded due to very high thresholds (on average 3 times higher than thresholds of others). Each participant performed 35 adjustment procedures. One session took from 30 to 100 min. Subjects were allowed to take a break whenever the felt tired. In total, we obtained 805 measured thresholds to which we fit our model.

6.4.6 Improved Response Prediction and Metric

Including information about luminance pattern to our disparity model allows to successfully detect the human inability to perceive changes of disparity when the luminance support is not adequate due to, for example low luminance contrast because of fog or depth-of-field. A comparison of the HVS response predicted by the model that takes into account luminance information and the disparity-only model is shown in Figure 6.14.

The new model, similarly to the one that does not account for luminance information, can be used to predict the perceived difference between two stereo images: a reference image and a second image which underwent a distortion, such as compression. The construction of this metric is similar to the construction of the previous one Section 6.3. We first use our model to map both input images into our perceptually-linear space. The transducer function is applied after the *phase uncertainty operation*. Per-band differences indicate the detectability of disparity changes, computed by a simple subtraction. All bands can be combined using Minkowski summation to produce a spatially-varying difference map. We use the same parameters as those that were obtained in the process of calibrating the disparity-only metric for both—phase uncertainty and Minkowski summation. The difference between the metric that uses the new luminance-contrast aware disparity model and the one that does not account for luminance pattern is shown in Figure 6.15. Previous metric is too conservative and report differences which are invisible due to weak luminance signal (false positives).

In order to evaluate the new model, we conducted an additional user study with 17 participants. We wanted to verify how well our metric predicts actual JND values. We used a stereo image from Figure 7.10 and applied a scaling to the disparity in order to create images that differ in depth perception. One image was modified to

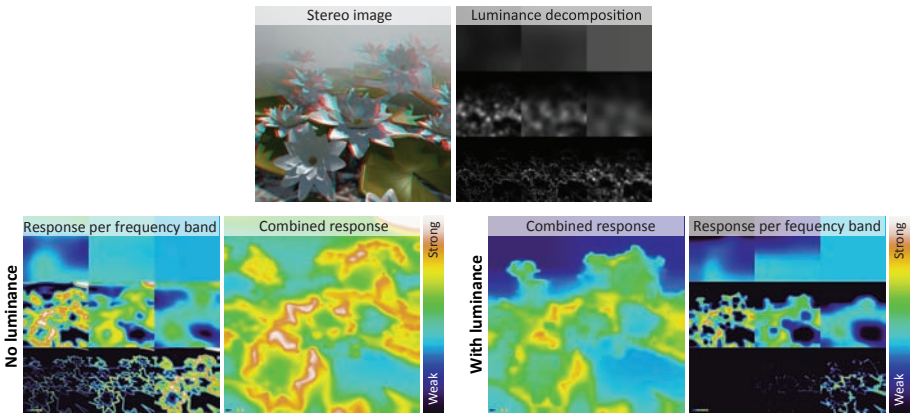


Figure 6.14: Comparison of perceived disparity as predicted by the previous model that ignores image content (left) and the new model that accounts for underlying luminance pattern(right). Responses per frequency band and the combined response are shown for both, as well as the original stereo image with the multi-band decomposition of the luminance pattern (middle).

match an average error of 0.5 JND (with minimum 0.4 JND and maximum 0.8 JND). For a second image the average difference was 3 JND (with minimum 2.5 JND and maximum 3.5 JND). We showed the modified images side by side (randomized) with the original image and asked about perceived differences. Each pair was shown ten times in randomized order. The 0.5 JND difference image was detected in 58 % cases, which is close to a random answer, as expected. For the 3 JND case the probability of the detection was 91 %.

6.5 Discussion

Previous experiments concerning depth discrimination thresholds [Julesz 1971; Blakemore 1970; Anstis, Howard and Rogers 1978; Poggio and Poggio 1984; Coutant and Westheimer 1993; Lee and Rogers 1997; Prince and Rogers 1998; Bradshaw and Rogers 1999; Hess, Kingdom and Ziegler 1999; Nishina 2003; Sato 2004] covered small ranges of parameters values (i. e., disparity frequency/amplitude and luminance-contrast magnitude/frequency). Usually, also parameters interdependence was ignored. In our models, we considered bigger ranges of parameters as well as their interdependence. Further, previous findings were based on mutually very different setups and viewing conditions e. g., they require participants to fixate points or bars, sometimes for only a short time. Our thresholds are mostly higher than what is reported for physical stimuli in the literature but we focused on current off-the-shelf stereo equipment. The difference implies that there is still room for improvement of modern equipment, but also that it is worth deriving thresholds for existing hardware explicitly.

Our disparity models are based on a couple of simplifying assumptions. We do not consider temporal effects as those described by Lee et al. [2007] although they are not only limited to high-level cues, but also present in low-level pre-attentive structures [Palmer 1999; Howard and Rogers 2002]. It would require adding additional

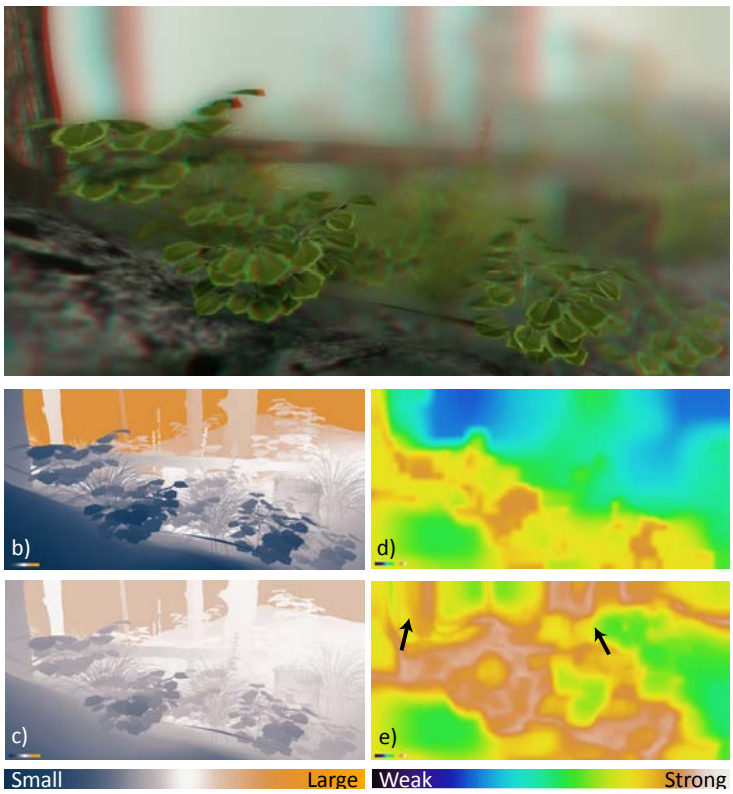


Figure 6.15: When stereo content *(luminance, a; disparity, b)* is manipulated *(disparity, c)* we quantify the perceived change considering luminance and disparity *(d)*. Ignoring luminance *(e)* produces wrong predictions e. g., for low-texture areas, fog, or depth-of-field *(arrows)*.

dimensionality to our experimental data, and we relegate such an extension as future work. Furthermore, our measurements are performed under the assumption that subjects accommodate onto the screen. This is valid for current equipment, but might not hold in the future. Our measurements consider only horizontal corrugations, while the stereoscopic anisotropy (lower sensitivity to vertical corrugations) can be observed for spatial corrugations below 0.9 cpd [Bradshaw and Rogers 1999], but our metric could easily accommodate for anisotropy by adding orientation selectivity into our channel decomposition [Daly 1993; Lubin 1995].

We do not include the influance of chromatic stereopsis as it is less contrast sensitive, leads to weaker stereoacuity, and feature a more-limited disparity range with respect to its luminance-contrast counterpart [Kingdom and Simmons 2000]. We also do not consider image brightness because stereoacuity weakly depends on luminance in mesopic and photopic levels (over 0.1 cd/m^2), which are typical for standard stereo 3D displays [Howard and Rogers 2002, Chapters 19.5.1]. To reduce dimensionality, we decided to exclude the influence of luminance-contrast magnitude from our measurements for the second model; stereo increment thresholds per luminance spatial frequency channel actually increase for low contrast as a power-law function [Rohaly and Wilson 1999, Fig. 6]. We considered this influence in a simplified form by ex-

pressing the signal in each luminance channel in JND units including its normalization via the CSF function. We then compute stereoacuity per channel using a compressive function (Eq. 6.4), which we derived based on the data from [Cormack, Stevenson and Schor 1991].

In our models we do not incorporate other depth cues than binocular disparity. For example we do not include the influence of color, whereas it is known for centuries [Livingstone 2002] how e. g., aerial perspective (the haze effect) greatly helps the depiction of space. As for most luminance perception models and metrics, higher-level processing is beyond the scope of this dissertation. A perceptual model that includes an analysis of the shape and its properties (e. g., its curvature, moments, etc.) would be an exciting avenue of future research.

Concerning the generality of our models, for the second model which accounts for luminance pattern we did not repeat the experiment for different display technologies (e. g., anaglyph, polarization), which may result in a slightly different stereoacuity. However, measurements with different equipment are not a problem and our model and techniques remain valid. For displays with different parameters (e. g., size, resolution, contrast ratio), both our models are directly applicable; they use physical values which can be computed from the display specification and viewing conditions. Furthermore, in Chapter 7 we present a number of techniques that use our model and are evaluated on a different group of people than the threshold measurements. The positive results of the study suggest that, although stereoacuity varies among people, our models are general enough to be successfully used in practice.

Please also note that our metrics measure perceived disparity differences, which is different from viewing comfort or immersion in the environment which are important problems when dealing with stereo. However, an automated computational model of perceived disparity like ours could be a critical component when developing dedicated algorithms. Similarly, the prediction of disparity distortions is merely one of many factors which contributes to the perceived realism of a 3D scene, image quality itself as well as the visual comfort (e. g., eye strain) [Meesters, IJsselsteijn and Seuntiens 2004] are further interesting aspects.

Finally, our models and metrics, once acquired, are easy to implement and efficient to compute, allowing a GPU implementation which was used to generate all results presented in this dissertation at interactive frame rates.

6.6 Conclusions

We identified the interdependence of disparity magnitude and spatial frequency in a consistent set of stimuli using a psycho-visual experiment. By fitting a model to the acquired data, we derived metrics that were shown to perform the challenging task of predicting human disparity perception.

A user study on the impact of luminance stimuli on disparity perception allowed us to derive a new disparity-sensitivity function, which enabled us to construct a model that captures and models the interaction between disparity and luminance. To our knowledge, this model is the first of its kind. We also explained how to integrate certain neighborhood-related effects, such as asymmetry. In the next chapter (Chapter 7), we show how powerful our disparity models are in the context of disparity manipulations.

We present there a number of applications where the models either improve results or allow for completely new edits.

In future work, one could consider temporal effects and higher-level cues (shading, texture, bas-relief ambiguity, etc.) that would complement our approach. The effects of conflicting stimuli (accommodation, image content, etc.), currently, remain mostly unclear. We believe that models such as ours will be crucial for stereo images and video processing.

7

Perceptually Driven Disparity Manipulations

While in the past only an anaglyph stereo was accessible on the consumer-level market, today, we find a variety of techniques to produce stereo effects ranging from polarization or shutter glasses to autostereoscopic displays [Onural et al. 2006; Matusik and Pfister 2004]. This trend is underlined by the increasing amount of stereo content in the form of TV broadcasts, feature films, and computer games. Although the quality of the stereo equipment is constantly improving, the reproducible depth range is smaller than what is observable in the real world (Section 2.3.4). Ignoring this limitation can result in viewing discomfort or even loss of stereo impression when the left and right images can no longer be fused.

Furthermore, the viewing conditions in which the content is observed do not necessarily correspond to conditions for which the content is prepared. For example, the distance between the virtual cameras might not correspond to the actual eye distance of the observer. Similarly, one might have made assumptions concerning the distance to the screen, or even the type of screen itself, which can substantially differ from later viewing conditions. Especially for movies, where stereo equipment, observers and their position are unknown such differences in viewing conditions can easily lead to perceivable distortions of the stereo content (Section 3.3.3).

Therefore, although the general creation of image pairs is scene- and artist-dependent, limitations in stereo content reproduction, as those presented above, make the process much more complex than producing regular 2D material [Mendiburu 2009]. Often, to assure viewing comfort, reduce distortions coming from different viewing conditions, or allow artistic adjustment, the stereo content needs to be manipulated (Section 3.3.3). In many situations, such modifications can be performed during the production step, however, sometimes they need to be customized once the viewing conditions are known. This motivates researchers to develop 3D stereo content manipulation techniques that are easy, intuitive and can be applied automatically.

The disparity models as well as the disparity metrics (Chapter 6) were shown to be powerful tools for analyzing stereo content manipulations. In this chapter, we show that they also enable designing new disparity manipulation techniques which allow for taking into account human perception. This, as demonstrated in our user studies, can significantly improve previous methods. The knowledge on human perception, especially human abilities in discriminating depth differences, is also interesting for compression applications. We show that using our models we can improve disparity compression without impairing visual quality of e. g., broadcast footage.

7.1 Overview

In this chapter, we show a number of techniques which utilize previously proposed disparity models and metrics (Chapter 6). We show advantages of using them at all stages of the stereo content post-processing.The methods presented here include disparity manipulations in the perceptual space, disparity optimization, disparity compression, apparent stereo manipulations as well as hybrid images. In this chapter we also describe a number of user studies which validate results of our techniques.

Most of the methods proposed here take an advantage of the perceptual space introduced in the previous chapter (Section 6.2.2). By performing manipulation in this space, we directly introduce changes to a predicted response of the HVS to disparity patterns, which automatically accounts for disparity sensitivity function of the HVS (Section 2.3.2). This makes the performed edits more meaningful. The advantage of using similar methods has been already shown for luminance manipulation [Mantiuk, Myszkowski and Seidel 2006]. In order to perform such edits, we first transform pixel disparity to the perceptual space using the pipeline depicted in Figure 6.5. After applying desired operations, we use the inverse pipeline (Section 6.2.2) to transform the modified response of the HVS back to pixel disparity, which is later used for resynthesizing the left and right view so the output stereo image reflects the manipulations. To perform the latter step we use our technique presented in Chapter 8. Besides the perceptual space, some of our applications (e. g., disparity optimization) utilize disparity metrics (Sections 6.3 and 6.4.6). This allows for stereo-content editing which minimizes visible distortions.

The rest of the chapter is organized as follows. First, in Section 7.2 we present simple operations in the perceptual space that can be applied either for artistic purposes or to fit the disparity range of stereo images into the comfort zone. Next, in Section 7.3, we present a more sophisticated method, where disparity is optimized using our disparity metrics, minimizing perceived distortions. Since one of our disparity metrics can account for underlying luminance pattern, we can also optimize content for multi-view autostereoscopic displays where a blur needs to be introduced in order to avoid inter-view aliasing. Minimizing distortions is also crucial for content compression. In Section 7.4, we present our disparity compression method, which can reduce disparity storage information without perceivable loss of quality. In Section 6.2.2, we showed that stereo perception depends on the viewer as well as the display device. Using our models we can perform so-called *personalization* (Section 7.5), which adjusts stereo content so it appears similarly to different observers regardless of the equipment. Often, in order to fit scene disparities within the comfort range, the scene needs to be flattened. In Section 7.6, we present a technique that can enhance stereo impression without expanding disparity range by exploiting the Cornsweet Illusion. The same method allows us to compute backward-compatible stereo images which appear almost ordinary when observed without stereo equipment, but convey a stereo impression if the equipment is used. Our luminance-disparity model predicts how depth perception is affected by underlying luminance pattern. This is utilized in Section 7.7, where we demonstrate how stereo impression can be improved when only luminance pattern is manipulated. In Section 7.8, we present a technique, which, using information about the HVS sensitivity to different disparity corrugation frequencies, computes stereo images that depict different stereo content depending on the viewing distances. This is similar to hybrid images for luminance proposed by Oliva et al. [2006]. In Section 7.9,

Figure 7.1: From left to right: Starting from an original depth map a pixel disparity map is computed and then a disparity pyramid is built. After multi-resolution disparity processing, the dynamic range of disparity is adjusted and the resulting enhanced disparity map is produced. The map is then used to create enhanced stereo image.

we conclude providing directions for future work.

In all user studies presented here, viewing conditions as well as the procedure of selecting subjects were the same as for the measurements performed in Section 6.2.1. We only restricted the experiments to the shutter-glasses-based screen that was used earlier. In this chapter, we describe only stimuli and tasks, and for other details (i. e., equipment and participants) please refer to Section 6.2.1.

7.2 Disparity Manipulation in Perceptual Space

Global operators [Pratt 1991] that map disparity values to new disparity values globally, can operate in our perceptually uniform space, and their perceived effect can be predicted using our metric. To this end, disparity is first perceptually linearized, i. e., converted into a perceptually uniform space via our disparity model (Section 6.2), then modified, and converted back. Below, we describe more precisely how it is done in practice.

7.2.1 Pipeline

An overview of our approach for manipulating disparity in the perceptual space is shown in Figure 7.1. As input of our algorithm we use a linearized depth buffer along with corresponding RGB color image. Based on this depth information, we derive, as an output, a disparity map that defines the stereo effect.

To compute the disparity map, we first convert the linearized depth into pixel disparity based on a mapping between scene-centric to viewer-centric model. The pixel disparity is converted then to the perceptually uniform space (Section 6.2.2), which also provides a decomposition into different frequency bands. Our methods will act on these bands to yield the output pixel disparity map which defines the enhanced stereo image pair. Given the new disparity map, we can then warp the color image according to this definition. Our approach is orthogonal to the particular technique used for warping. Here, we use our method described in Chapter 8. Below, we present two global operations that can be performed using such an approach.

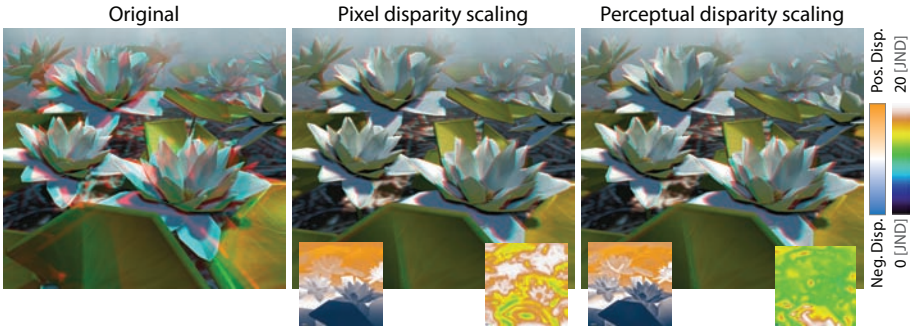


Figure 7.2: Disparity re-scaling preformed in our perceptually-uniform space preserves important disparities that can be lost when global scaling is applied. Our scaling compresses big disparities more, as our sensitivity in such regions is small, and preserves small disparities where the sensitivity is higher. In the lower insets, pixel disparities (left) and the difference to the original (right), as predicted by our metric, are shown. Simple scaling of pixel disparity results in loss of small disparities, flattening objects as correctly indicated by our metric in the flower regions. Our perceptual scaling preserves detailed disparity resulting in smaller and more uniformly distributed differences, again correctly detected by our metric.

7.2.2 Non-linear disparity-retargeting

Non-linear disparity-retargeting allows us to match pixel disparity in 3D content to specific viewing conditions and hardware, and provides artistic control [Lang et al. 2010]. The original technique uses a non-linear mapping of pixel disparity, whereas with our model, one can work directly in a perceptual uniform disparity space, making editing more predictable. Furthermore, our difference metric can be used to quantify and spatially localize the effect of a retargeting (Figure 7.2).

7.2.3 Histogram equalization

Histogram equalization can use our model to adjust pixel disparity to optimally fit into the perceived range [Pratt 1991; Mantiuk, Myszkowski and Seidel 2006]. Again, after transforming into our space, the inverse cumulative distribution function $c^{-1}(y)$ is built on the absolute value of the perceived disparity in all levels of the Laplacian pyramid and sampled at the same resolution. Then, every pixel value y in each level, at its original resolution is mapped to $\mathrm{sgn}(y)c^{-1}(y)$, which preserves the sign.

7.3 Disparity Optimization

One of our new applications is perceptual disparity optimization, which automatically fits the disparity of stereo content into a limited range by analyzing disparity via our disparity model. The objective is to achieve a small difference between the original and the re-mapped content according to our disparity metrics presented in Chapter 6. Due to many non-linearities of human disparity-luminance perception the optimization

is challenging and the search space of all possible disparity re-mappings is difficult to tackle.

7.3.1 General Case

To make the problem tractable, we restrict the search space to the subset of all global and piecewise-defined mapping curves, as done for automatized tone mapping [Mantiuk, Daly and Kerofsky 2008] (Figure 7.3). Such curves can be defined using a small number of n (we use $n = 7$) control points with values at fixed locations $P := \{(0, y_0), \ldots, (1.0, y_n)\}$ combined with a simple (e. g., piecewise-cubic) reconstruction. Given the original stereo content A and a remapping $r(A, P)$ of A using the control

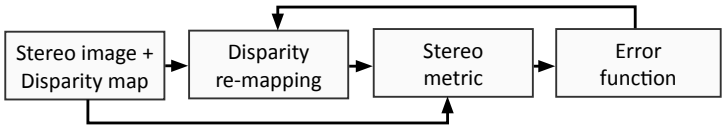


Figure 7.3: Our disparity optimization. From left to right: Input is a stereo image and a disparity map. A disparity mapping P is applied to the input. Our metric computes the difference between input and remapped content. The difference is converted into a single error value, and a new mapping P is chosen. The process is repeated until the error is low enough or a fixed iteration number is reached.

points P, simulated annealing is used to minimize the integrated perceived difference over the image domain Ω

$$\min_{P \in \mathbb{R}^n} \int_{\Omega} A \ominus r(A, P) \mathrm{d}\mathbf{x},$$

where the $\ominus$ operator denotes our perceptual metric of disparity difference. By implementing our method on a GPU, the disparity optimization can be performed at interactive speeds e. g., while a user navigates inside the scene (Figure 7.4). In order to maintain temporal coherence, we use the last frame's solution as the initial guess for P in the next frame. We can further smoothly interpolate previous solutions over a couple of frames to improve the smoothness of the animation. A similar approach was recently used in [Oskam et al. 2011].

7.3.2 Multi-view Autostereoscopic Display

Disparity optimization is particularly important for multi-view auto-stereoscopic displays, where the affordable disparity range is very shallow. Beyond this range depth-of-field is usually applied in order to avoid interperspective aliasing as described by Zwicker et al. [2006]. Therefore, two extreme strategies (Figure 7.5, top and middle) are possible. Either, the whole scene needs to fit into the small range where everything can be sharp or a bigger range can be used, but then prefiltering (blur) is necessary. The trade-off between these two solutions is not obvious. Our metric which accounts for underlying luminance pattern can predict the strength of perceived depth in the presence of blur due to depth-of-field. Therefore, using our optimization scheme along with this metric, leads to an optimal trade-off between the sharpness and depth range. Two modifications are required: First, based on the display specification, the focal

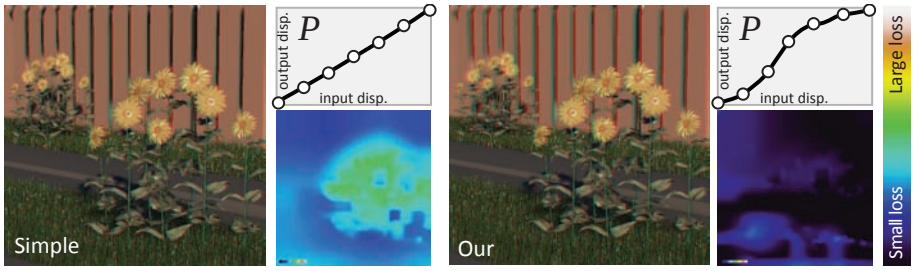


Figure 7.4: Our optimization compared to linear disparity mapping. Insets visualize mapping curves and disparity perception loss compared to the original stereo image, as reported by our metric.

range (ϕ_0, ϕ_1) has to be computed. Second, a depth of field operator $d(A, \phi_0, \phi_1)$ has to be applied to the luminance content A [Zwicker et al. 2006]. The solution is given by:

$$\arg\min_{P \in \mathbb{R}^n} \int_{\Omega} A \ominus d(r(A,P), \phi_0, \phi_1))\mathrm{d}\mathbf{x}$$

An example of this optimization is presented in Figure 7.5, bottom.

7.3.3 Validation

To evaluate our disparity optimization, we compared it to other techniques in a pair-wise comparison with three different scenes (Figure 7.6) and four different techniques: camera-parameter adjustment [Jones et al. 2001; Oskam et al. 2011] (CAM), disparity scaling in the perceptual space (PCT), the proposed here optimization scheme without (OPT-D), as well as with accounting for the luminance support (OPT-CD). For each method we ensured that the resulting disparities spanned the same range. In total, 18 pairs of stereo images were shown in a randomized order to the 17 participants who were asked to indicate which stereo image exhibits a better depth impression. In order to analyze the obtained data we computed scores (the average number of times each method was preferred) and computed a two-way ANOVA test. To reveal the differences between the methods, we performed a multiple comparison test. Detailed results of the study are presented in Figure 7.7.

The study showed, that for the scenes *Dinos* (courtesy of [Lee, Eisemann and Seidel 2009]) and *Gates*, our optimization was preferred over all other methods and the effect was significant. The lower performance of CAM, as well as PCT is due to the inability to effectively compress disparities in regions that are less crucial for depth perception. In the *Comic* scene, the difference between OPT-CD and CAM is not statistically significant for $p = 0.05$, but it is when assuming $p = 0.1$. This observation indicates that in some cases our optimization may perform similarly to others. The *Comic* scene is also interesting for another reason; the biggest depth-range compression can be obtained in the back, due to the low luminance frequency in the sky, which is correctly detected by our model. The CAM solution mostly affects the background, actually even a bit too much. The optimization more evenly distributes the depth impression (refer to the images in additional materials) and while the foreground looks similar, the background has more depth information. Nonetheless, this difference is

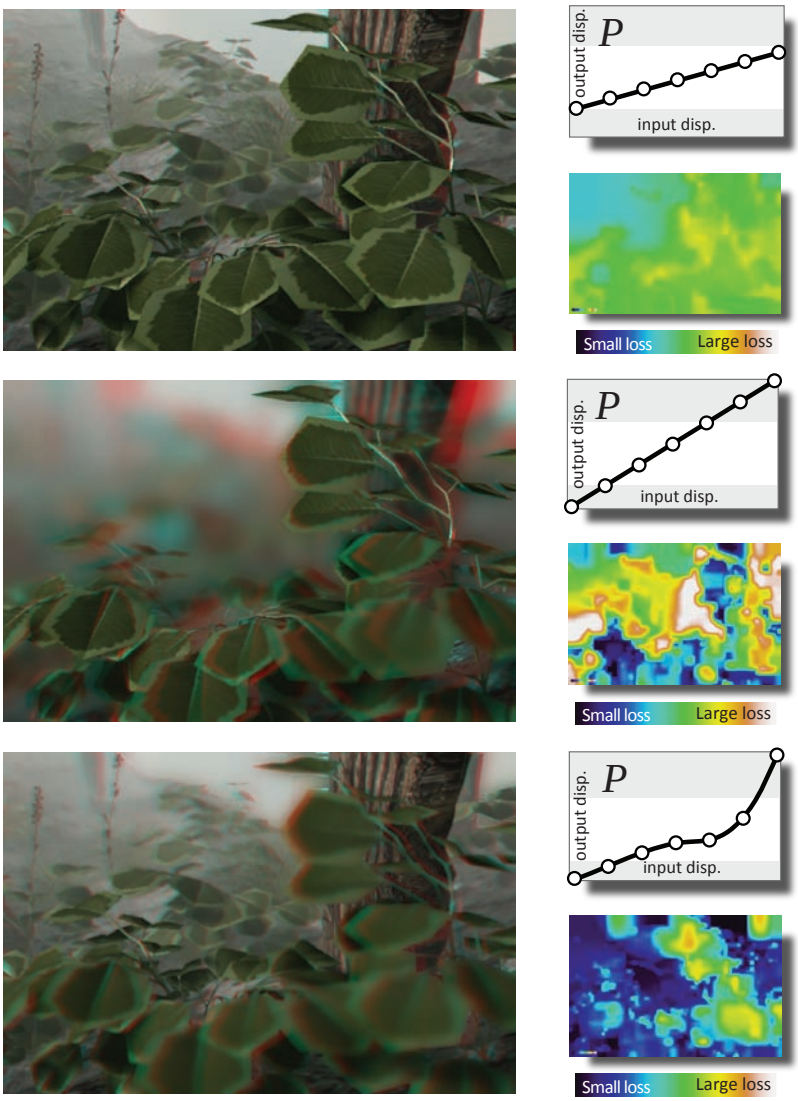


Figure 7.5: Trade-off between the depth range and sharpness on a multi-view autostereoscopic display. The insets show disparity mapping functions and the loss of depth perception due to blur. Top to bottom: simple mapping that fits entire scene in the depth-of-field region (marked in white on curve plots), disparity mapping using the entire pixel disparity range, our mapping. Our mapping leads to a good balance between depth perception and depth-of-field constraints.

very localized in the scene. When telling people afterwards to consider the farther tree and clouds, they saw the previously-missed improvement. Generally, the results show that including luminance in the model improves the performance of the disparity optimization significantly.

We also illustrate the usefulness of our optimization for autostereoscopic displays, where depth-of-field and disparity perception are linked and, hence, using our

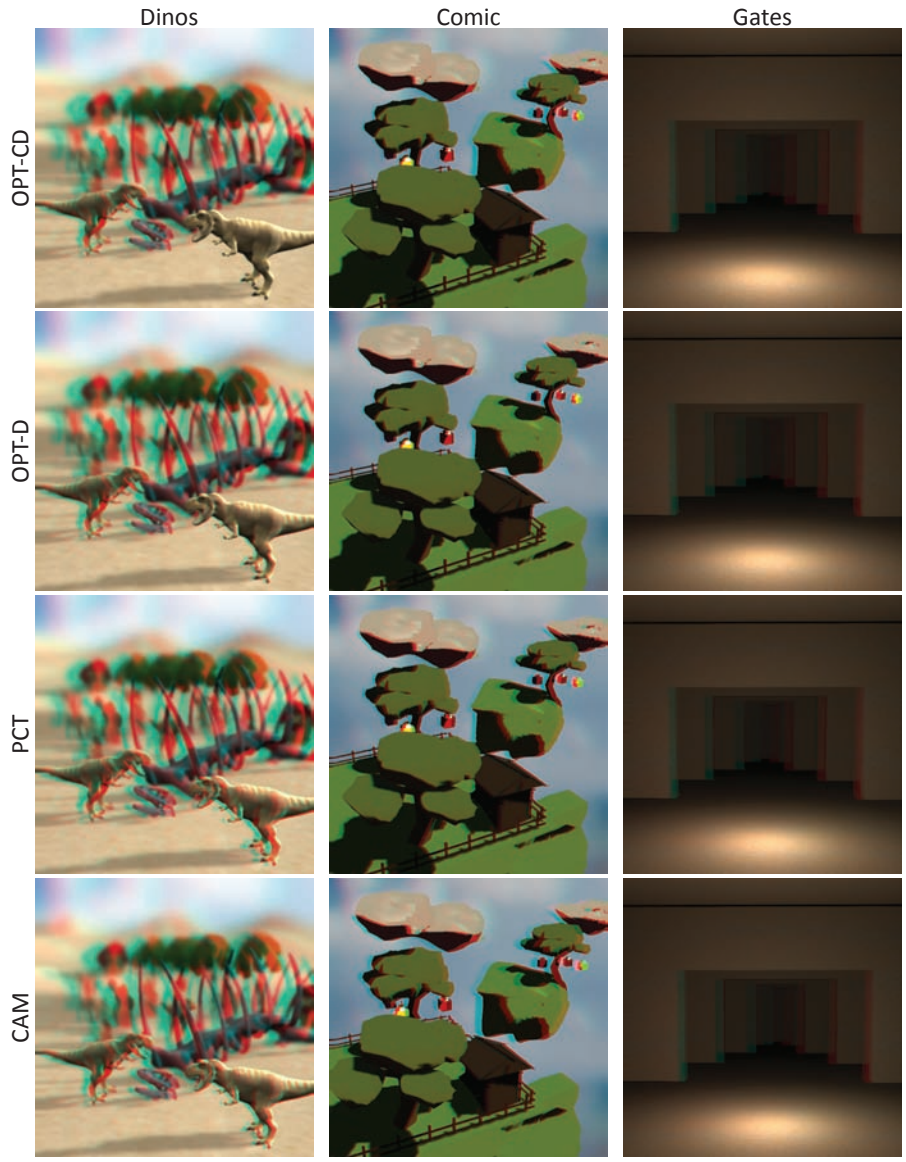


Figure 7.6: All stimuli used for our study evaluating the disparity optimization method.

luminance-disparity model presented in Section 6.4 is crucial. We used the examples from Figure 7.5. We compared our method separately to the mapping that linearly fits everything into the depth-of-field region and the one that uses the full display-disparity range. 13/14 out of 17 participants preferred the depth impression delivered by our method to using the entire depth-of-field/disparity range. A two-sided binomial statistical test revealed that this result is statistically significant with $p < 0.05$.

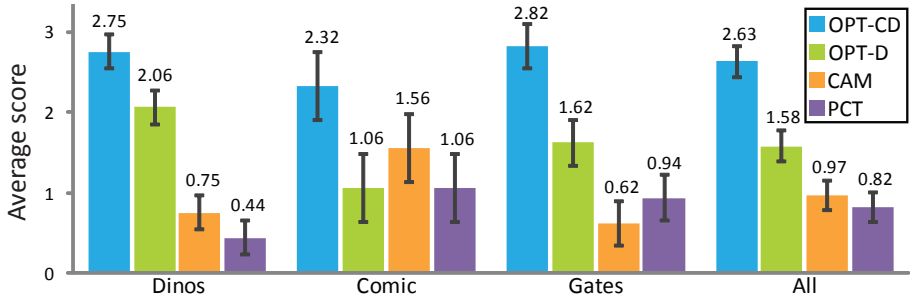


Figure 7.7: Statistical data obtained in our study. The error bars show 95 % confidence intervals.

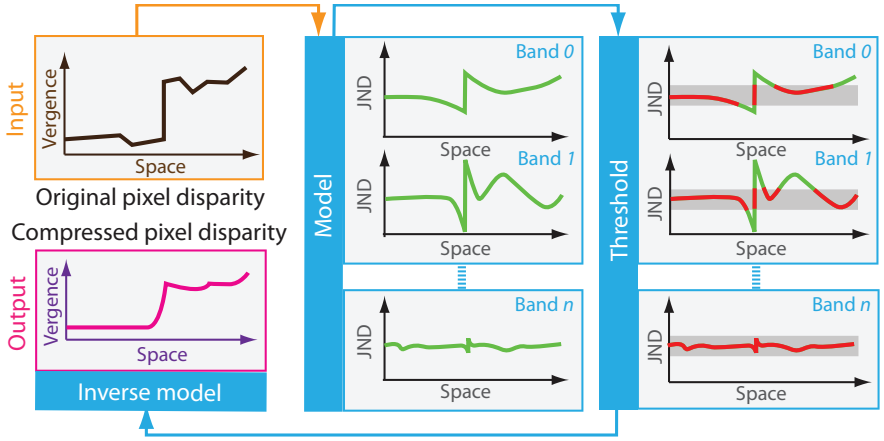


Figure 7.8: Perceptual disparity compression pipeline: An original pixel disparity (vergence) image (top left), is transformed into JND (middle). In this space, disparities which are below one JND (red dotted line) can be identified and removed, because they are not perceived (right). Optionally, a threshold of more than one JND can achieve more aggressive compression. The compressed disparity will have less details, as those which are not perceived are removed (bottom left).

7.4 Stereo Image and Video Compression

Our models can be used to improve the compression efficiency of stereo content. Key to many perceptual compression approaches is to map the signal into a perceptually uniform space, such that the perception of artifacts can be reliably controlled. This is for example the idea behind classic image compression such as JPEG [Taubman and Marcellin 2001]

We follow this idea and assuming a disparity image as input, we first convert physical disparity into perceived disparity (Figure 7.8). Here, using the model that accounts only for disparity signal. In perceptual space, disparity below the detection threshold (one JND) can be safely removed without changing the perceived stereo effect (Figure 7.2). More aggressive results are achieved when using multiple JNDs. It would further be possible to remove disparity frequencies beyond a certain value. As

shown by Tyler [1975] subjects cannot perceive disparity corrugations with a frequency above 3-5 cpd. This, however, requires further verification and was not used in our results, e. g., Figure 7.9.

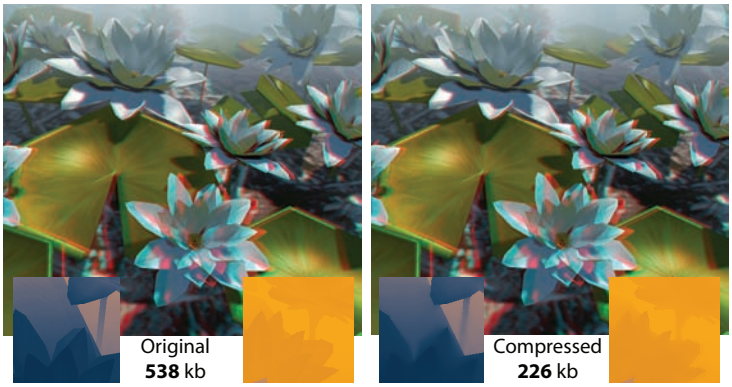


Figure 7.9: Disparity compression can be improved by operating in our perceptually-uniform space. The figure shows a stereo image, and the same image with disparities below 1 JND removed. The insets show pixel disparity and file size when compressing with LZW. Our method detects small, unperceived disparities and removes them. Additionally it can remove spatial disparity frequencies that humans are less sensitive to.

Above results can be improved when luminance information is taken into account. For this purpose our model that accounts for luminance pattern can be used. This leads to more aggressive compression in places where luminance signal weakens disparity perception. An example comparing both compression methods is presented in Figure 7.10.

Comparison In order to show that taking into account luminance information improves results, we compared compression performed using both models. For this purpose we used the examples from Figure 7.10. We compared the original stereo images to ones where all disparities below 2 JND were removed using the disparity-only model, as well as the model that accounts for underlying luminance pattern. We showed the modified images side by side (randomized) with the original image and asked about perceived differences. Each pair was shown ten times in randomized order. We asked 17 participants which compression technique produces images that are closer to the original in terms of depth. In 51 % the method with the model accounting for luminance was chosen as the one closer to the original. This suggests that, although the compression with luminance taken into account reduces storage size, it does not introduce additional perceivable artifacts.

7.5 Personalized Stereo

When displaying stereo content with a given physical disparity, its perception largely depends on the viewing subject and the equipment used. It is known that stereoacuity varies drastically for different individuals, even more than for luminance [Coutant

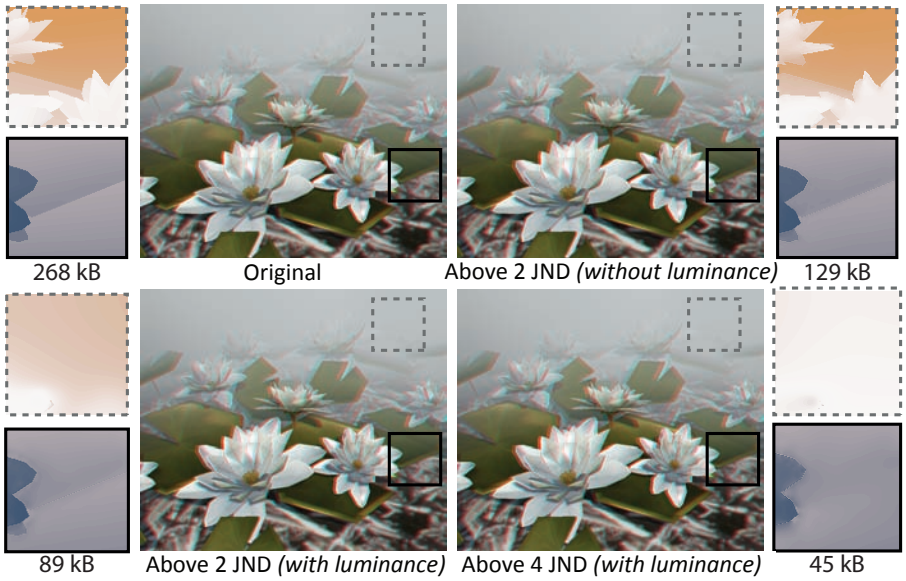


Figure 7.10: Comparison of disparity compression using model with luminance and the more conservative one without luminance. The first method can account for regions where poor luminance pattern reduces sensitivity to depth changes. Therefore, it can remove more imperceptible signal than previous techniques. In insets we show zoomed-in parts of pixel disparity maps. The size corresponds to the size of our disparity representation compressed using LZW.

and Westheimer 1993; Richards 1971]. In our applications we use an average model derived from the data obtained during experiments. Although it has the advantage of being a good trade-off in most cases, it can significantly over- or underestimate discrimination thresholds for some users. This may have an impact especially while adjusting disparity according to user-preferences. Therefore, our model provides the option of converting perceived disparity between different subjects, between different equipment, or even both. To this end a transducer (Section 6.2.2), acquired for a specific subject or equipment, converts disparity into a perceptually uniform space. Applying an inverse transducer acquired for another subject or equipment achieves a perceptually equivalent disparity for this other subject or equipment.

7.6 Apparent Stereo

As mentioned in the beginning of this chapter, it is both a technical and artistic challenge to depict three-dimensional content using flat two-dimensional screens. On the one hand, the content needs to fit within the limits of a given display technology and at the same time achieve a comfortable viewing experience. Given the technological advances of 3D equipment, especially the latter increases in importance. Modifications to stereo content become necessary that aim at flattening or even removing binocular disparity to adjust the 3D content to match the comfort zone in which the clash between accommodation and vergence stays acceptable. However, applying such modifications

can lead to a reduction of crucial depth details.

In this section we present a disparity manipulation technique that does not expand overall disparity range of a stereo image but still is able to enhance depth impression. This method builds upon the Craik-O'Brien-Cornsweet effect (Section 2.3.2), a visual illusion, which uses so-called Cornsweet profiles to produce an illusion of depth. Applying it skilfully at depth discontinuities allows, as shown in this part, to either enhance depth impression or reduce the overall disparity range to ensure a comfortable yet convincing stereo experience. An interesting case is our backward-compatible stereo, for which the disparity is low enough that overlaid stereo pairs seem almost identical, however, stereo can be experienced when the images are viewed stereoscopically. One additional advantage of the Cornsweet disparity is its locality that enables apparent depth accumulation by cascading subsequent disparity discontinuities. This way the need to accumulate global disparity is avoided leading to smaller disparity range of stereo image.

We illustrate effectiveness and usefulness of our technique by showing that Cornsweet illusion, as previously applied to brightness, can increase stereo perception without introducing a large overall disparity. We present a way of respecting potential limits of a given display technology improving at the same time depth impression. Furthermore, a user study measures the performance of backward-compatible stereo and our disparity enhancement.

In the following part, we present the various manipulations we apply to the initial disparity map. All of them are performed in the perceptually uniform space, therefore, before applying them, similar pipeline to the one presented in Section 7.2.1 needs to be used. Depending on the purpose (retargeting, enhancement, backward-compatible stereo...), the applied operations differ.

7.6.1 Retargeting

One of our main applications of the Conrsweet Illusion is to use it in the context of stereo content retargeting. Hereby, we mean modifying the pixel disparity to fit into the range that is appropriate for the given device and user preferences (distance to the screen and eye distance). Typically, such retargeting implies that the original reference pixel disparity D^r is scaled to a smaller range D^s. Consequently, in D^s some of the information may get lost or become invisible during this process. Inspired by previous work [Krawczyk, Myszkowski and Seidel 2007] in the field of tone-mapping, we want to compensate this loss by adding Cornsweet profiles P_i to enhance the apparent depth contrast.

As the perceptual decomposition is performed using a Laplacian pyramid, the bands correspond to Cornsweet profile coefficients (each level is a difference of two gaussian levels, which remounts to unsharp masking). Hence, modifying higher bands in the pyramid remounts to modifications in form of Cornsweet profiles. E.g., adding the sum of these higher bands would directly yield unsharp masking. In practice, it is a good choice to only involve the top five bands of the perceptual decomposition to add the lost disparities.We estimate the *loss* of disparity in D^s with respect to D^r by comparing the disparity change in each band of a Laplacian pyramid:

$$R_i = C_i^r - C_i^s$$

Figure 7.11: We can change the effect of depth perception by increasing JNDs. In this way, we can uniformly exaggerate the depth impression ("Big Buck Bunny" © by Blender Foundation).

where R_i are the corrections in a given band i, C_i^r and C_i^s are the bands of the reference and distorted disparity respectively.

In theory, one might be tempted to simply add all R_i directly on top of D^s. Effectively, this would add Cornsweet profiles to the signal, but care has to be taken that the resulting pixel disparity does not create disturbing deformation artifacts and remains within the given disparity bounds. In order to prevent disturbing distortions, we limit the Corsnweet profiles directly in the perceptual space, as detailed in the following.

7.6.2 Limiting Cornsweet profiles

To assure that added Cornsweet profiles do not yield a too large disparity range, we manipulate the corrections R_i. A first observation is that all values are in JND units, hence, we can limit the maximum influence of the Cornsweet profiles, by clamping individual coefficients in R_i so they do not exceed a limit given in JND units. Clamping is a good choice, as the Laplacian decomposition of a step function exhibits the same maxima over all bands situated next to the edge, is equal zero on the edge itself, and decays quickly away from the maxima. Because each band has a lower resolution with respect to the previous, clamping of the coefficients lowers the maxima to fit into the allowed range, but does not significantly alter the shape. The combination of all bands together leads to an approximate smaller step function, and, consequently, choosing the highest bands leads to a Cornsweet profile of limited amplitude. In Figure 7.11, we show how different limits result in different enhancement strength.

Unfortunately, this will not yet ensure that the enhancement layer R (composed of all R_i) combined with D^s will not result in too large value. Clamping is a straightforward

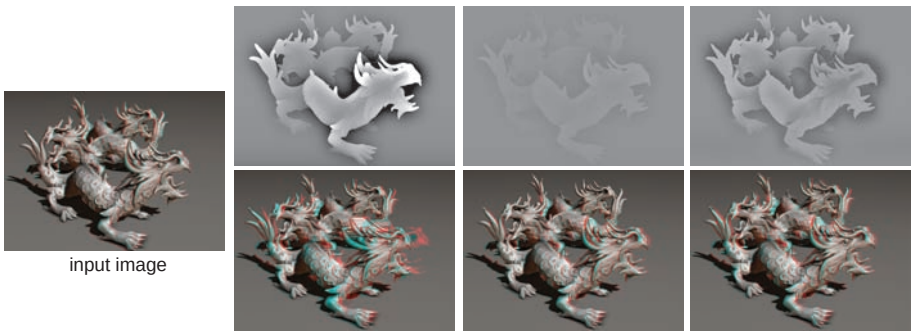


Figure 7.12: Different approaches for limiting Cornsweet profiles. All of examples were limited to the range $[-0.5, 0.5]$. A simple unsharp-masking profile can exceed the range of possible disparities while the image is enhanced (a). Logarithmic suppression (b) limits big profiles but at the same time those that could stay bigger (in the far plane) get almost invisible. Our local method (c) limits profiles locally preserving small ones.

way of limiting the profiles R, but it results in flat areas whenever the disparity bounds are exceeded. The second possibility is to scale profiles using a monotonic mapping function. Here, a good mapping seems to be a logarithmic function that favors small variations, which we do not need to clamp as they usually do not result in an exceeded disparity range. Nonetheless, an important observation is that some parts of D^s might allow for more aggressive Cornsweet profiles than others without exceeding the comfort zone. Therefore, instead of using a global method, we propose to locally scale the Cornsweet profiles to best exploit local disparity variations and to make sure that most of the lost contrast is restored. Wherever the limits are respected, these scaling factors are simply one, otherwise, we ensure that the multiplication resolves the issue of discomfort. Scaling is an acceptable operation because the Cornsweet profiles vary around zero.

Deriving a scale factor for each pixel independently is easy, but if each pixel were scaled independently of the others, the Cornsweet profiles might actually disappear. In order to maintain the profile shape, scaling factors should not vary with higher frequencies than the scaled corresponding band. Hence, we compute scale factors per band.

One observation is that we relied on a pyramidal decomposition, consequently, R_i has a two times higher resolution than R_{i+1}. This is important because when deriving a scaling S_i per band, it will automatically exhibit a reduced frequency variation. Hence, we derive per-pixel-per-band scaling factors S_i that ensures that each band R_i when added to D^s would not exceed the limit. Next, these scaling factors are "pushed down" to the highest resolution from the lowest level by always keeping the minimum scale factor of the current and previous levels. This operation results in a high-resolution scaling image S. We finally divide each S by the number of bands to transfer (here, five). This ensures that $D^s + \sum_i R_i S$ respects the given limits and maintains the Cornsweet profiles. Figure 7.12 illustrates our local scaling in comparison to other approaches and shows that it best preserves the Cornsweet profiles, while reproducing most of the original contrast.

7.6.3 Artistic enhancement

Our previously described retargeting ensures that contrast is preserved as much as possible. Although this enhancement is relatively uniform, it might not always reflect an artistic intentions. E.g., some depth differences between objects or particular surface details might be considered important, while other regions are judged unimportant. Figure 7.13 (bottom) shows an example where the distance between the two dragons in the background has been enhanced, as well as the details in the foreground where the dragon scales appear more detailed. It is also possible to increase the overall depth impression in the scene by increasing disparity scaled in JNDs units (see Figure 7.11).

To give control over the enhancement, we developed a simple interface that allows an artist to specify which scene elements should be enhanced and which ones are less crucial to preserve. Precisely, we allow the user to specify weighting factors for the various bands which gives an intuitive control over the frequency content. Using a brush tool, the artist can directly draw on the scene and locally decrease or increase the effect. By employing a context-aware brush, we can achieve ensure edge-stopping behavior to more easily apply the modifications.

7.6.4 Backward-compatible Stereo

The need for specialized equipment is one of the main problems when distributing stereo content. As an example, consider printing an anaglyph stereo image on paper: the stereo impression can be enjoyed with special anaglyph glasses, but the colors are ruined for spectators with no such glasses. Similarly, observers without shutter glasses see a blur of two images when sharing a screen with users wearing adapted equipment. We approach this backward-compatibility problem, in a way that is equipment and image content independent, by employing our model.

Using our technique, we can produce backward-compatible stereo that “hides” 3D information from observers without 3D equipment. The observation is that a zero disparity leads to a perfectly superposed image for both eyes. Unfortunately, this also implies that no 3D information is experienced anymore. Therefore, our goal is to reduce disparity where possible to make both images converge towards the same location, hereby it appears closer to a monocular image. In particular, this technique can transform anaglyph images and makes them appear close to a monocular view.

The implementation follows the same process as for the retargeting, but we do not add the scaled disparity. In this case, the Cornsweet profiles will create apparent depth discontinuities, while the overall disparity remains low. This is naturally achieved because Cornsweet profiles are centered around zero. The example comparing our backward-compatible stereo with original 3D stereo rendering is presented in Figure 7.14

The solution is very effective, and has other advantages. The reduction leads to less ghosting for imperfect shutter or polarized glasses (which is often the case for cheaper equipment). Furthermore, more details are preserved in the case of anaglyph images because less content superposes. This is particularly visible for the grass and sky in the foreground of Figure 7.15. Furthermore, it is important to realize that much of the scene structure remains understandable because the HVS is capable of propagating some of the perceived differences over the neighboring surfaces. When comparing

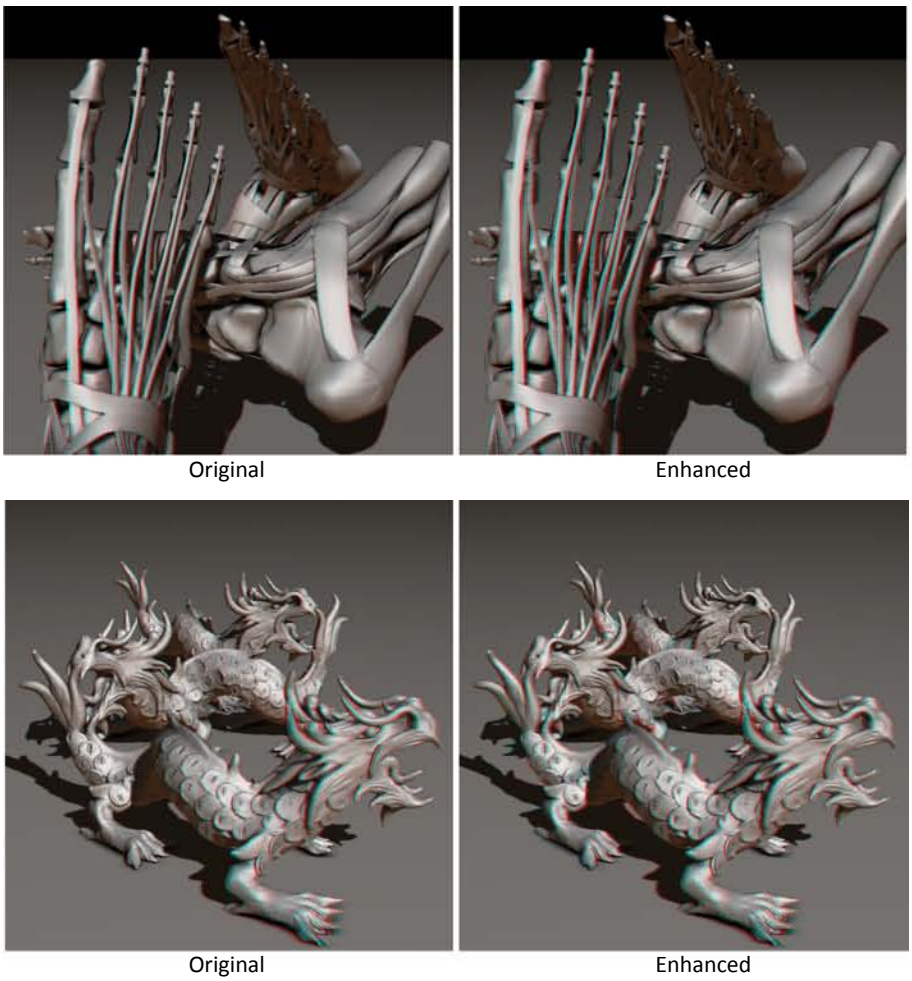


Figure 7.13: Depth enhancement using the Cornsweet illusion. Original and enhanced anaglyph images are shown for two different scenes with significant depth range. Note a better separation between the foreground and background objects and a more detailed surface structure depiction.

to an image of equivalent disparity (scaled to have the same mean), almost all depth cues are lost. In contrast, to produce a similar relative depth perception, the disparity can become very large in some regions even causing problems with eye convergence. Finally, our backward-compatible approach could be used to reduce visual discomfort for cuts in video sequences that exhibit changing disparity [Lang et al. 2010].

7.6.5 Photo Manipulation

Finally, converting 2D photos into 3D [Saxena, Chung and Ng 2005] is never perfect. To minimize and facilitate the user interaction, we can concentrate on local discontinuities and avoid a global depth depiction. According to our findings even a localized depth

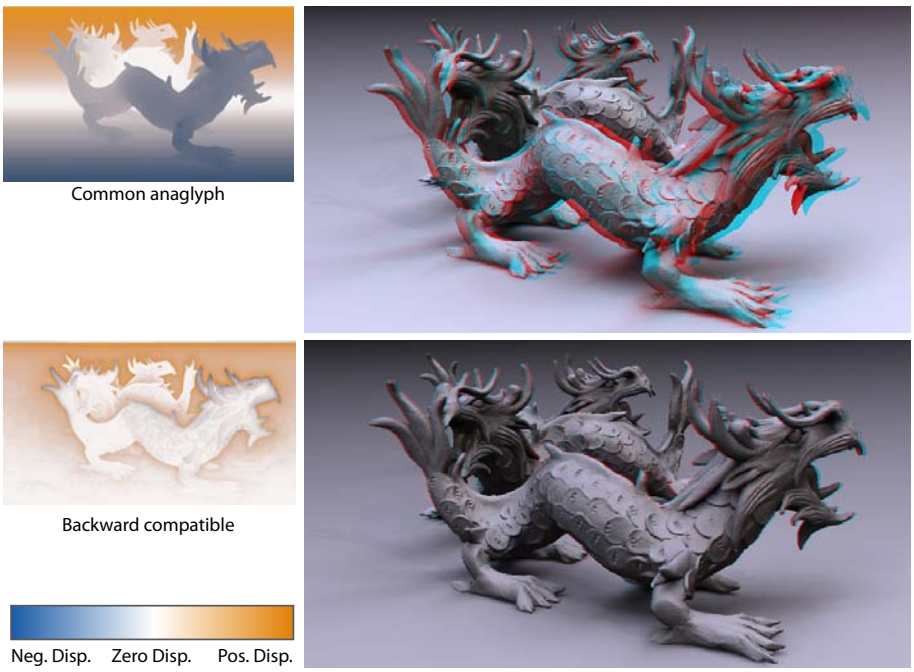


Figure 7.14: Backward compatible stereo provides just-enough disparity cues to perceive stereo, but minimizes visible artifacts when seen without special equipment.

Figure 7.15: Converting a photo (Left) into a 3D image (Middle, anaglyph), just by using the blue channel as depth. Our enhancement (Right, anaglyph) can be used to put stereo cues only where depth contrast exists, minimizing the global error due to the naïve 3D reconstruction, but with locally plausible cues. The insets on the right depict the corresponding disparity maps.

representations can deliver a good scene understanding (refer to Figure 7.15). This is not surprising, as it is an observation that has been used for centuries in the form of bas-relief depictions. In fact, again the Cornsweet profile seems to be a very effective shape in this context.

7.6.6 Evaluation

To evaluate the backward-compatible approach, we performed two user studies. In our first study, ten naïve subjects participated. First, we investigated the overall quality of our method. For this, we handed a backward-compatible stereo image with "hidden" anaglyph content. We asked each subject for flaws or particularities in the image. None of those that received our output reported the artifacts produced by the stereo information within the first minute. Furthermore, only two subjects reported this observation within two minutes. After two minutes, the subjects received anaglyph glasses and were asked to report their observation concerning the stereo impression of the backward-compatible stereo image and the standard 2D image shown side by side. All 10 subjects agreed that the backward-compatible stereo image exhibits a 3D effect whereas the standard image does not. Obviously, such results depend on the underlying image content, but the findings give a clear indication that 3D content can be hidden to a large extent.

The second study was conducted to measure the depth effect of our solution and to show that it reduces disturbing artifacts when not using special equipment. To this extent, we let six participants compare the depth percept of two stereo images, one with our backward-compatible stereo and one with standard stereo. We then asked them to adjust the disparity in the standard stereo image (by approaching the two cameras), such that the depth impression was equivalent to our backward-compatible version. Such an adjustment of camera distances is similar to performing micro-stereopsis [Siegel and Nagata 2000]. In Figure 7.16, we show comparison of the backward-compatible version and the average result.

7.7 Joint Luminance and Disparity Manipulations

The disparity metric presented in Section 6.4.5 can predict the perceived change of distorted disparity, just like the effect of luminance distortions on perceived depth. Hence, we can identify image regions, where the stereo impression is weak due to poor luminance support. We can quantify this effect by comparing two stereo images with the same disparity pattern but an assumed-perfect luminance pattern in one of them.

By improving the luminance contrast in areas where the original support proves insufficient, we re-introduce the impression of depth as shown in Figure 7.17. In Figure 7.18, we also use this technique to illustrate the successful detection of asymmetries described in Section 6.4.

We also tested whether our luminance pattern in Figure 7.17 improved depth perception. We compared both images, i. e., with and without introduced luminance pattern, and asked people to choose image that exhibits more depth. 16 out of 17 participants chose the solution images where the additional luminance pattern was used. A two-sided binomial statistical test revealed that this result was statistically significant with $p < 0.05$.

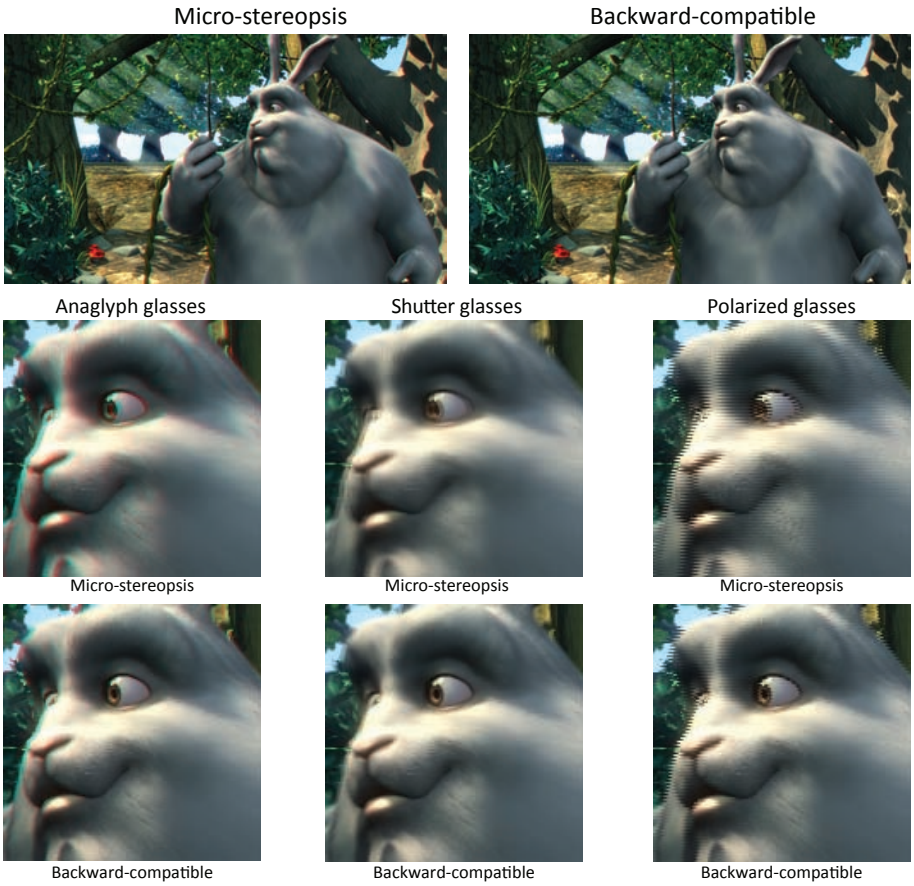


Figure 7.16: The figure presents a comparison between the backward-compatible stereo and the micro-stereopsis technique. The second method was adjusted in a way that both versions exhibit the same depth impression. The insets present zoomed-in versions of images displayed using differed stereo equipment. It can be seen that although the depth impression in both cases is very similar, the backward-compatible version reveals less disturbing artifacts while watched without stereo equipment. This is well visible especially for areas without depth discontinuities such as body of the bunny (notice in particular the shadows), or inside the tree ("Big Buck Bunny" © by Blender Foundation)

7.8 Hybrid Images

Hybrid images change interpretation as a function of viewing distance [Oliva, Torralba and Schyns 2006]. They are created, by decomposing the luminance of two pictures into low and high spatial frequencies and mutually swapping them. The same procedure can be applied to stereo images by using our disparity band-decomposition and perceptual scaling (Figure 7.19).

Figure 7.17: An insufficient luminance support in the original stereo image *(left)*, lowers its depth perception (top right). By adding a hatching pattern, guided by our metric, the resulting stereo image *(middle)* shows significantly less stereo loss (bottom right).

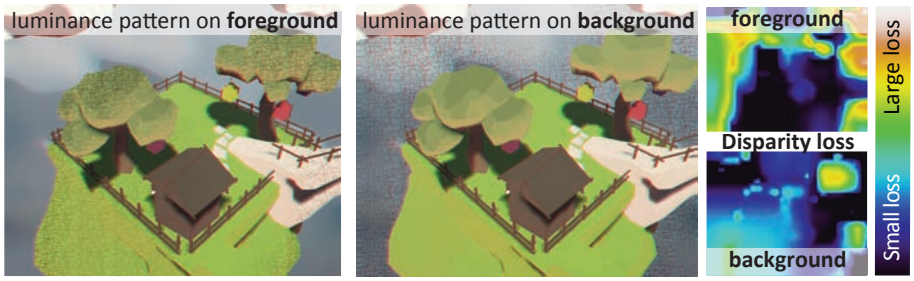


Figure 7.18: To illustrate the prediction of asymmetries, we show two cases: hatching on the foreground *(left)* and the background *(middle)*. Compared to foreground hatching (right top), background hatching creates more pronounced differences due to disocclusions, leading to better depth perception (right bottom), as correctly predicted by our metric.

7.9 Conclusions

In this chapter, we proposed a number of perceptually-motivated disparity manipulation techniques, which are based on the disparity models described before (Chapter 6). By using them, it is possible to improve existing, but also develop new compelling applications, such as an image optimization for multi-view autostereoscopic displays, backward-compatible stereo, personalization or joint luminance-disparity processing. Those techniques demonstrate that even simple operations, can be enhanced when human perception is taken into account.

While modern rendering effects (depth of field, lens flare, motion blur, veiling glare, participating media, as well as poor visibility conditions – rain, night, ...) increase realism or artistic/aesthetic value, they also affect luminance contrast, which in turn influences the disparity perception. With techniques presented in this chapter, for the first time, an adequate disparity handling becomes possible in all these situations. The disparity optimization method is a good alternative to previous methods for disparity-range control. It was shown that considering luminance significantly improves the results of the proposed mapping technique.

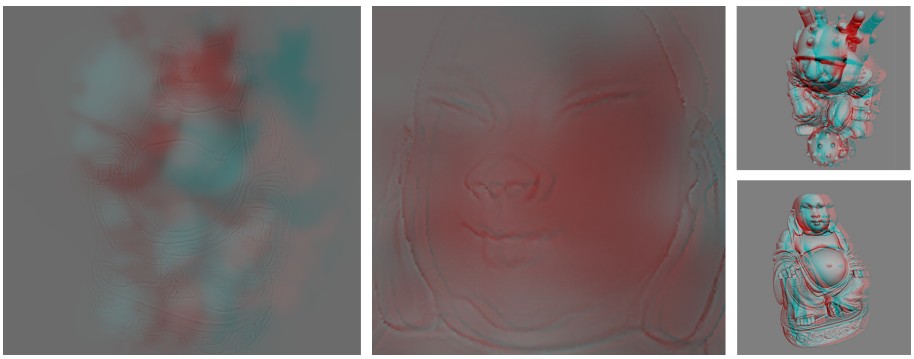


Figure 7.19: A hybrid stereo images: nearby, it shows the Buddha; from far away, the Grog model.

Especially interesting for perceptually driven manipulations is to use the Consweet Illusion. We showed that it is a practical tool for depth impression enhancement and that the possibility of Corsnweet profiles cascading enables a good backward-compatible stereo. One limitation is that, similarly to the Cornsweet Illusion in luminance, the manipulation might change the appearance of the shape or even material to some extent. On the other hand, we do not manipulate colors in the rendered image itself, which means that we preserve many of the original cues (lighting, material) that are particularly helpful in conveying a satisfactory overall appearance. This is particularly visible in complex stimuli (Figure 7.16) where the spatial layout is convincingly captured without introducing large disparities. These properties make backward-compatible stereo an interesting trade-off.

The solutions presented here are general in the sense that they do not depend on the way the input images were captured, be it 3D rendering, a depth camera, or a multi-view surface reconstruction. Further our techniques are independent of the 3D display technology used to present the stereo color image pair. All the manipulations can be also performed at interactive framerates. Because all operations are realizable on a GPU and are applied to textures, the solution performs almost independently of the geometric complexity of the scene and could be potentially implemented as a small computational unit in TV-sets.

There are many interesting avenues for future research. In our work, we did not show how all of our disparity manipulations perform with both our metrics, i. e., with/without luminance taken into account. In particular, it would be interesting to investigate how apparent stereo manipulations can benefit from the luminance-disparity model.

Not all stereo cues are equally important for all distances, thus, other stereo cues could be enhanced, when disparity becomes ineffective. For example, warm-cold shading might distort colors, but helps in conveying spatial organization. Similarly, motion parallax becomes a strong depth cue at certain distances. In fact, generating exactly those stereo cues that are actually used for a certain depth, while minimizing their distorting effect, would allow to improve rendering performance and maximize the perceptual effectiveness. Disparity enhancement methods could be improved by taking into account some inabilities of the HVS, which can be detected using our disparity

models. Similar idea was previously used in the context of brightness enhancement [Didyk et al., 2008], where human inability to perceive details in bright regions was exploit.

We believe that in the future, models such as ours will be crucial for stereo images and video processing. Many other applications are possible; combined tone and disparity remapping for HDR stereo content, or luminance hatching could be combined with other styles of non-photorealistic rendering. We also believe that our models could be integrated in a 3D video-conference system, as, especially in architectural environments, regions with weak luminance variations are common. Further, our way of optimizing 3D content could be used to consider different viewing conditions or even viewers.

8

Stereo Upsampling

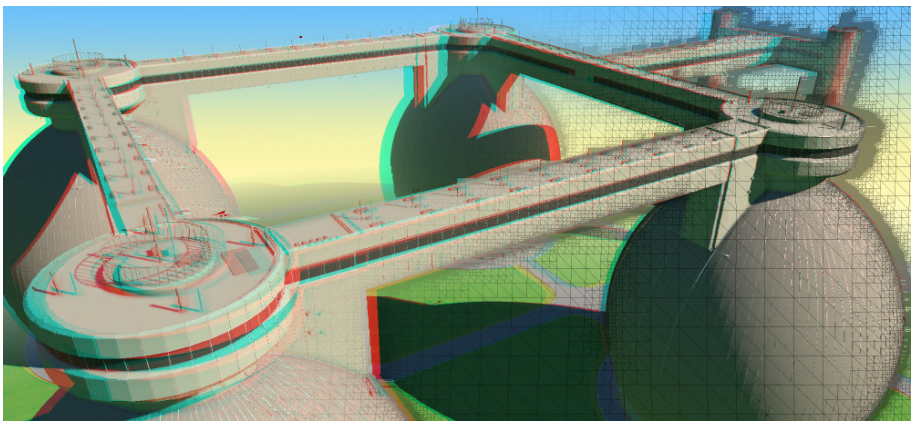


Figure 8.1: Stereo image created using our techniques, which uses adaptive grid for artifact-free warping.

As pointed out in Section 3.1.3, although stereo vision has received recently much attention due to its broad success in feature films, visualization and interactive applications such as computer games, it does not come for free and often implies that two images need to be rendered instead of a single one, as for standard rendering. This can have a high impact on performance which is an issue for real-time applications. Therefore, it becomes a good idea to use image-based techniques to lower the cost of producing two views. Such techniques play also an important role in disparity manipulations as those presented in Chapter 7, where the adjusted disparity does not correspond to the actual depth of the scene. Therefore, recapturing or re-rendering the scene with the new depth is impossible and image-based techniques are necessary to resynthesize new views.

8.1 Overview

In this chapter, we propose to create only a single view of the scene, together with its depth buffer and use image-based techniques to generate two individual images for the left and the right eye. The resulting stereo effect is of a high quality, but our

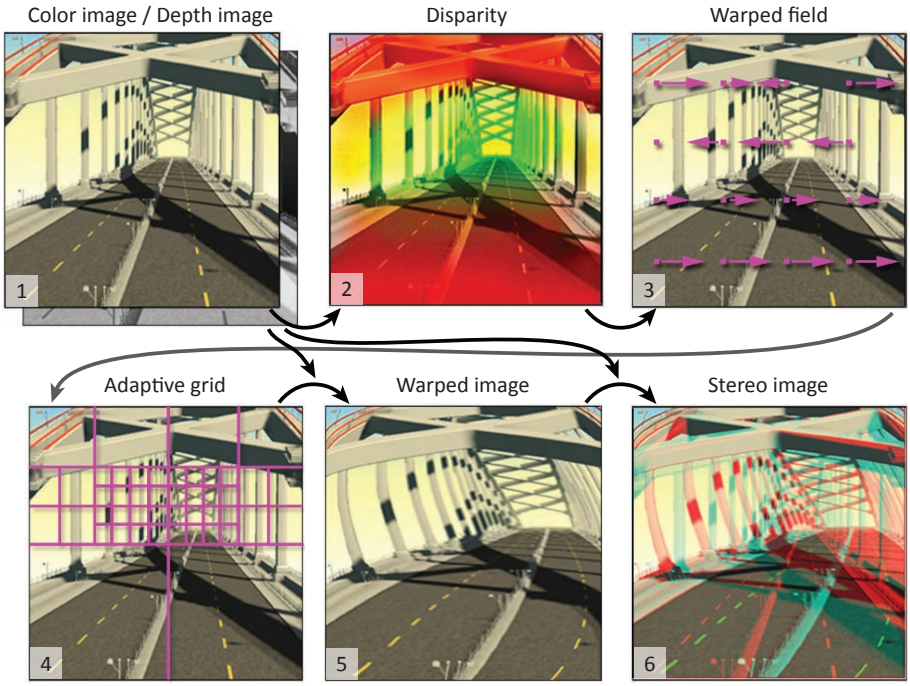


Figure 8.2: Our image-based stereo view synthesis pipeline: We assume a rendered image with depth buffer, as well as a disparity map as the input of our method (1). If desired, the disparity map can be computed from the input depth image (2). Next, we build a warp field of this disparity mapping (3). This field is discretized adaptively: Areas with similar disparity are warped as large blocks, areas of different disparity are warped as small blocks (4). Finally, the input image and the new warped image (5) are used as a stereo image pair (6), here, presented in anaglyph stereo.

approach avoids the cost of rendering two individual frames. In this context, we address two major challenges. First, our stereo view-synthesis should show a performance behavior that approaches the rendering time for a single view. Second, the stereo image pair should have as few artifacts as possible. Our solution addresses both issues via an adaptive algorithm that respects depth disparity, exploits temporal and spatial consistency, and maps well to the GPU. This method is an extension of our temporal upsampling presented in Chapter 4, however, here, we present a number of improvements that target directly stereo-image synthesis.

This chapter is structured as follows. We first propose our algorithm in Section 8.2. Then we present results in Section 8.3. Strengths and limitations are discussed in Section 8.4, before we conclude in Section 8.5.

8.2 Our Approach

In this section, we propose a pipeline (Section 8.2.2) to turn a rendered image with depth into a stereo image pair as shown in Figure 8.1. To this end, we first show how a

pixel disparity mapping (Section 8.2.1) from an image location in one eye to the image location of the other eye can be computed for rendered content. We observe that this mapping is piecewise smooth, and exploit this fact to efficiently create a high-quality stereo image pair using an adaptive approach (Section 8.2.3). Finally, we discuss how to improve the result further by warping not only between the left and right eye, but also between the current and previous frames (Section 8.2.5). In particular, this modification also ensures convergence (Section 8.2.6) to the reference in the case of a static scenes and a decelerating camera.

8.2.1 Pixel Disparity Mapping

Let $\mathbf{y} \in \mathbb{R}^3$ be a point in world space and $\mathbf{x}_{\text{left}} \in \mathbb{R}^2$ its projection into the left eye's view as well as $\mathbf{x}_{\text{right}} \in \mathbb{R}^2$ it's projection into the right eye's view. For the purpose of this section, we call the mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$ which maps every left image position $\mathbf{x}_{\text{left}}$ to its right image position $\mathbf{x}_{\text{right}}$ the *the pixel disparity mapping* from left to right. Further, we simply call the distance $\|\mathbf{x}_{\text{left}} - \mathbf{x}_{\text{right}}\| \in \mathbb{R}^+$ the *pixel disparity* of $\mathbf{y}$. This definition is analogous to the definition given in Section 2.3.2.

Given a depth map, the simplest method to generate a pixel disparity map is to apply a scale and bias to all values. In the case of a rendered scene, the depth can be directly output by the GPU, but our method does not rely on this particular feature and would support alternatively determined depth/disparity. We use a simple fragment program that applies a scale and bias to the depth in order to derive a disparity map. We adjusted our results in such a way, that both negative and positive parallax is present, as preferred by most viewers.

8.2.2 Pipeline

Our basic approach follows the pipeline depicted in Figure 8.2. In order to facilitate the explanations, we will focus on how to produce a right image out of a given left image. Later in Section 8.2.6, we will extend this setting. We assume that the pixel disparity mapping f is an input to this process and use it to convert a single image with depth information $I_{\text{left}}(\mathbf{x})$, into a pair of stereo images $I_{\text{left}}(\mathbf{x})$ and $I_{\text{right}}(\mathbf{x}) = I_{\text{left}}(f(\mathbf{x}))$.

Simply applying f in a pixel-wise fashion as done in previous approaches Section 3.1.3, can lead to holes and is not efficient to compute on a GPU, as it involves data scattering. Therefore, we represent f as a quad *grid*, i. e., a mapping from areas to areas instead of points to points. By doing so, we avoid holes and allow a parallel computation based on reverse reprojection (*gathering*) instead of forward reprojection (*scattering*) (Section 3.1.2), which is preferred for GPUs. To this end, we follow approach described in Chapter 4: We start with a regular grid much coarser than the screen resolution and sample f at every vertex, we then warp this grid as textured quads into I_{right} and use I_{left} as a texture. While a grid-based approach avoids many holes, special considerations are required for the case of *occlusions* and *discoclusions*.

Occlusions occur when multiple locations $\mathbf{x}$ in I_{left} map to the same location in I_{right}. This happens for example, when a nearby object with a strong disparity covers a background object with low disparity in I_{right}. Indeed f might not have a unique inverse for some locations. However, such ambiguities can be resolved completely by

using the depth information from $I_{\text{left}}(\mathbf{x})$: Whenever a pixel is written to $I_{\text{right}}(\mathbf{x})$, we compare its depth to the depth in $I_{\text{right}}(\mathbf{x})$ and omit the writing if its depth is bigger. In practice, this can be achieved using standard GPU depth buffering similarly to our temporal upsampling technique.

Contrary to occlusions, disocclusions lead to holes because the originally hidden information is missing, but needed. Using the described grid warping, such holes are essentially filled with content from the input image by stretching the grid. A better solution, using multiple-image warping, is discussed in Section 8.2.5.

8.2.3 Adaptive Grid

While the described so far approach succeeds in producing stereo image pairs (Section 8.3), it has two main drawbacks. First, if the image has many details in depth, a regular, coarse grid representation of f leads to undersampling and aliasing problems, i. e., low quality (Figure 8.8). Second, just increasing the grid resolution (or keeping any fixed resolution), wastes an excessive amount of grid vertices in areas which are essentially simple to warp using a low number of vertices, i. e., achieving low performance. We will now alleviate these two shortcomings by introducing an adaptive discretization of f.

As f is smooth over large areas, except at a few discontinuities, we construct a grid that *adapts* to the structure of f. We start from an initially regular grid (in practice, we start with a 32×32 grid to achieve enough parallelism, in theory one could start with 1×1 as well). The grid's quads are stored as a list of quad centers in an OpenGL vertex buffer object. A geometry shader traverses all these quads in parallel, and either

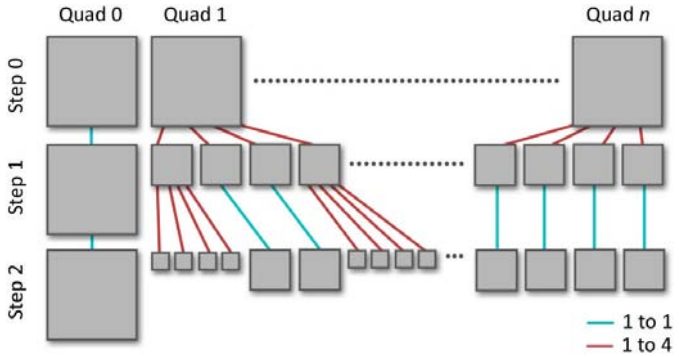


Figure 8.3: Multiple quads (horizontal) subdivided in parallel using multiple steps (vertical). In every step, every thread produces either a single quad (1-to-1, blue) or four (1-to-4, red) new quads. In the next step, each quad is again processed in parallel. We repeat this until quads are pixel-sized.

outputs the same quad/center again, or refines this quad into four new quads/centers (Figure 8.3). This process is iterated until all quads are sufficiently refined and the structure well reflects the discontinuities in f.

The decision whether a subdivision should be applied is based on the difference between minimal and maximal disparity inside the quad. If this difference is larger

than some threshold four subquads are produced, otherwise the quad is left unchanged. The output is captured in a second vertex buffer object using the OpenGL transform feedback extension. This subdivision process is iterated until the level 0 of 1×1-pixel-sized quads is reached in the regions where needed (hence, the number of steps depends logarithmically on the resolution of the input frame).

An alternative approach would be to directly refine a quad to many subquads, without recursion and without transform feedback. This leads to strongly varying output sizes (between one and several hundred vertices) which is not recommend for the geometry shader. Distributing the work amongst as-many-as-possible new threads after each subdivision is the preferred approach and allows for much more parallelism [Meyer et al. 2009].

Finally, when the subdivision is finished, we transform the vertex buffer object (VBO) quad centers back into a grid. For this, we use a second geometry shader that consumes quad centers and produces quads. f is evaluated for each corner of a quad, and each quad is drawn to I_{right} using I_{left} as a texture, as described in the previous Section 8.2.2.

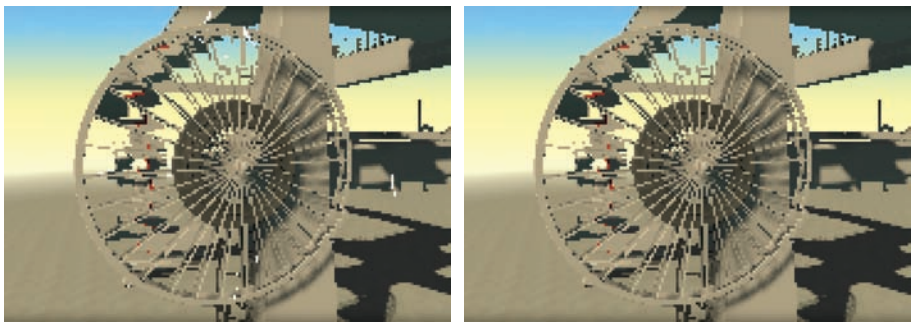
In order to avoid holes when disocclusions occur, it is important to realize that the grid vertices always fall on locations *between* two pixels (i. e. at level 0, a 1×1 quad maps to the corner of a pixel). We select the preferred pixel to fetch f and I_{left} based on its depth. That is, we fetch all four adjacent pixels around a vertex in I_{left} and use depth and disparity from the pixel with the smallest depth. By doing so, vertices adjacent to disocclusions effectively stretch the background avoiding holes.

8.2.4 Implementation Details

The position and level information for each quad is packed into an 8-bit RGB texture ($10 + 10$-bit position, 4-bit level).

To efficiently bound the amount of difference between minimal and maximal disparity inside a quad we use a min/max MIP-map. This map is similar to a common MIP-map, alas instead of storing the average, it stores the minimum and the maximum of all pixels below a pixel on higher levels. Such a map can efficiently be constructed in a parallel recursive fashion. Starting from level 0 at full resolution, a fragment program visits every pixel of the next-lower level and stores the minimum and the maximum of the four pixels from the lower level. This process is repeated until arriving at a single-pixel image, which, in our case, would store the minimum and maximum of all disparity values.

We set the subdivision threshold to 3 pixels which basically leaves only a low number of spurious single-pixel holes due to T-junctions, which occur if one quad is neighbor to a quad that is subdivided more. While a T-junction removal method could fix such problems, it usually generates again a higher and varying number of output vertices form the geometry shader. Doing so would significantly lower the geometry shader throughput, which is the bottleneck in our computation. We found the most efficient and simplest solution is, to just fill the undefined pixel via inpainting. In practice one can chose a random neighbor pixel in image space (Figure 8.4).



Before hole filling After hole filling

Figure 8.4: We stop subdividing before reaching a pixel exact result (left) and fill the few remaining holes (right). Note, that this is an inset and pixel-sized holes are proportionally much smaller in multi-megapixel images.

8.2.5 Using multiple images

Changing from a regular grid to an adaptive grid results in speed and quality improvements. Disocclusions remain the *only* visible artifact. By stretching the grid quads, the artifacts become less visible, but they can be perceived in certain configurations.

While disocclusions can ultimately not be solved without re-rendering, in this section, we will discuss how to use multiple images and multiple mappings to produce an improved stereo image pair.

We will use a previously rendered image I_{old} together with a mapping g which maps from the past view into the current view of the same eye (Figure 8.5). While f was defined to be a pixel disparity mapping, g is not. Nonetheless, it is a mapping from $\mathbb{R}^2$ to $\mathbb{R}^2$ as well. g is also constructed rapidly via a fragment program which is executed on all depth-buffer pixels in parallel. These are unprojected from the old view and re-projected into the new view. The resulting 2D displacement is stored. As for f, g is not defined everywhere, for example if a location in the current frame was clipped in the previous frame.

We can now produce an alternative right stereo image $I_{\text{right}}(\mathbf{x}) = I_{\text{old}}(g(\mathbf{x}))$. I_{old} should be used whenever a disocclusion is present. To get the best result of both we carefully choose between the two sources. In practice, we use the stretching difference inside a quad: If a quad undergoes varying stretching, it is likely to cause a disocclusion (it "tears up" the space) and should therefore not be used. Precisely, we use a *preference* operator w, arriving at

$$I_{\text{right}}(\mathbf{x}) = \frac{w(f)(\mathbf{x}) \cdot I_{\text{left}}(f(\mathbf{x})) + w(g)(\mathbf{x}) \cdot I_{\text{old}}(g(\mathbf{x}))}{w(f)(\mathbf{x}) + g(f)(\mathbf{x})},$$

with

$$w(h)(\mathbf{x}) : (\mathbb{R}^2 \to \mathbb{R}^2) \to (\mathbb{R}^2 \to \mathbb{R}).$$

The operator w turns the (disparity) mapping h into a spatially varying preference for that mapping.

Although there is no guarantee, that all occlusions will be resolved. This strategy performs rather well because a disocclusion in one mapping will often not be a dis-

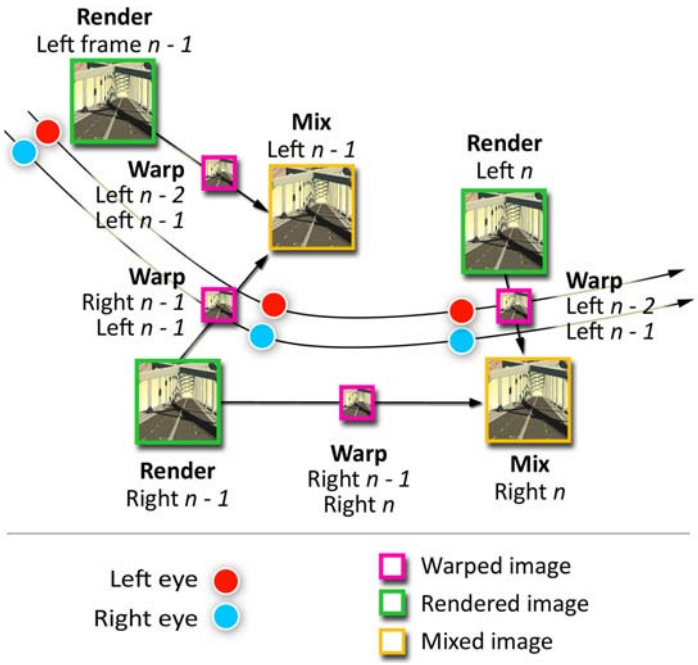


Figure 8.5: Using multiple images to reduce disocclusions and improve quality. Consider the two eyes (red and cyan circle) of a moving observer in a virtual world (arrow). Ground truth would produce two images in each frame. Instead, we produce one frame only (green), warp (magenta) from the past and the other eye, and merge (yellow) according to the one with the lower error. To achieve convergence when slowing down or halting, we alternate the rendered and the synthesized image.

occlusion in another. Following the same strategy, we can also avoid the T-junction holes nearly completely. Only such holes that are present in *both* images remain holes, which is never the case in practice when relying on a three pixel threshold in a multi-megapixel image.

8.2.6 Convergence

One final step can further improve the result: Instead of always rendering the left eye view and creating a right eye view, we can *swap* the eye roles and either warp from left to right or right to left. Swapping eyes in every frame does not lead to a strong improvement as long as the viewer is moving, nonetheless, also no temporal artifacts are introduced. However, already in this setting, if the speed of the motion decreases, w will prefer the past image, and ultimately, when no animation is present, w will always pick the past right eye for the current right eye and the past left eye for the current left eye, i. e. the result converges to the static reference.

In order to further improve the quality in the case the camera is moving, instead of toggling, it is best to choose the most distant eye view from the previously rendered. In such a way we minimize the potential disocclusion. In order to visualize the advantage of this choice, one can imagine a constant panning movement. If the left eye always

falls on the old position of the right eye, a toggling would be harmful, as it would lead to the same view being rendered twice. Choosing the most distance view eases the handling of disocclusion. In this particular case, in combination with the operator w, our algorithm even produces the reference result, although the camera is no longer static.

8.3 Results

In this section we evaluate quality and performance of our approach. We used an NVIDIA Quadro FX 5800.

To test our approach, we have chosen mostly architectural models because they represent an excellent stress test with many occlusions, disocclusions and fine details. All models are rendered using shadow mapping, per-pixel deferred shading, fog, depth of field and screen-space ambient occlusion. With such a set-up it takes around 40 ms to produce a frame. We excluded the computation of the disparity from all timings as we assume it to be an input of our method.

We compare our method to three other approaches. First, straightforward mapping of a 1×1 grid, including handling of occlusions in the same way as the method presented here does. This approach is our *reference* solution in terms of speed. Using our method by morphing only one image we can only approach the quality of such solution. An improvement is possible using more views as described in Section 8.2.5. Second, we show that our method produces better results in terms of speed and quality than using pixel-wise re-projection. We also compare our method to approach presented in Chapter 4 to which we refer as "Simple Grid". This method although targets temporal upsampling, can be used directly for producing stereo images (Section 4.4.3). It is significantly faster than the reference approach, but has lower quality. We will substantially improve upon this method in terms of quality, and in some cases even in terms of speed.

8.3.1 Quality and Performance

To show the importance of using an adaptive approach we compared our one view morphing method to the naïve, reference solution. Although we cannot improve the quality, we can bound an error by setting the subdivision threshold properly. Doing so, the solutions of both methods become indistinguishable but due to the adaptivity, our solution is several times faster.

In Figure 8.8 we compare the performance and the quality of our approaches as well as method from Chapter 4 ("Simple Grid") to ground truth rendering. First, we see how our approach speeds up the process of producing stereo content compared to rendering two frames. On average, for all scenes used for our experiments, the morphing of one frame in resolution 2048×1024 takes around 7 ms.

Second, our method achieves quality similar to the ground truth, while "Simple Grid" approach falls short in doing so for complex details (spikes, ghosting). In particular, when comparing to the trivial approach (Figure 8.9) of mapping individual pixels and filling the holes using pull-push, the quality is worse and the performance

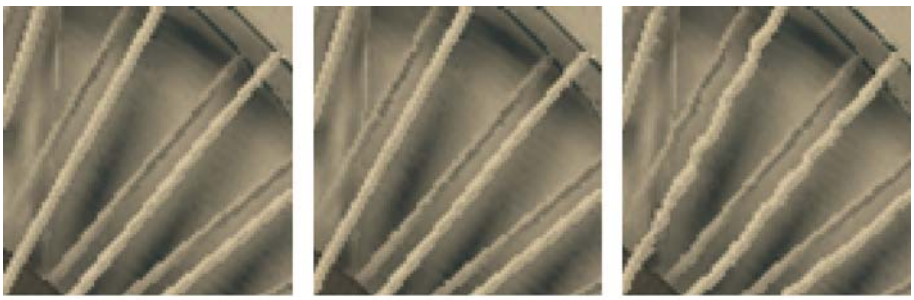


Figure 8.6: Decreasing (resp. increasing) the threshold generates a higher (resp. lower) grid resolution, therefore lower (resp. higher) speed but also higher (resp. lower) quality.

is three times lower. This is easy to see, as warping a grid of vertices which form a small subset of all pixels in the image is obviously faster than warping all pixels. This performance difference underlines the importance of supporting modern fine grained parallelism (i. e. gathering) over straightforward approaches which require scattering.

Third, we see how the use of multiple images avoids disocclusions and improves the quality by comparing the two rightmost columns. This is most visible for the "Antenna" scene in the second row, where the thin features are stretched across disocclusions when using only a single image. As our approach is orthogonal to the used surface representation, we can apply our technique also directly to iso-surface ray-casting [Levoy 1988] (last row).

8.3.2 Adaptation Quality

Further, we seek to illustrate the influence of the subdivision threshold by keeping all parameters fixed and varying only this threshold. In Figure 8.6, we show high, medium and low-quality thresholds, the respective subdivision, as well as some details that represent typical problems also encountered with a trivial approach (Figure 8.9).

8.3.3 Analysis

In Figure 8.7, the variation of performance over time for the reference, "Simple Grid" and the new method is plotted for the "Crane" scene. We see, how the new method has varying efficiency over time. This is because the adaptation creates a varying number of quads in our grid. However, it is almost never slower than previous work, at much higher quality, as discussed in the previous paragraph. Tighter bounding of this time interval is desirable in interactive applications such as games and remains future work.

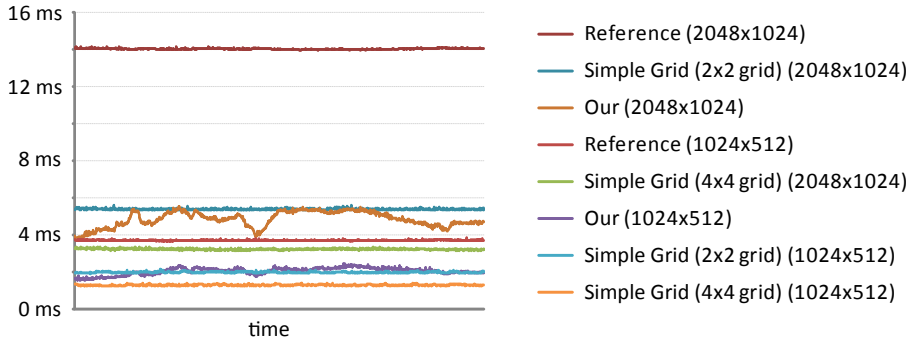


Figure 8.7: Variation of performance over time for several different strategies. Although our performance varies due to the adaptivity, it is nearly as high as for method designed for temporal upsampling but at a quality comparable to the reference solution.

8.4 Discussion

Similar to many other upsampling methods (e. g., the one presented in Chapter 4) this approach is limited to non-transparent surfaces. We do not account for view dependent-effects such as specular highlights.

The improvement when using previous frames (Section 8.2.5) depends on the camera path. In case of camera movement in the plane to which the eye axis is normal, no additional information is won, but such movements are less likely than e. g., human walking animations. Put in another way, human eyes are placed horizontal to each other and not vertically because of the movements performed by humans [Ross 1974]. In future work, more advanced view selection techniques are worth investigating.

Lacking a suitable output device, we were not able to test our method for generating more than two views out of one. However, the time-benefit of image-based upsampling would be even more pronounced. Also, we envision upsampling in time as well as in stereo and other image-based re-use e. g., for anti-aliasing or motion blur.

8.5 Conclusions

In this chapter, we described an approach to upsample a stream of monocular images with depth information to a stereo-image streams, exploiting modern GPUs and human perception. We demonstrated its application to a number of problems, in which the approach drastically reduces the rendering time compared to rendering an image pair. The approach is independent of the underlying surface representation and can be easily integrated into existing software as a post-process to deliver high-quality stereo-image pairs. Recently, our approach has been improved by Bowles et al. [2012], also mentioned in Section 4.5.

In future work we would like to address stereo view synthesis of images with transparent surfaces, such as volume rendering with full transfer functions, or clouds and steam in interactive applications such as games. Besides technical improvements

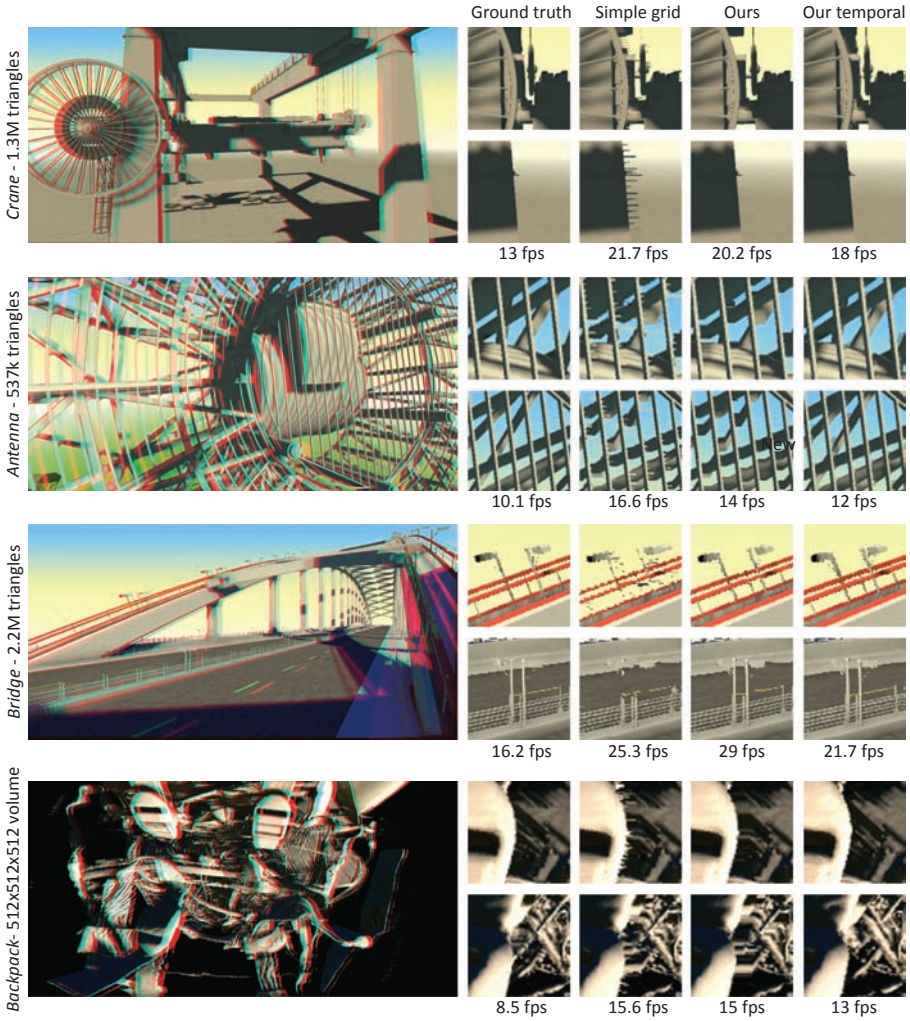


Figure 8.8: Results produced by our algorithm (Left) for different scenes in resolution 2048×1024, presented in anaglyph stereo. On the right, we show scene details computed using four different approaches: Ground truth; "Simple Grid" method; Ours using only single images; Ours using multiple images. We achieve similar quality to ground truth at a performance similar to "Simple Grid" method (see the fps insets).

to produce stereo, many questions of stereo perception and stereo content control are not yet answered, including the depiction of specularities and transparency as well as the disagreement of lens accommodation and other stereo cues or how depth of field and stereo should be combined. In this contexts, recently in [Templin et al., 2012], a new method for specular highlights rendering was presented. Instead of reproducing highlights as they would be captured by a two-camera setup, we proposed to render all reflections with small disparities (so-called *microdisparities*). This improves comfort comparing to a physically-correct rendering and enhances material depiction comparing to highlights located at the depth of objects. In order to render highlights with microdisparities an image-based warping technique is required. For this purpose the

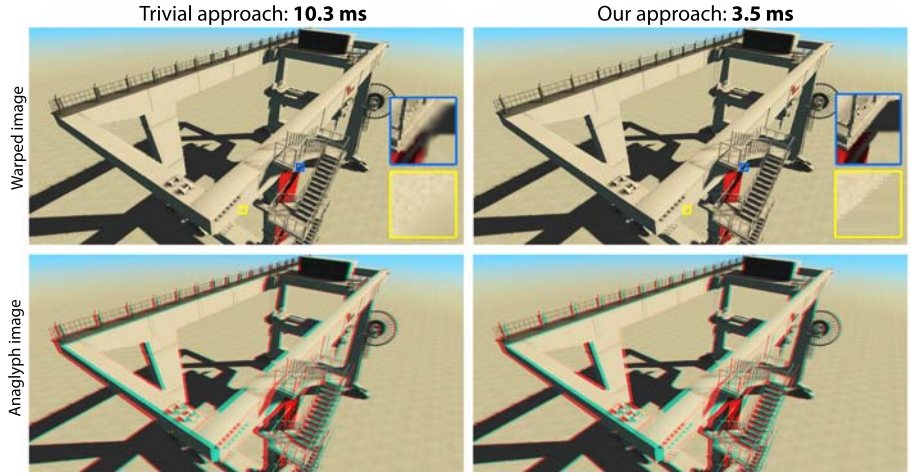


Figure 8.9: Using pixel-wise re-projection (trivial approach, below-reference) results in many holes, that have to be filled using pull-push which leads to blur. At the same time, the performance is approximately three times lower than for our approach.

technique proposed in this chapter could be used.

In the future, we plan to exploit the fact that humans have a dominant and a recessive eye which allows for lowering the quality of the left or the right image without introducing visible artifacts [Stelmach, Tam and Meegan 1999]. One could display a perfect rendering for the dominant eye and a warped (maybe even blurred) imperfect rendering for the recessive eye, which would further improve efficiency of stereo image computation.

9

Summary

In this part, we summarize all contributions that are presented in this dissertation and propose directions for future work.

9.1 Conclusions

The continuous demand for faithfully reproducing the real world and for great visual sensation, has forced computer graphics researchers, artists, display and capture-device manufacturers to constantly improve their techniques and products. The role of today's cinema, television, video games or visualization techniques is not only to illustrate but also to immerse the viewer into a crafted or captured world. Although it may seem that mostly entertainment applications benefit from the advances, realistic and convincing content presentation finds great exploitation in fields such as medical imaging or product design. In this dissertation, we concentrated on display devices whose quality is of high importance to today's visualization techniques.

Usually, the improvement of display devices is achieved by the means of better or new hardware designs. However, as shown in this dissertation, the quality of reproduced content should be always considered in the context of *perceived quality*, hence, it is important to take properties of the HVS into account. In the end, the result is not what one can reproduce on screen but rather the mental image created by the human brain. Therefore, instead of considering new display designs, we explicitly employed perceptual effects to improve the quality of display devices, often beyond their physical capabilities. By capitalizing on various aspects of the human visual system, display qualities have, at least perceptually, been significantly enhanced in a stable and persistent way. Similar enhancements could often only be achieved by improving physical parameters of displays, which might be impossible without fundamental design changes in the existing display technology and clearly may lead to an overall higher display cost. Instead, we studied apparent improvements that are enabled by properties of the HVS.

Most of the here-presented techniques achieve their goals by means of relatively simple image processing operations, which often rely on skillful signal manipulation. We interleaved frames in the temporal domain and combined high-quality sharpened and low-quality blurred frames on high-refresh rate displays to reduce the perceptual hold-type blur and to improve apparent motion smoothness. Also, we reduced rendering cost as blurred low-quality images are derived from high-quality images without causing any perceivable quality degradation. Similar in spirit, the principles of temporal

signal integration in the retina can be used to enhance apparent resolution, where high frequency information that cannot be shown directly on screen is included in many consecutive frames in form of aliasing which later canceled by the HVS produces impression of looking at higher resolution content. More complex, but implemented as simple image processing operations, disparity models allow for a prediction of disparity perception and effective disparity and luminance-disparity manipulations control become possible. Finally, the Cornsweet gradient profile inserted across depth discontinuities can enhance perceived depth. Hereby, we can produce stereo images that look ordinary when viewed without glasses but create a stereo impression when special equipment is used. The simplicity of these techniques makes them very efficient but can also be integrated within the display panels of the future, in form of small computational units, which would be an interesting alternative to standard hardware solutions.

In this dissertation, we successfully showed that taking into account human perception during display stage can indeed improve perceived quality significantly. Our techniques considered only a few properties of HVS leaving many of them still not explored in this context. Therefore, we believe, that further research taking into account other aspects of human vision as well as new display designs will lead not only to new software techniques for perceived quality enhancement but also to new display designs.

9.2 Future Work

We believe that the idea of exploiting properties of the HVS in the context of image quality enhancement is appealing and will stimulate other researchers to pursue further investigations and there are many directions for future work. Here, we list a few of them.

In our work quality dimensions were considered mostly separately. In the future work, interactions between them could be investigated and possible advantages of their interplay could be used for further improvements. With our joint luminance-disparity manipulation, where we investigated how luminance contrast can enhance perceived depth, we made a first step in this direction. Interestingly, their relationship is mutual and disparity can also enhance the luminance-contrast perception. Hence, we could find a good balance between visible luminance contrast and disparity, when processing them simultaneously. Similarly, interactions of qualities, such as brightness, contrast, spatial or temporal resolution with depth might be promising. The main observation here is that the HVS tolerates differences between left and right views to some extent. In the same spirit, as we split high frequency information among many consecutive frames in our apparent resolution enhancement method, one could think of splitting information between the left and right views. Such an approach could be beneficial not only in the context of resolution enhancement but also for high dynamic range imaging.

In the future, considering not only different quality dimensions but also different modalities could bring a significant improvement of the viewing experience. However, so far, not much research exists in this direction. For example, it is possible that audio has a big influence on the human depth sensation. A skillful audio signal adjustment could potentially be used to compensate for current 3D stereo display limitations, where the allowed depth range is to a great extent limited by the accommodation-vergence

conflict.

In this dissertation, we focused on techniques that do not attempt to modify current displays designs. Instead, we provided software solutions for off-the-shelf displays, which can achieve apparent quality enhancements. However, we believe that making a step back and considering display design under the light of the contributions in this dissertation may result in new and better solutions. An interesting technique has been recently proposed by Berthouzoz and Fattal [2012]. Instead of moving images in order to enhance perceived resolution as we proposed, the authors introduce a periodic movement of a display device. Hereby the perceived resolution is increased using a similar approach to ours for static content.

Apart from combining our techniques with hardware development one can consider designing new software techniques that would take advantage of the data provided by additional devices. For instance, eye trackers, which receive an increasing amount of attention. These can provide not only 2D information about the point of gaze in the display plane, but also the exact vergence location in 3D space. Such information can potentially be used to improve viewing comfort in the context of 3D stereo displays or it could serve as an additional information for our resolution enhancement or temporal upsampling techniques which currently assume that the observer closely follows the movement of dynamic objects on the screen.

Furthermore, we proposed techniques that use display features for different purpose than what they were initially designed for. For example, we used high-framerate displays that were primarily designed for low-cost 3D stereo and showed that by taking advantage of the framerate, an apparent resolution enhancement can be achieved. This matches the current trend of computational display techniques, which, by analogy to computational photography, modify standard display devices or provide additional software in order to achieve new functionalities. This is a very appealing concept for future customers who will be able to extend the functionality of new displays according to their needs. For example, one could consider new uses of multi-view autostereoscopic displays. Instead of reproducing 3D stereo and parallax, such displays could be potentially used for better material reproduction.

Material reproduction is on its own an exiting direction for the future work, which recently started getting much attention. People do not only consider displaying different materials with spatially varying properties but they also fabricate them using modern 3D printers. It is, however, still a tedious task to reproduce materials that exhibit properties that closely match the real world exemplars. The biggest problem is the curse of dimensionality. The HVS judges material appearance based on many factors such as brightness, illumination, 3D information, and different views. Understanding how this information is combined might be crucial for a faithful material reproduction.

Bibliography (Own work)

BANTERLE, F., ARTUSI, A., AYDIN, T., DIDYK, P., EISEMANN, E., GUTIERREZ, D., MANTIUK, R. AND MYSZKOWSKI, K. (2011): Multidimensional Image Retargeting. In *ACM SIGGRAPH Asia 2011 Courses* 2

BANTERLE, F., ARTUSI, A., AYDIN, T., DIDYK, P., EISEMANN, E., GUTIERREZ, D., MANTIUK, R. AND RITSCHEL, T. (2012): Mapping images to target devices: spatial, temporal, stereo, tone, and color. In *Eurographics 2012 Tutorials* 2

DIDYK, P., EISEMANN, E., RITSCHEL, T., MYSZKOWSKI, K. AND SEIDEL, H.-P. (2010a): Apparent Display Resolution Enhancement for Moving Images. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2010, Los Angeles)*, 29 (4), 113:1–113:8 3

DIDYK, P., EISEMANN, E., RITSCHEL, T., MYSZKOWSKI, K. AND SEIDEL, H.-P. (2010b): Perceptually-motivated Real-time Temporal Upsampling of 3D Content for High-refresh-rate Displays. *Computer Graphics Forum (Proc. Eurographics)*, 29 (2), 713–722 3

DIDYK, P., EISEMANN, E., RITSCHEL, T., MYSZKOWSKI, K. AND SEIDEL, H.-P. (2011a): Apparent Display Resolution Enhancement for Moving Images., WO Patent Application WO/2011/135,052 3

DIDYK, P., MANTIUK, R., HEIN, M. AND SEIDEL, H.-P. (2008): Enhancement of Bright Video Features for HDR Displays. *Computer Graphics Forum (Proceedings Eurographics Symposium on Rendering 2008, Sarajevo, Bosnia and Herzegovina)*, 27 (4), 1265–1274 120

DIDYK, P., RITSCHEL, T., EISEMANN, E. AND MYSZKOWSKI, K. (2012a): CHAP. EXCEEDING PHYSICAL LIMITATIONS: APPARENT DISPLAY QUALITIES. IN Perceptual Digital Imaging: Methods and Applications. 2

DIDYK, P., RITSCHEL, T., EISEMANN, E., MYSZKOWSKI, K. AND SEIDEL, H.-P. (2010c): Adaptive Image-space Stereo View Synthesis. In *Vision, Modeling and Visualization Workshop*, 299–306 4

DIDYK, P., RITSCHEL, T., EISEMANN, E., MYSZKOWSKI, K. AND SEIDEL, H.-P. (2011b): A Perceptual Model for Disparity. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2011, Vancouver)*, 30 (4), 96:1–96:10 3

DIDYK, P., RITSCHEL, T., EISEMANN, E., MYSZKOWSKI, K. AND SEIDEL, H.-P. (2012b): Apparent Stereo: The Cornsweet Illusion Can Enhance Perceived Depth. In *Human Vision and Electronic Imaging XVII, IS&TSPIE's Symposium on Electronic Imaging*, 1–12 3

DIDYK, P., RITSCHEL, T., EISEMANN, E., MYSZKOWSKI, K. AND SEIDEL, H.-P. (2012c): Methods and device for processing digital stereo image content., WO Patent Application 3

DIDYK, P., RITSCHEL, T., EISEMANN, E., SEIDEL, H.-P., MYSZKOWSKI, K. AND MATUSIK, W. (2012d): A Luminance-Contrast-Aware Disparity Model and Applications. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2012, Singapore)*, 31 (5), (Conditionally accepted) 4

TEMPLIN, K., DIDYK, P., RITSCHEL, T., EISEMANN, E., MYSZKOWSKI, K. AND SEIDEL, H.-P. (2011): Apparent Resolution Enhancement for Animations. In *27th Spring Conference on Computer Graphics*, 85–92 73, 74

TEMPLIN, K., DIDYK, P., RITSCHEL, T., EISEMANN, E., MYSZKOWSKI, K. AND SEIDEL, H.-P. (2012): Highlight Microdisparity for Improved Gloss Depiction. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2012, Los Angeles, CA)*, 31 (4), 1–5 131

Bibliography

ADELSON, E. AND BERGEN, J. (1991): The plenoptic function and the elements of early vision. *Computational models of visual processing*, 1, 3–20 32

ALLEN, W. AND ULICHNEY, R. (2005): Wobulation: Doubling the addressed resolution of projection displays. In *Proceedings of the Symposium Digest of Technical Papers (SID)* Volume 47.4,, 1514–1517 39

ANSTIS, S., HOWARD, I. AND ROGERS, B. (1978): A Craik-O'Brien-Cornsweet illusion for visual depth. *Vision Research*, 18 (2), 213–217 18, 95

ARTAMONOV, O. (2004): X-bit's Guide: Contemporary LCD Monitor Parameters and Characteristics. Page 11. `http://www.xbitlabs.com/articles/monitors/display/lcd-guide\protect_11.html` 36

BARTEN, P. G. J. (1989): The Square Root Integral (SQRI): A New Metric to Describe the Effect of Various Display Parameters on Perceived Image Quality. In *Proceedings of SPIE* Volume 1077,, 73–82 20, 89

BEIER, T. AND NEELY, S. (1992): Feature-based image metamorphosis. *Computer Graphics*, 26 (2), 35–42 31

BENOIT, A., CALLET, P. L., CAMPISI, P. AND COUSSEAU, R. (2008): Quality Assessment of Stereoscopic Images. *EURASIP Journal on Image and Video Processing* 2008 (659024) 42

BERTHOUZOZ, F. AND FATTAL, R. (2012): Resolution enhancement by vibrating displays. *ACM Transactions on Graphics*, 31 (2), 15:1–15:14 74, 135

BEZERRA, H., EISEMANN, E., DÉCORET, X. AND THOLLOT, J. (2008): 3D Dynamic Grouping for Guided Stylization. In *NPAR '08: Proceedings of the 6th International Symposium on Non-photorealistic Animation and Rendering*, 89–95 40

BIJL, P., SCHUTTE, K. AND HOGERVORST, M. A. (2006): Applicability of TOD, MTDP, MRT and DMRT for dynamic image enhancement techniques. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* Volume 6207, 38

BLAKEMORE, C. (1970): The range and scope of binocular depth discrimination in man. *The Journal of Physiology*, 211 (3), 599–622 17, 78, 95

BOWLES, H., MITCHELL, K., SUMNER, R., MOORE, J. AND GROSS, M. (2012): Iterative Image Warping. In *Computer Graphics Forum* Volume 31, The Eurographics Association and Blackwell Publishing Ltd., 237–246 59, 130

BRADLEY, A. AND OHZAWA, I. (1986): A comparison of contrast detection and discrimination. *Vison Research*, 26 (6), 991–997, ISSN 0042–6989 80

BRADSHAW, M. F. AND ROGERS, B. J. (1999): Sensitivity to horizontal and vertical corrugations defined by binocular disparity. *Vision Research*, 39 (18), 3049–56, ISSN 0042–6989 16, 17, 76, 77, 78, 79, 81, 95, 96

BROOKES, A. AND STEVENS, K. (1989): The analogy between stereo depth and brightness. *Perception*, 18 (5), 601–614 16, 17

BRUCKNER, S. AND GRÖLLER, E. (2007): Enhancing Depth-Perception with Flexible Volumetric Halos. *IEEE Transactions on Visualization and Computer Graphics*, 13 (6), 1344–51 40

BURR, D. C. (1979): Acuity for apparent vernier offset. *Vision Research*, 19 (7), 835 – 837 13

BURR, D. (1981): Temporal summation of moving images by the human visual system. In *Proceedings of the Royal Society of London* Volume B 211,, 321–339 10, 45

BURT, P. AND ADELSON, E. (1983): The Laplacian pyramid as a compact image code. *IEEE Transactions Communic.* 31 (4), 532–540 83, 91

CALABRIA, A. AND FAIRCHILD, M. (2003): Perceived image contrast and observer preference I: The effects of lightness, chroma, and sharpness manipulations on contrast perception. *Journal of Imaging Science and Technology*, 47, 479–493 11

CHEN, H., KIM, S.-S., LEE, S.-H., KWON, O.-J. AND SUNG, J.-H. (2005): Nonlinearity compensated smooth frame insertion for motion-blur reduction in LCD. In *Proceedings of Multimedia Signal Processing, 2005 IEEE 7th Workshop on*, 1–4 28, 46, 49

CHEN, S. E. AND WILLIAMS, L. (1993): View interpolation for image synthesis. In *Proceedings of SIGGRAPH*, 279–288 32

COLEMAN, T. F. AND LI, Y. (1996): A Reflective Newton Method for Minimizing a Quadratic Function Subject to Bounds on some of the Variables. *SIAM Journal on Optimization*, 6 (4), 1040–1058 66

CORMACK, L., STEVENSON, S. AND SCHOR, C. (1991): Interocular correlation, luminance contrast and cyclopean processing. *Vison Research*, 31 (12), 2195–2207 20, 21, 88, 89, 97

COUTANT, B. AND WESTHEIMER, G. (1993): Population distribution of stereoscopic ability. *Ophthalmic and Physiological Optics*, 13 (1), 3–7, ISSN 1475–1313 78, 95, 108

CURCIO, C. A., SLOAN, K. R., KALINA, R. E. AND HENDRICKSON, A. E. (1990): Human photoreceptor topography. *The Journal of Comparative Neurology*, 292 (4), 497–523 6, 7, 8, 64

CUTTING, J. AND VISHTON, P. (1995): Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In EPSTEIN, W. AND ROGERS, S., EDITORS: *Perception of Space and Motion (Handbook Of Perception And Cognition)*, 69–117 14, 15, 24, 76, 92

DALY, S. (1998): Engineering observations from spatiovelocity and spatiotemporal visual models. In *Human Vision and Electronic Imaging III* Volume 3299,, 180–191 10, 52

DALY, S. (1993): The visible differences predictor: an algorithm for the assessment of image fidelity. *Digital images and human vision*, 4, 179 19, 40, 76, 83, 96

DAMERA-VENKATA, N. AND CHANG, N. L. (2009): Display Supersampling. *ACM Transactions on Graphics*, 28 (1), 9:1–9:19 39

DEERING, M. F. (2005): A photon accurate model of the human eye. *ACM Transactions on Graphics (Proceedings of SIGRAPH 2005)*, 24 (3), 649–658 7

DIXON, N. F. AND SPITZ, L. (1980): The detection of auditory visual desynchrony. *Perception* 9 (6) 29

DOMONKOS, B., EGRI, A., FÓRIS, T., SZIRMAY-KALOS, L. AND TAMÁS, J. (2007): Isosurface Ray-casting for Autostereoscopic Displays. In *WSCG, Short Papers* 35

FAHLE, M. AND POGGIO, T. (1981): Visual Hyperacuity: Spatiotemporal Interpolation in Human Vision. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 213 (1193), 451–477 13

FARBMAN, Z., FATTAL, R., LISCHINSKI, D. AND SZELISKI, R. (2008): Edge-Preserving Decompositions for Multi-Scale Tone and Detail Manipulation. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 27 (3), 67:1–67:10 39, 40

FATTAL, R., LISCHINSKI, D. AND WERMAN, M. (2002): Gradient Domain High Dynamic Range Compression. *ACM Transactions on Graphics*, 21 (3), 249–256 39, 41

FENG, X.-F. (2006): LCD motion blur analysis, perception, and reduction using synchronized backlight flashing. In *Human Vision and Electronic Imaging XI*, M1–14 11, 28

FUJIBAYASHI, A. AND BOON, C. S. (2008): A Masking model for motion sharpening phenomenon in video sequences. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E91-A (6), 1408–1415 13

GOREA, A. AND TYLER, C. W. (1986): New look at Bloch's law for contrast. *Journal of the Optical Society of America A*, 3 (1), 52–61 8, 9

GRAHAM, C. (1965): Vision and visual perception. 9

GREEN, C. S. AND BAVELIER, D. (2003): Action video game modifies visual selective attention. *Nature*, 423 (6939), 534–537, ISSN 0028–0836 53

HARA, Z. AND SHIRAMATSU, N. (2000): Improvement in the picture quality of moving pictures for matrix displays. *Journal of the Society for Information Display*, 8 (2), 129–137 13

HARTLEY, R. I. AND ZISSERMAN, A. (2000): Multiple View Geometry in Computer Vision. 34

HATEREN, J. H. VAN (2005): A cellular and molecular model of response kinetics and adaptation in primate cones and horizontal cells. *Journal of Vision*, 5 (4), 331–347 8, 63

HATEREN, J. H. VAN (2006): Encoding of high dynamic range video with a model of human cones. *ACM Transactions on Graphics*, 25 (4), 1380–1399 63

HECKMANN, T. AND SCHOR, C. M. (1989): Is edge information for stereoacuity spatially channeled? *Vison Research*, 29 (5), 593–607 20, 22

HEINZLE, S., GREISEN, P., GALLUP, D., CHEN, C., SANER, D., SMOLIC, A., BURG, A., MATUSIK, W. AND GROSS, M. (2011): Computational stereo camera system with programmable control loop. *ACM Transactions on Graphics*, 30, 94:1–94:10 41

HELD, R. AND BANKS, M. (2008): Misperceptions in stereoscopic displays: A vision science perspective. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization* ACM, 23–32 43

HELD, R., COOPER, E., O'BRIEN, J. AND BANKS, M. (2010): Using blur to affect perceived distance and size. *ACM Transactions on Graphics*, 29 (2), 19:1–19:16 24

HERZOG, R., EISEMANN, E., MYSZKOWSKI, K. AND SEIDEL, H.-P. (2010): Spatio-Temporal Upsampling on the GPU. In *Proceedings of ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 91–98 33, 35

HESS, R., KINGDOM, F. AND ZIEGLER, L. (1999): On the relationship between the spatial channels for luminance and disparity processing. *Vison Research*, 39 (3), 559–68 21, 88, 95

HOFFMAN, D., GIRSHICK, A., AKELEY, K. AND BANKS, M. (2008): Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of vision*, 8 (3), 1–30 24, 78

HOWARD, I. P. AND ROGERS, B. J. (2002): Seeing in Depth. Volume 2: Depth Perception, 15, 16, 19, 75, 76, 77, 78, 81, 86, 95, 96

IOANNOU, G., ROGERS, B., BRADSHAW, M. AND GLENNERSTER, A. (1993): Threshold and supra-threshold senssitivity functions for stereoscopic surfaces. *Investigative Ophthalmology & Visual Science*, 34, 1186 16, 77

ISHIHARA, S. (1987): Test for colour-blindness. 80

JANSSEN, R. (2001): Computational Image Quality. 11

JONES, G., LEE, D., HOLLIMAN, N. AND EZRA, D. (2001): Controlling perceived depth in stereoscopic images. *Proceedings of SPIE*, 4297, 42–53 41, 76, 104

JULESZ, B. (1964): Binocular Depth Perception without Familiarity Cues: Random-dot stereo images with controlled spatial and temporal properties clarify problems in stereopsis. *Science*, 145 (3630), 356–362 19

JULESZ, B. (1971): Foundations of Cyclopean Perception. 16, 95

KALLONIATIS, M. AND LUU, C. (2009): Temporal Resolution. *http://webvision.med.utah.edu/temporal.html* 9, 36, 68

KINGDOM, F. AND MOULDEN, B. (1988): Border Effects on Brightness: A Rreview of Findings, Models and Issues. *Spatial Vision*, 3 (4), 225–62 18

KINGDOM, F. AND SIMMONS, D. (2000): The relationship between colour vision and stereoscopic depth perception. *Journal of Society for 3-D Broadcasting and Imaging*, 1, 10–19 96

KLOMPENHOUWER, M. A. AND HAAN, G. DE (2003): Subpixel image scaling for color-matrix displays. *Journal of the Society for Information Display*, 11 (1), 99–108 38

KLOMPENHOUWER, M. A. AND VELTHOVEN, L. J. (2004): Motion blur reduction for liquid crystal displays: Motion-compensated inverse filtering. In *Proceedings of SPIE* 11, 29

KNORR, S., KUNTER, M. AND SIKORA, T. (2008): Stereoscopic 3D from 2D Video with Super-Resolution Capability. *Signal Processing: Image Communication*, Vol. 23 (9), 665–676 34

KOOI, F. AND TOET, A. (2004): Visual comfort of binocular and 3D displays. *Displays*, 25 (2), 99–108 24

KOPF, J., UYTTENDAELE, M., DEUSSEN, O. AND COHEN, M. (2007): Capturing and Viewing Gigapixel Images. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2007)* 26 (3) 37

KRAPELS, K., DRIGGERS, R. G. AND TEANEY, B. (2005): Target-acquisition performance in undersampled infrared imagers: static imagery to motion video. *Applied Optics*, 44 (33), 7055–7061 38

KRAUZLIS, R. AND LISBERGER, S. (1994): Temporal properties of visual motion signals for the initiation of smooth pursuit eye movements in monkeys. *Journal of Neurophysiology*, 72 (1), 150–162 10

KRAWCZYK, G., MYSZKOWSKI, K. AND SEIDEL, H.-P. (2007): Contrast Restoration by Adaptive Countershading. *Computer Graphics Forum*, 26 (3), 581–590, ISSN 0167–7055 18, 40, 110

KURITA, T. (2001): Moving picture quality improvement for hold-type AM-LCDs. In *Society for Information Display (SID) '01*, 986–989 29

LAIRD, J., ROSEN, M., PELZ, J., MONTAG, E. AND DALY, S. (2006): Spatio-velocity CSF as a function of retinal velocity using unstabilized stimuli. In *Human Vision and Electronic Imaging XI* Volume 6057,, 32–43 10

LAMBOOIJ, M., IJSSELSTEIJN, W., FORTUIN, M. AND HEYNDERICKX, I. (2009): Visual Discomfort and Visual Fatigue of Stereoscopic Displays: A Review. *Journal of Imaging Science and Technology*, 53 (3), 030201:1–12, ISSN 10623701 41, 76, 78

LANG, M., HORNUNG, A., WANG, O., POULAKOS, S., SMOLIC, A. AND GROSS, M. (2010): Nonlinear disparity mapping for stereoscopic 3D. *ACM Transactions on Graphics (Proceedings SIGGRAPH)*, 29 (4), 751–760 35, 42, 75, 76, 102, 114

LANGE, H. DE (1958): Research into the Dynamic Nature of the Human Fovea - Cortex Systems with Intermittent and Modulated Light. I. Attenuation Characteristics with White and Colored Light. *Journal of the Optical Society of America*, 48 (11), 777–783 9

LEE, B. AND ROGERS, B. (1997): Disparity modulation sensitivity for narrow-band-filtered stereograms. *Vison Research*, 37 (13), 1769–77 21, 22, 88, 95

LEE, S., SHIOIRI, S. AND YAGUCHI, H. (2007): Stereo channels with different temporal frequency tunings. *Vison Research*, 47 (3), 289–97 95

LEE, S., EISEMANN, E. AND SEIDEL, H.-P. (2009): Depth-of-field rendering with multiview synthesis. *ACM Transactions on Graphics (Proceedings of of SIGGRAPH Asia)* 28 (5) 104

LEE, S., CHWA, K., HAHN, J. AND SHIN, S. (1996): Image morphing using deformation techniques. *Journal of Visualization and Computer Animation*, 7 (1), 3–23 31

LEGGE, G. AND GU, Y. (1989): Stereopsis and contrast. *Vison Research*, 29 (8), 989–1004 20, 22, 88, 89

LEVITIN, D. J., MACLEAN, K., MATHEWS, M., CHU, L. AND JENSEN, E. (2000): The perception of cross-modal simultaneity. In *Proceedings of CASYS*, 323–329 29

LEVOY, M. (1988): Display of Surfaces from Volume Data. *IEEE Comput. Graph. Appl.* 8 (3), 29–37 129

LI, Y., SHARAN, L. AND ADELSON, E. H. (2005): Compressing and companding high dynamic range images with subband architectures. *ACM Transactions on Graphics*, 24 (3), 836–844 40

LIN, W., GAI, Y. AND KASSIM, A. (2006): Perceptual impact of edge sharpness in images. *Vision, Image and Signal Processing, IEE Proceedings*, 152 (2), 215–223 11

LIU, F., GLEICHER, M., JIN, H. AND AGARWALA, A. (2009): Content-Preserving Warps for 3D Video Stabilization. *ACM Transactions on Graphics (Proceedings SIGGRAPH)* 28 31

LIVINGSTONE, M. (2002): Vision and Art: The Biology of Seeing. 75, 97

LUBIN, J. (1995): A visual discrimination model for imaging system design and development. In PELI, E., EDITOR: *Vision models for target detection and recognition*, 245–283 40, 76, 85, 87, 96

LUEBKE, D., WATSON, B., COHEN, J. D., REDDY, M. AND VARSHNEY, A. (2002): Level of Detail for 3D Graphics., ISBN 1558608389 29

LUFT, T., COLDITZ, C. AND DEUSSEN, O. (2006): Image enhancement by unsharp masking the depth buffer. *ACM Transactions on Graphics*, 25 (3), 1206–13 40

LUNN, P. AND MORGAN, M. (1995): The analogy between stereo depth and brightness: a reexamination. *Perception*, 24 (8), 901–4 16, 19

MACKAY, D. M. (1973): Lateral Interaction between Neural Channels sensitive to Texture Density? *Nature*, 245 (5421), 159–161 18

MAHAJAN, D., HUANG, F.-C., MATUSIK, W., RAMAMOORTHI, R. AND BELHUMEUR, P. (2009): Moving Gradients: A Path-Based Method for Plausible Image Interpolation. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 28 (3), 42:1–42:11 32, 34

MAJUMDER, A. AND BROWN, M. S. (2007): Practical Multi-Projector Display Design. 37

MÄKELÄ, P., ROVAMO, J. AND WHITAKER, D. (1994): Effects of luminance and external temporal noise on flicker sensitivity as a function of stimulus size at various eccentricities. *Vision Research*, 34 (15), 1981–91 9, 68

MANTIUK, R., MYSZKOWSKI, K. AND SEIDEL, H. (2006): A perceptual framework for contrast processing of high dynamic range images. *ACM Transactions on Applied Perception*, 3 (3), 286–308, ISSN 1544–3558 39, 40, 77, 83, 84, 90, 91, 100, 102

MANTIUK, R., DALY, S. AND KEROFSKY, L. (2008): Display adaptive tone mapping. *ACM Transactions on Graphics* 27 (3) 39, 103

MARK, W. R., MCMILLAN, L. AND BISHOP, G. (1997): Post-Rendering 3D Warping. In *Proceedings of ACM I3D*, 7–16 33

MARR, D. AND POGGIO, T. (1979): A computational theory of human stereo vision. *Proceedings of the Royal Society London Ser. B*, 204, 301–28 22

MARSHALL, J., BURBECK, C., ARIELY, D., ROLLAND, J. AND MARTIN, K. (1996): Occlusion edge blur: a cue to relative visual depth. *Journal of the Optical Society of America A*, 13 (4), 681–688 23

MARTINEZ-CONDE, S., MACKNIK, S. L. AND HUBEL, D. H. (2004): The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience*, 5 (3), 229–239 10

MATHER, G. AND SMITH, D. (2002): Blur discrimination and its relation to blur-mediated depth perception. *Perception*, 31 (10), 1211–1220 23

MATUSIK, W. AND PFISTER, H. (2004): 3DTV: A scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. *ACM Transactions on Graphics*, 23 (3), 814–824 75, 99

MCKEE, S. P. AND TAYLOR, D. G. (1984): Discrimination of time: comparison of foveal and peripheral sensitivity. *Journal of the Optical Society of America A*, 1 (6), 620–628 9, 68

MCMILLAN, L. AND BISHOP, G. (1995): Plenoptic modeling: An image-based rendering system. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques* ACM, 39–46 32

MEESTERS, L., IJSSELSTEIJN, W. AND SEUNTIENS, P. (2004): A survey of perceptual evaluations and requirements of three-dimensional TV. *IEEE Transactions on Circuits and Systems for Video Technology*, 14 (3), 381 – 391, ISSN 1558–2205 42, 43, 97

MENDIBURU, B. (2009): 3D movie making: stereoscopic digital cinema from script to screen. 99

MESSING, D. S. AND KEROFSKY, L. J. (2006): Using optimal rendering to visually mask defective subpixels. In *Human Vision and Electr. Imaging XI* Volume 6057,, 236–247 38

MEYER, Q., EISENACHER, C., STAMMINGER, M. AND DACHSBACHER, C. (2009): Data-Parallel Hierarchical Link Creation for Radiosity. In *Proceedings of EPGV*, 65–69 125

MITCHELL, D. P. AND NETRAVALI, A. N. (1988): Reconstruction filters in computer-graphics. *Proceedings of SIGGRAPH*, 22 (4), 221–228 37, 70, 71

MORLAND, D. V. (1976): Computer-generated stereograms: a new dimension for the graphic arts. *SIGGRAPH Comput. Graph.* 10 (2), 19–24 34

NEHAB, D. F., SANDER, P. V., LAWRENCE, J., TATARCHUK, N. AND ISIDORO, J. (2007): Accelerating real-time shading with reverse reprojection caching. In *Graphics Hardware*, 25–35 33

NISHINA, S. (2003): Spatio-temporal dynamics of depth propagation on uniform region. *Vison Research*, 43 (24), 2493–2503 95

OLIVA, A., TORRALBA, A. AND SCHYNS, P. G. (2006): Hybrid images. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 25, 527–532, ISSN 0730–0301 100, 117

ONURAL, L., SIKORA, T., OSTERMANN, J., SMOLIC, A., CIVANLAR, M. R. AND WATSON, J. (2006): Assessment of 3DTV Technologies. In *NAB Broadcast Engineering*, 456–467 80, 99

OSKAM, T., HORNUNG, A., BOWLES, H., MITCHELL, K. AND GROSS, M. (2011): OSCAM - optimized stereoscopic camera control for interactive 3D. *ACM Transactions on Graphics*, 30, 189:1–189:8 41, 103, 104

O'SULLIVAN, C. AND DINGLIANA, J. (2001): Collisions and Perception. *ACM Transactions on Graphics*, 20, 151–168 29

PAJAK, D., HERZOG, R., EISEMANN, E., MYSZKOWSKI, K. AND SEIDEL, H.-P. (2011): Scalable Remote Rendering with Depth and Motion-flow Augmented Streaming. *Computer Graphics Forum*, 30 (2), 415–424 33

PALMER, S. E. (1999): Vision Science: Photons to Phenomenology. 14, 15, 75, 95

PAN, H., FENG, X.-F. AND DALY, S. (2005): LCD motion blur modeling and analysis. In *Proceedings of ICIP*, 21–24 11, 28

PARK, S., PARK, M. AND KANG, M. (2003): Super-resolution image reconstruction: A technical overview. *IEEE Signal Processing Magazine*, 20 (3), 21–36 37

PLATT, J. (2000): Optimal filtering for patterned displays. *Signal Processing Letters, IEEE*, 7 (7), 179 –181 38

POGGIO, G. F. AND POGGIO, T. (1984): The analysis of stereopsis. *Annual review of neuroscience*, 7, 379–412 95

PRATT, W. K. (1991): Digital Image Processing., 720 18, 101, 102

PRINCE, S. J. AND ROGERS, B. J. (1998): Sensitivity to disparity corrugations in peripheral vision. *Vison Research*, 38 (17), 2533–7, ISSN 0042–6989 95

PULLIAM, K. (1981): Spatial frequency analysis of three-dimensional vision. In *Proceedings of SPIE* Volume 303,, 71–77 20

PURVES, D., SHIMPI, A. AND LOTTO, B. R. (1999): An Empirical Explanation of the Cornsweet Effect. *Journal of Neuroscience*, 19 (19), 8542–8551 2

RAMACHANDRAN, V. S., RAO, V. M. AND VIDYASAGAR, T. R. (1974): Sharpness constancy during movement perception (Short note). *Perception*, 3 (1), 97–98 13

RAMANARAYANAN, G., FERWERDA, J., WALTER, B. AND BALA, K. (2007): Visual Equivalence: Towards a new standard for Image Fidelity. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 26 (3), 76 40

REINHARD, E., WARD, G., PATTANAIK, S., DEBEVEC, P., HEIDRICH, W. AND MYSZKOWSKI, K. (2010): High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting. 40

RICHARDS, W. (1971): Anomalous stereoscopic depth perception. *Journal of the Optical Society of America*, 61 (3), 410–14 80, 109

RITSCHEL, T., SMITH, K., IHRKE, M., GROSCH, T., MYSZKOWSKI, K. AND SEIDEL, H. (2008): 3D Unsharp Masking for Scene Coherent Enhancement. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 27 (3), 90:1–8 40

ROGERS, B. AND GRAHAM, M. (1983): Anisotropies in the perception of three-dimensional surfaces. *Science*, 221 (4618), 1409–11, ISSN 0036–8075 18

ROGERS, B. J. AND GRAHAM, M. E. (1979): Motion parallax as an independent cue for depth perception. *Perception*, 125 (8), 125–134 14

ROHALY, A. M. AND WILSON, H. R. (1999): The effects of contrast on perceived depth and depth discrimination. *Vison Research*, 39 (1), 9 – 18 23, 96

ROSS, J. (1974): Stereopsis by binocular delay. *Nature*, 248, 363–364 34, 130

SATO, M. (2004): A psychophysical study on the anisotropy and individual differences in human depth perception. *International Congress Series*, 1269, 97–100, Brain-Inspired IT I, ISSN 0531–5131 19, 95

SAWHNEY, H. S., GUO, Y., HANNA, K., KUMAR, R., ADKINS, S. AND ZHOU, S. (2001): Hybrid stereo camera: An IBR approach for synthesis of very high resolution stereoscopic image sequences. In *Proceedings of SIGGRAPH*, 451–460 35

SAXENA, A., CHUNG, S. H. AND NG, A. Y. (2005): Learning depth from single monocular images. In *In NIPS 18* 114

SAZZAD, Z., YAMANAKA, S., KAWAYOKEITA, Y. AND HORITA, Y. (2009): Stereoscopic image quality prediction. In *Quality of Multimedia Experience, Intl. Workshop on* IEEE, 180–185 42

SCHERZER, D., YANG, L., MATTAUSCH, O., NEHAB, D., SANDER, P. V., WIMMER, M. AND EISEMANN, E. (2011): A Survey on Temporal Coherence Methods in Real-Time Rendering. In *EUROGRAPHICS 2011 State of the Art Reports*, 101–126 33

SCHÜTZ, A. C., BRAUN, D. I., KERZEL, D. AND GEGENFURTNER, K. R. (2008): Improved visual sensitivity during smooth pursuit eye movements. *Nature Neuroscience*, 11 (10), 1211–1216 12

SEUNTIENS, P., MEESTERS, L. AND IJSSELSTEIJN, W. (2006): Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation. *ACM Transactions on Applied Perception*, 3, 95–109, ISSN 1544–3558 42

SHIBATA, T., KIM, J., HOFFMAN, D. AND BANKS, M. (2011): The zone of comfort: Predicting visual discomfort with stereo displays. *Jurnal of Vision* 11 (8) 24, 25, 41, 76

SIEGEL, M. AND NAGATA, S. (2000): Just enough reality: comfortable 3-D viewing via microstereopsis. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10 (3), 387–396 41, 116

SITTHI-AMORN, P., LAWRENCE, J., YANG, L., SANDER, P. V., NEHAB, D. AND XI, J. (2008): Automated reprojection-based pixel shader optimization. *ACM Transactions on Graphics* 27 (5), ISSN 0730–0301 33

SMYTHE, D. (1990): A two-pass mesh warping algorithm for object transformation and image interpolation. *Rapport technique* 1030 31

STELMACH, L. B., TAM, W. J. AND MEEGAN, D. V. (1999): Stereo image quality: Effects of spatio-temporal resolution. In MERRITT, J. O., BOLAS, M. T. AND FISHER, S. S., EDITORS: *Stereoscopic Displays and Virtual Reality Systems VI* Volume 3639,, 4–11 132

STICH, T., LINZ, C., WALLRAVEN, C., CUNNINGHAM, D. AND MAGNOR, M. (2011): Perception-motivated interpolation of image sequences. *ACM Transactions on Applied Perception*, 8 (2), 11:1–11:25 32, 34

SUNG, K., PEARCE, A. AND WANG, C. (2002): Spatial-Temporal Antialiasing. *IEEE Transactions on Visualization and Computer Graphics*, 8 (2), 144–153 54

TAKEUCHI, T. AND VALOIS, K. D. (2005): Sharpening image motion based on the spatio-temporal characteristics of human vision. In *Proceedings of SPIE*, 83–94 13

TAM, W. J. AND ZHANG, L. (2004): Nonuniform smoothing of depth maps before image-based rendering., 173–183 41

TAUBMAN, D. S. AND MARCELLIN, M. W. (2001): JPEG 2000: Image Compression Fundamentals, Standards and Practice., ISBN 079237519X 107

TAYLOR, M. AND CREELMAN, C. (1967): PEST: Efficient estimates on probability functions. *Journal of Acoustical Society of America*, 41, 782 80, 87

TEKALP, A. (1995): Digital Video Processing. 37, 38

TYLER, C. W. (1975): Spatial organization of binocular disparity sensitivity. *Vision Research*, 15 (5), 583 – 590 16, 76, 77, 78, 89, 91, 108

TYLER, C. (1973): Stereoscopic vision: cortical limitations and a disparity scaling effect. *Science*, 181 (4096), 276–278 16

VISBOX, INC.: Innovative Display and Interaction Technologies. `http://www.visbox.com` 7

WALTER, B., DRETTAKIS, G. AND PARKER, S. (1999): Interactive Rendering using Render Cache. In *Proceedings of EGSR*, 19–30 33

WANDELL, B. (1995): Foundations of Vision. 5, 6, 7, 10

WANG, Z., BOVIK, A. C., SHEIKH, H. R. AND SIMONCELLI, E. P. (2004): Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image processing*, 13 (4), 600–612 40, 42

WANGER, L., FERWERDA, J. AND GREENBERG, D. (1992): Perceiving spatial relationships in computer-generated images. *Computer Graphics and Applications, IEEE*, 12 (3), 44 –58, ISSN 0272–1716 75

WATSON, A. B. AND PELLI, D. G. (1983): QUEST: a Bayesian adaptive psychometric method. *Perception and Psychophysics*, 33 (2), 113–120 94

WATSON, A. (1987): The Cortex transform: rapid computation of simulated neural images. *Comp. Vision Graphics and Image Processing*, 39, 311–327 83

WESTERINK, J. AND TEUNISSEN, C. (1995): Perceived sharpness in complex moving images. *Displays*, 16 (2), 89–96 13

WEYRICH, T., DENG, J., BARNES, C., RUSINKIEWICZ, S. AND FINKELSTEIN, A. (2007): Digital Bas-relief from 3D Scenes. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 26 (3), 32, ISSN 07300301 41

WILSON, H. (1980): A transducer function for threshold and suprathreshold human vision. *Biological Cybernetics*, 38, 171–8 20, 40, 77, 83, 90

WINKLER, S. (2005): Digital video quality: vision models and metrics., ISBN 0470024046 84

WOLBERG, G. (1998): Image morphing: A survey. *The Visual Computer*, 14 (8), 360–372 31

WOODS, A. J. (1993): Image distortions in stereoscopic video systems. *Proceedings of SPIE*, 1915 (February 1993), 36–48 24

YANG, L., TSE, Y., SANDER, P., LAWRENCE, J., NEHAB, D., HOPPE, H. AND WILKINS, C. (2011): Image-based bidirectional scene reprojection. In *ACM Transactions on Graphics (TOG)* Volume 30, ACM, 150 59

YEH, Y.-Y. AND SILVERSTEIN, L. D. (1990): Limits of fusion and depth judgment in stereoscopic color displays. *Hum. Factors*, 32 (1), 45–60, ISSN 0018–7208 16

ZACH, C., POCK, T. AND BISCHOF, H. (2007): A Duality Based Approach for Realtime TV-L1 Optical Flow. In *DAGM-Symposium*, 214–223 58

ZAVAGNO, D. AND CAPUTO, G. (2001): The glare effect and the perception of luminosity. *Perception*, 30 (2), 209–222 2

ZHANG, G., HUA, W., QIN, X., WONG, T.-T. AND BAO, H. (2007): Stereoscopic Video Synthesis from a Monocular Video. *IEEE Transactions Visualization and Comput. Graph.* 13 (4), 686–696 35

ZWICKER, M., MATUSIK, W., DURAND, F., PFISTER, H. AND FORLINES, C. (2006): Antialiasing for Automultiscopic 3D Displays. In *Proceedings of EGSR*, 73–82 103, 104