–High-Fidelity Imaging–

The Computational Models of the Human Visual System in High Dynamic Range Video Compression, Visible Difference Prediction and Image Processing

Dissertation

zur Erlangung des Grades des
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
der Naturwissenschaftlich-Technischen Fakultäten
der Universität des Saarlandes

vorgelegt von

Rafał Mantiuk

Eingereicht am 10. Juli 2006 in Saarbrücken

Datum des Kolloqiums: 14.12.2006

Betreuender Hochschullehrer – Supervisor:
Dr.-Ing. habil. Karol Myszkowski, MPI für Informatik, Saarbrüken, Germany

Gutachter – Reviewers:
Dr.-Ing. habil. Karol Myszkowski, MPI für Informatik, Saarbrüken, Germany
Prof. Dr. Hans-Peter Seidel, MPI für Informatik, Saarbrüken, Germany
Prof. Dr. Sumanta N. Pattanaik, University of Central Florida, USA

Dekan – Dean:
Prof. Dr. Thorsten Herfet, Universität des Saarlandes, Saarbrüken, Germany

Abstract

As new displays and cameras offer enhanced color capabilities, there is a need to extend the precision of digital content. High Dynamic Range (HDR) imaging encodes images and video with higher than normal bit-depth precision, enabling representation of the complete color gamut and the full visible range of luminance.

This thesis addresses three problems of HDR imaging: the measurement of visible distortions in HDR images, lossy compression for HDR video, and artifact-free image processing. To measure distortions in HDR images, we develop a visual difference predictor for HDR images that is based on a computational model of the human visual system. To address the problem of HDR image encoding and compression, we derive a perceptually motivated color space for HDR pixels that can efficiently encode all perceivable colors and distinguishable shades of brightness. We use the derived color space to extend the MPEG-4 video compression standard for encoding HDR movie sequences. We also propose a backward-compatible HDR MPEG compression algorithm that encodes both a low-dynamic range and an HDR video sequence into a single MPEG stream. Finally, we propose a framework for image processing in the contrast domain. The framework transforms an image into multi-resolution physical contrast images (maps), which are then rescaled in just-noticeable-difference (JND) units. The application of the framework is demonstrated with a contrast-enhancing tone mapping and a color to gray conversion that preserves color saliency.

Kurzfassung

Aktuelle Innovationen in der Farbverarbeitung bei Bildschirmen und Kameras erzwingen eine Präzisionserweiterung bei digitalen Medien. High Dynamic Range (HDR) kodieren Bilder und Video mit einer grösseren Bittiefe pro Pixel, und ermöglichen damit die Darstellung des kompletten Farbraums und aller sichtbaren Helligkeitswerte.

Diese Arbeit konzentriert sich auf drei Probleme in der HDR-Verarbeitung: Messung von für den Menschen störenden Fehlern in HDR-Bildern, verlustbehaftete Kompression von HDR-Video, und visuell verlustfreie HDR-Bildverarbeitung. Die Messung von HDR-Bildfehlern geschieht mittels einer Vorhersage von sichtbaren Unterschieden zweier HDR-Bilder. Die Vorhersage basiert dabei auf einer Modellierung der menschlichen Sehens. Wir addressieren die Kompression und Kodierung von HDR-Bildern mit der Ableitung eines perzeptuellen Farbraums für HDR-Pixel, der alle wahrnehmbaren Farben und deren unterscheidbaren Helligkeitsnuancen effizient abbildet. Danach verwenden wir diesen Farbraum für die Erweiterung des MPEG-4 Videokompressionsstandards, welcher sich hinfort auch für die Kodierung von HDR-Videosequenzen eignet. Wir unterbreiten weiters eine rückwärts-kompatible MPEG-Kompression von HDR-Material, welche die übliche YUV-Bildsequenz zusammen mit dessen HDR-Version in einen gemeinsamen MPEG-Strom bettet. Abschliessend erklären wir unser Framework zur Bildverarbeitung in der Kontrastdomäne. Das Framework transformiert Bilder in mehrere physikalische Kontrastauflösungen, um sie danach in Einheiten von just-noticeable-difference (JND, noch erkennbarem Unterschied) zu reskalieren. Wir demonstrieren den Nutzen dieses Frameworks anhand von einem kontrastverstärkenden Tone Mapping-Verfahren und einer Graukonvertierung, die die ursprünglichen Farbkontraste bestmöglich beibehält.

Summary

As new displays and cameras offer enhanced color capabilities, there is a need to extend the precision of digital content, specifically images and video. High Dynamic Range Imaging (HDRI) encodes images and video with higher bit-depth precision, enabling representation of the complete color gamut and the full visible range of luminance, which makes this technology a successor to traditional 8-bit-per-color-channel imaging. However, to realize transition from the traditional to HDR imaging, it is necessay to develop imaging algorithms that work with the high-precision data. To make such algorithms effective and usable in practice, it is necessary to take advantage of the limitations of the human visual system by reducing the storage and processing precision so that it matches the performance of the human eye. Therefore, human visual perception is the key component in the solutions we present in this dissertation. We address three important problems in this dissertation: the measurement of visible distortions in HDR images, lossy compression for HDR video, and an HDR image processing framework, suitable for contrast compression.

To facilitate assessment of the visual quality of HDR content, we develop a visual difference predictor for HDR images. Given two images, the predictor can detect differences that would be noticeable to the human observer. The metric is based on a computational model of the human visual system, which we extend and adapt for HDR content. We included several aspects that are important in the perception of high contrast images, such as distortions of the eye's optics, photoreceptor response under a broad range of luminance adaptation conditions, and contrast sensitivity in the presence of the local adaptation. The metric is calibrated for natural images in a subjective experiment.

The key component of an imaging pipeline is standardized and effective image and video encoding. To address the problem of HDR image encoding and compression, we derive a color space for HDR pixels from perceptual measurements. The color space can efficiently encode all perceivable colors and distinguishable shades of brightness that are visible under all illumination conditions. The proposed color space, which requires only twelve bits to encode luminance and two eight-bit channels to encode chrominance, offers a straightforward extension of existing image and video compression standards.

We use the derived color space for HDR pixels to extend the MPEG-4 video compression standard for encoding HDR movie sequences. The extended encoder offers a special treatment of sharp contrast edges, which can have higher contrast than traditional video material. The proposed compression method proves to be an effective as well as novel extension to the existing MPEG standard (ISO/IEC 14496-2 and 14496-10).

To facilitate a smooth transition from traditional to HDR content, we propose a backward-compatible HDR MPEG compression algorithm. Within a single MPEG stream, the algorithm encodes two video sequences, one low-dynamic range (LDR – traditional video) and the other HDR, into a single MPEG stream. Naive applications recognize this stream as an ordinary MPEG video, however advanced software or hardware can decode HDR video. The algorithm requires only 8-bit software or hardware MPEG coders. The LDR and HDR video sequences are decorrelated to achieve the best compression performance. To further improve compression, invisible noise is removed from the HDR data stream using a multi-band perceptual filter. The filter estimates

visibility thresholds, taking into account luminance masking, the contrast sensitivity function, phase uncertainty and contrast masking.

The multi-resolution representations of images, such as wavelets, pyramids or band-pass channels, offer an attractive tool for image processing and editing. However, these representations often lead to unwanted artifacts and artificial looking resulting images, especially when each band or resolution is modified separately. To avoid such artifacts while benefiting from the advantages of the multi-resolution representation, we propose a contrast-domain image processing framework. The framework transforms an image into several resolutions of physical contrast. The contrast is then rescaled using a specially derived transducer function in perceptually plausible just-noticeable-difference (JND) units. The resulting image is constructed from the modified contrast by solving an optimization problem. All components of the framework are designed to work with high contrast HDR images. We demonstrate the application of the framework on a contrast-enhancing tone mapping and a color to gray conversion that preserves color saliency. The framework is especially effective for operations that heavily distort contrast, such as extreme sharpening of images.

The proposed solutions constitute the central part of the HDR pipeline. The predictor enables the evaluation of HDR image quality and thus was instrumental in developing a color space for HDR pixels that is free of contouring artifacts, as well as the compression algorithms. Lossy HDR video compression is indispensable for efficient storage and transmission of HDR content. Finally, the contrast-domain image processing framework enables rendering such content on existing low-dynamic range displays.

In summary, this dissertation contributes primarily to the fields of encoding and compression of HDR image and video, computational models of visual system for HDR images and multi-resolution image processing. The proposed solutions can help in standardizing color spaces and compression algorithms for HDR content. The visual difference metric contributes to a better understanding of the perception of high contrast images and is useful as a tool for validating imaging and computer graphics algorithms. The multi-resolution image processing framework facilitates image editing in a perceptually plausible contrast domain, which, unlike existing methods, does not lead to unwanted artifacts.

Zusammenfassung

Aktuelle Innovationen in der Farbverarbeitung bei Bildschirmen und Kameras erzwingen eine Präzisionserweiterung bei digitalen Medien, besonders bei Bild- und Videodaten. High Dynamic Range (HDR) kodiert Bilder und Video mit einer grösseren Bittiefe pro Pixel, und ermöglicht damit die Darstellung des kompletten Farbraums und aller sichtbaren Helligkeitswerte. Damit wird es den Nachfolger der traditionellen 8 bit-Verarbeitung in den Farbkanaelen stellen.

Für den reibungslosen Übergang von der traditionellen Bildverarbeitung zu HDR-Verfahren werden Bildverarbeitungsalgorithmen benötigt, die mit hoch auflösenden Daten umgehen können. Diese Algorithmen sind in der Praxis nur dann effizient und anwendbar, wenn sie sich der Beschränkungen des menschlichen Sehens bedienen und die Datenrepräsentation in ähnlichen Zügen führen, um den Speicherbedarf und die Verarbeitungsgenauigkeit klein zu halten. Deswegen ist das menschliche Sehen einer der Schlüsselpunkte für die Problemlösungsansätze in dieser Dissertation. Diese Arbeit konzentriert sich auf drei Probleme in der HDR-Verarbeitung: Messung von für den Menschen störenden Fehlern in HDR-Bildern, verlustbehaftete Kompression von HDR-Video, und visuell verlustfreie HDR-Bildverarbeitung.

Die Messung von HDR-Bildfehlern geschieht mittels einer Vorhersage von sichtbaren Unterschieden zweier HDR-Bilder. Der Vorhersage-Operator kann dabei mit Hilfe zweier Bilder die Unterschiede erkennnen, die auch einem menschlichen Beobachter auffallen würden. Diese Metrik basiert auf einem rechnerischen Modell des menschlichen Sehens, das wir für HDR-Medien angepasst und erweitert haben. Wir inkludieren mehrere Aspekte, die beim visuellen Erfassen von Hochkontrast-Aufnahmen eine Rolle spielen, darunter optische Verzerrungen im menschlichen Auge, Sehzellenverhalten in stark verschiedenen Zuständen der Helligkeitsanpassung, und Kontrastempfindlichkeit unter Rücksichtnahme auf lokale Anpassung. Die Metrik wird in einem subjektiven Experiment auf natürliche Bilder kalibriert.

Der wichtigste Baustein einer Bildverarbeitungspipeline ist die standardisierte und effiziente Bild- und Videokodierung. Wir addressieren die Kompression und Kodierung von HDR-Bildern mit der Ableitung eines perzeptuellen Farbraums für HDR-Pixel. Dieser Farbraum kann alle wahrnehmbaren Farben und deren unterscheidbaren Helligkeitsnuancen effizient für alle möglichen Lichtverhältnisse abbilden. Der vorgeschlagene Farbraum benötigt weiter nur zwölf Bit zur Abbildung von Helligkeit, und zwei Achtbit-Kanäle zur Abbildung der Chrominanz, und bietet damit eine logische Erweiterung von existierenden Bild- und Videokodierungsverfahren.

Danach verwenden wir diesen Farbraum für die Erweiterung des MPEG-4 Videokompressionsstandards, welcher sich hinfort auch für die Kodierung von HDR-Videosequenzen eignet. Der neue Kodierer bietet dafür eine Spezialbehandlung von kontrastreichen Bilddetails, die in normalem Videomaterial so nicht auftreten würden. Diese Kodierungsmethode hat sich als effiziente und geradlinige Erweiterung des existierenden MPEG-Standards erwiesen (ISO/IEC 14496-2 und 14496-10).

Um den Übergang von traditionellem zu HDR-Material zu erleichtern, bieten wir eine rückwärts-kompatible MPEG-Kompression von HDR-Material. Der Algorithmus kodiert dabei zwei Videosequenzen in einen gemeinsamen MPEG-Strom, eine traditionelle / LDR Sequenz, und eine HDR-Sequenz. Software oder Hardware neueren Schlages können damit HDR-Video dekodieren, während alte oder einfache Deco-

der den MPEG-Strom weiterhin als traditionelles MPEG-Video betrachten. Der Algorithmus benötigt dabei weiterhin nur 8-bit-fähige MPEG-Encoder (egal ob Software oder Hardware). Die LDR und HDR-Videosequenzen werden datenmässig dekorreliert, um die bestmögliche Kompression zu erreichen. Weitere Kompressionseffizienz wird mit Hilfe eines perzeptuellen Multiband-Filters erreicht, welches nicht unsichtbares Bildrauschen aus dem HDR-Datenstrom entfernt. Der Filter schätzt Sichtbarkeitsschwellen, indem er Helligkeitsmaskierung, Kontrastempfindlichkeit, Phasenungenauigkeit und Kontrastmaskierung einrechnet.

Bildrepräsentationen in multiplen Auflösungen, z.B. Wavelets, Pyramids oder Bandpasskanal-Repräsentationen, bieten ein nützliches Werkzeug für Bildverarbeitung und Bildbearbeitung. Leider führen diese Repräsentationen oft zu ungewollten Artefakten und Bildern mit künstlichem Aussehen, besonders wenn Bänder oder Auflösungsstufen einzeln modifiziert werden. Unsere Bildverarbeitungs-Framework in der Kontrast-Domäne ermöglicht es, solche Artefakte zu vermeiden. Das Framework transformiert zuerst Bilder in mehrere physikalische Kontrastauflösungen. Danach reskaliert es den Bildkontrast mit Hilfe einer speziellen Übertragungsfunktion in Einheiten von just-noticeable-difference (JND, noch erkennbarem Unterschied). Das Ausgabebild entsteht am Ende aus dem modifizierten Kontrast durch die Lösung eines Optimierungsproblems. Alle Komponenten des Frameworks können mit Hochkontrast-HDR-Bildern arbeiten. Wir demonstrieren den Nutzen dieses Frameworks anhand von einem kontrastverstärkenden Tone Mapping-Verfahren und einer Graukonvertierung, die die ursprünglichen Farbkontraste bestmöglich beibehält. Das Framework zeigt seine besonderen Stärken bei Operationen mit starken Kontrastveränderungen, wie dem extremen Schärfen von Bilddetails.

Die genannten Lösungsansätze bilden den Kern der HDR-Pipeline. Der Vorhersage-Operator ermöglicht die Auswertung der HDR-Bildqualität, und spielte eine wichtige Rolle bei der Suche nach einem HDR-Farbraum ohne Kontur-Artefakte, und bei der Entwicklung des Videokompressionsverfahrens. Verlustbehaftete HDR-Videokompression ist für die effiziente Lagerung und Übertragung von HDR-Material unabdingbar. Danach können mit Hilfe der Bildverarbeitung in der Kontrastdomäne auch traditionelle LDR-Displays (Low Dynamic Range) für die Anzeige von HDR-Inhalten verwendet werden.

Diese Doktorarbeit trägt also vorrangig zu folgenden Bereichen bei: Repräsentation und Kompression von HDR-Video und HDR-Bildmaterial, Berechnungsmodelle des menschlichen Sehens für HDR-Bilder und Bildverarbeitung in multiplen Auflösungen. Die vorgeschlagenen Lösungen können bei der Standardisierung von Farbräumen und Kompressionsverfahren von HDR-Material behilflich sein. Die Metrik für noch erkennbare Bildunterschiede (JND) erweitert das Verständnis des Sehvorganges für HDR-Bildmaterial mit hohem Kontrast, und eignet sich zur Validierung von verwandten Bildverarbeitungs- und Computergraphikalgorithmen. Das Bildverarbeitungs-Framework in multiplen Auflösungen erleichtert die Bildbearbeitung in einer perzeptuell plausiblen Kontrastdomäne, die, ungleich existierenden Methoden, nicht zu ungewollten Artefakten führt.

Acknowledgements

First of all, I would like to thank my supervisor Dr.-Ing. habil. Karol Myszkowski for his interest in this work, his valuable comments, his continuous support, and giving me freedom to pursue my own ideas. Dr. Myszkowski is responsible for making me interested in computer graphics and especially high dynamic range imaging and human visual perception.

I would like to thank Prof. Dr. Hans-Peter Seidel for creating an excellent work environment at the Max-Plank Institute, and his great support for our projects in the novel field of high dynamic range imaging.

I would also like to thank the external reviewer Prof. Dr. Sumanta Pattanaik who agreed to reviews this thesis. I had the pleasure of spending a semester working with Prof. Pattanaik at the University of Central Florida, during which I decided to further my studies in the area of computer graphics.

I would also like to thank Prof. Dr. Wolfgang Heindrich for hosting me at his group in Vancouver and allowing me to work on a prototype of the HDR display. I would especially like to thank Scott Daly for many insightful discussions, valuable comments and recently inviting me for an internship with his group at Sharp Laboratories in America. I am very grateful to Helge Seetzen for fruitful collaboration in several HDR projects and his support for the work on the backward-compatible HDR MPEG compression. Special thanks to Greg Ward for many comments on our work.

I would especially like to thank Grzegorz Krawczyk, Akiko Yoshida and Alexander Efremov, who co-authored many of my previous publications. Many projects described in this dissertation would not have been possible without their help and contributions.

Finally, I would like to thank all my present and former colleagues at the Computer Graphics Group at the MPI, who make it such a great place. Special thanks to Kaleigh Smith, Gernot Ziegler, Christina Scherbaum and Michael Neff for their help and comments on some of the publications, and to Martin Fuchs and Carsten Stoll for technical support, particularly on the days before deadlines.

Contents

Chapter 1

Introduction

The majority of existing digital imagery and video material capture only a fraction of the visual information that is visible to the human eye and are not of sufficient quality for reproduction by the future generation of display devices. The limiting factor is not the resolution, since most consumer level digital cameras can take images of higher number of pixels than most of displays can offer. The problem is the limited color gamut and even more limited dynamic range (contrast) captured by cameras and stored by the majority of image and video formats.

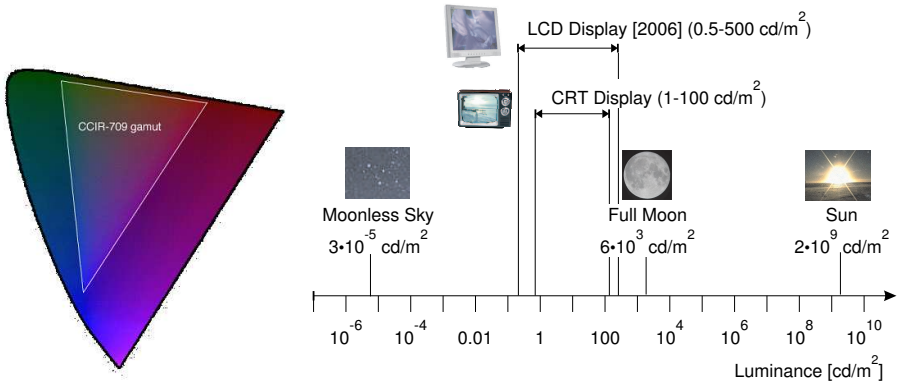


Figure 1.1: Left: the standard color gamut frequently used in traditional imaging (CCIR-705), compared to the full visible color gamut. Right: real-world luminance values compared with the range of luminance that can be displayed on CRT and LDR monitors. Most digital content is stored in a format that at most preserves the dynamic range of typical displays.

For instance, each pixel value in the JPEG image encoding is represented using three 8-bit integer numbers (0-255) using the YC_rC_b color space. This color space is able to store only a small part of visible color gamut (although containing the colors most often encountered in the real world), as illustrated in Figure 1.1-left, and an even smaller part of the luminance range that can be perceived by our eyes, as illustrated in Figure 1.1-right. The reason for this is that the JPEG format was designed to store as much information as can be displayed on the majority of displays, which were at that

time Cathode Ray Tube (CRT) monitors or TV sets. This assumption is no longer valid, as the new generations of LCD and Plasma displays can depict a much broader color gamut and dynamic range than their CRT ancestors. Every new generation of displays offers better color reproduction and requires higher precision of image and video content. The traditional low-dynamic range and limited color gamut imaging, which is confined to three 8-bit integer color channels, cannot offer the precision that is needed for the upcomming developments in image capture, processing, storage and display technologies.

High Dynamic Range Imaging (HDRI) overcomes the limitation of traditional imaging by performing operations on color data with much higher precision. Pixel colors are specified in HDR images as a triple of floating point values (usually 32-bit per color channel), providing accuracy that exceeds the capabilities of the human visual system [Reinhard et al. 2005]. Moreover, while traditional imaging assumes that content is already profiled for a particular display medium (paper, LDR/CRT display), HDRI operates on colors of original scenes. By its inherent colorimetric precision, HDRI can represent all colors found in real world that can be perceived by the human eye.

HDRI has recently gained momentum and is revolutionizing almost all fields of digital imaging. One of the breakthroughs of the HDR revolution was the development of an HDR display, which proved that the visualization of color and the luminance range close to real-world scenes is possible [Seetzen et al. 2004]. One of the first to adopt HDRI were video game developers together with graphics card vendors. Today most of the state-of-the art video game engines perform rendering using HDR precision to deliver more believable and appealing virtual reality imagery. Computer generated imagery used in special effect production uses HDR techniques to achieve the best match between synthetic and realistic objects. High-end cinematographic cameras, both analog and digital, already provide significantly higher dynamic range than most of the displays today. This dynamic range can be retained after digitalization only if a form of HDR representation is used. HDRI is also a strong trend in digital photography, mostly due to the multi-exposure techniques that allow an HDR image to be made using a consumer level digital camera. HDR cameras that can directly capture higher dynamic range are available, for example *SheroCamHDR* from *SheronVR*, *Origin*®from *Dalsa* or *Viper FilmStream*™. To catch up with the HDR trend, many software vendors announce their support of the HDR image formats, taking Adobe® Photoshop® CS2 as an example. In general, the products start to appear at both ends of the imaging pipeline: HDR cameras on the acquisition side, and commercial tone-mapping and rendering algorithms on the display side. However, the storage and transmission stage lacks any well defined standards and no products are available. There are almost no solutions for lossy, and thus efficient, HDR image and video compression. The lack of standards can result in a multitude of incompatible image and video formats. This situation is already happening in the case of cameras' RAW formats, which are different from vendor to vendor. Moreover, HDR is likely to be misinterpreted by the industry, which can develop and standardize another device dependent format, which offers nothing more than slightly extended color gamut and dynamic range, but is still insufficient to cover the entire range of HDR applications. This way, the huge advantage of HDR, which is device independence, would be lost.

HDRI does not only provide higher precision, but also enables the synthesis, storage and visualization of a range of perceptual cues that are not achievable with traditional imaging. Most of the imaging standards and color spaces have been developed to match

the needs of office or display illumination conditions. When viewing such scenes or images in such conditions, our visual system operates in a mixture of day-light and dim-light vision state, so called the mesopic vision. When viewing out-door scenes, we use day-light perception of colors, so called the photopic vision. This distinction is important for digital imaging as both types of vision shows different performance and result in different perception of colors. HDRI can represent images of luminance range fully covering both the photopic and the mesopic vision, thus making distinction between them possible. One of the differences between mesopic and photopic vision is the impression of colorfulness. We tend to regard objects more colorful when they are brightly illuminated, which is the phenomenon that is called Hunt's effect. To render enhanced colorfulness properly, digital images must preserve information about the actual level of luminance of the original scene, which is not possible in the case of traditional imaging. Real-world scenes are not only brighter and more colorful than their digital reproductions, but also contain much higher contrast, both local between neighboring objects, and global between distant objects. The eye has evolved to cope with such high contrast and its presence in a scene evokes important perceptual cues. Traditional imaging, unlike HDRI, is not able to represent such high-contrast scenes. Similarly, traditional images can hardly represent common visual phenomena, such as self-luminous surfaces (sun, shining lamps) and bright specular highlights. They also do not contain enough information to reproduce visual glare (brightening of the areas surrounding shining objects) and a short-time dazzle due to sudden increase of the brightness of a scene (e.g. when exposed to the sunlight after staying indoors). To faithfully represent, store and then reproduce all these effects, the original scene must be stored and treated using high fidelity HDR techniques.

Besides its significant impact on existing imaging technologies that we can observe today, HDRI has the potential to radically change the methods by which imaging data is processed, displayed and stored in several fields of science. Computer vision algorithms can greatly benefit from the increased precision of HDR images, which lack over- or under-exposed regions, which are often the cause of the algorithms failure. Medical imaging has already developed image formats (e.g. the DICOM format) that partly cope with the shortcomings of traditional images, however they are supported only by specialized hardware and software. HDRI gives the sufficient precision for medical imaging and therefore its capture, processing and rendering techniques can be used also in this field. For instance, HDR displays can show even better contrast than high-end medical displays and therefore facilitate diagnosis based on CT scans. HDR techniques can also find applications in astronomical imaging, remote sensing, industrial design and scientific visualization.

1.1 Problem Statement

In our work we strive to realize the concept of an imaging pipeline that would not be restricted by any particular imaging technology and, if efficiency of storing data is required, is limited only by the capabilities of the human visual system.

The concept of an imaging pipeline is illustrated in Figure 1.2. At the first stage digital images are acquired, either with cameras or computer rendering methods. At the second stage, digital content is efficiently compressed and encoded either for storage or transmission purposes. Finally, digital video or images are displayed on display de-

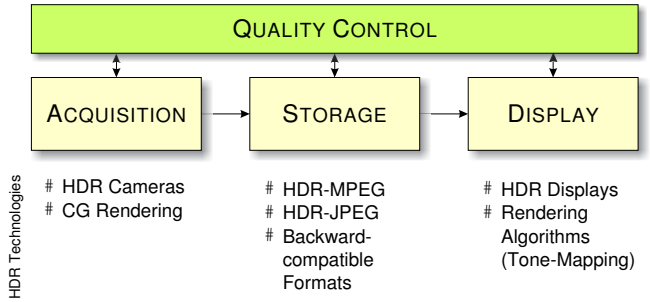


Figure 1.2: Imaging pipeline and available HDR technologies.

vices. Additionally, to verify algorithms at all stages of the pipeline, quality metrics are employed. The difference between HDRI and traditional imaging is that HDRI operates on device-independent and high-precision data throughout all the stages of the pipeline, so that the quality of the content is reduced only at the display stage, and only if a device cannot faithfully reproduce the content. This is contrary to traditional imaging, where the content is usually profiled for particular device and thus stripped from useful information as early as at the acquisition stage or latest at the storage stage. For example, most consumer level digital cameras store images in the JPEG format, which offers sufficient quality for print, but not sufficient quality for wide-gamut and high-dynamic range displays. Another example is color spaces used in traditional imaging that are often based on the spectral response of the red, green and blue phosphors in CRT displays. Since CRT technology is being replaced by LCD and plasma technologies, the use of CRT primaries can be questioned. HDRI, on the other hand, can offer an image-independent representation of images and video, so that the content can be rendered on any display device. The proper rendering of the content is the responsibility of a device, since only the device has all the information related to its limitation and sometimes also viewing conditions (e.g. ambient illumination), which is necessary to render the content properly.

The major focus of this dissertation is the encoding and compression of HDR content. In order to make HDR compression efficient, we devote much effort to better understand the human visual perception, especially in the context of high contrast images, where local adaptation and dark-to-daylight vision plays an important role. One of the outcomes of such perceptual considerations is a visual difference metric that can be applied to real-world scenes. Besides image and video formats, the dynamic range reduction, necessary to display HDR content on LDR displays, is another and still not fully solved problem. We address this problem by proposing a contrast processing framework, which is a robust tool for producing believable renderings of HDR scenes on LDR displays.

1.2 Main Contributions

Parts of this dissertation have already been published at several conferences and in various journals [Mantiuk et al. 2004a, Mantiuk et al. 2004b, Mantiuk et al. 2005a, Mantiuk et al. 2005b, Mantiuk et al. 2006c, Mantiuk et al. 2006a, Mantiuk et al. 2006d]. These

publications are the foundation of this thesis, which unites them under the concept of the HDR imaging and presents improvements and updated results.

The main contributions of this dissertation can be summarized as follows:

- A method for perceptual linearization of luminance values. The method can be used for a range of applications, such as prediction of photoreceptor response in models of the human visual system (Section 4.2.2), image and video compression (Section 5.3.1) and prediction of perceived brightness.
- Two algorithms for encoding HDR video content. The first method is an extension of the MPEG-4 standard (ISO/IEC 14496-2) and the second offers backward compatibility with any MPEG compression. Both algorithms are viable solutions for future generation wide color gamut and high dynamic range video encoding.
- An extension of the visual difference metric capable of handling real-world viewing conditions. The metric is based on the model of human visual system and can predict visible differences between a pair of images for the full range of colors and luminance values visible to the human eye.
- A computational framework for the processing of images in perceptualy plausible visual contrast space. The framework offers an image representation, that, unlike the wavelet or the Fourier domains, does not lead to contrast reversal artifacts when spatial bands are modified separately. The framework is demonstrated to be effective in the tasks of tone mapping and color salience preserving color-to-gray conversion.

1.3 Chapter Overview

This dissertation is organized as follows: Chapter 2 gives background information on the digital representation of images and the photometric and colorimetric description of light and color. Chapter 3 summarizes the components of the computational models of the visual system and their applications. In Chapter 4 we describe our extension to the visual difference predictor that enables the prediction of differences in HDR images. The most extensive chapter, Chapter 5, introduces the concepts of HDR image and video compression, starting with a summary of existing solutions (Section 5.2), followed by the derivation of the novel color space for HDR pixels (Section 5.3), the HDR extension to MPEG-4 video compression (Section 5.4) and finally the backward-compatible HDR MPEG video compression (Section 5.6). The framework for image processing in the contrast domain is described in Chapter 6. We conclude this dissertation and give an outlook for future work in Chapter 7. In Appendix A we describe software packages we developed for processing of HDR images and video that have been made available as an open source project.

Chapter 2

Physical, Photometric and Colorimetric Image Representation

This chapter explains several physical and perceptual quantities important for digital imaging, such as radiance, luminance, luminance factor, luma, and color. It does not give a complete or exhaustive introduction to radiometry, photometry or colorimetry, since these are described in full extent elsewhere [Hunt 1995, Wyszecki and Stiles 2000, Reinhard et al. 2005]. The focus of this chapter is on the concepts that are confusing or vary in terminology between disciplines, and also those that are used in the following chapters.

2.1 Light

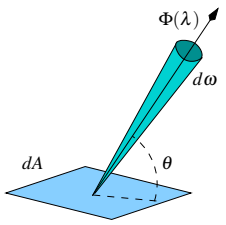


Figure 2.1: Spectral radiance. Spectral radiance is a differential measure, defined for infinitely small area dA, infinitely small solid angle $d\omega$, radiant flux Φ and an angle between the rays and the surface θ.

The physical measure of light that is the most appropriate for imaging systems is either *luminance* (used in photometry) or *spectral radiance* (used in radiometry). This is because both measures stay constant regardless of the distance from a light source to a sensor (assuming no influence of the medium in which the light travels). The sensor can

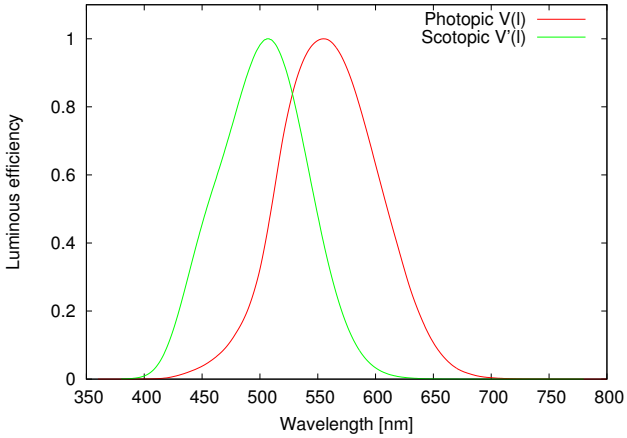


Figure 2.2: CIE spectral luminous efficiency curve for photopic (day light) and scotopic (night) vision. Data downloaded from http://www.cvrl.org/.

be either camera's CCD chip or a photoreceptor in the eye. The quantities measured by photoreceptors or digital sensors are related to either of these measures.

Spectral radiance is a radiometric measure, defined by:

$$L(\lambda) = \frac{d^2\Phi(\lambda)}{d\omega \cdot dA \cdot \cos\theta} \tag{2.1}$$

where $L(\lambda)$ is spectral radiance for the wavelength λ, Φ is radiant flux flowing through a surface per unit time, ω is a solid angle, θ is an angle between the rays and the surface, and A is the area of the surface, as illustrated in Figure 2.1. Although spectral radiance is commonly used in computer graphics, images are better defined with photometric units of *luminance*. *Luminance* is spectral radiance integrated over the range of visible wavelengths with the weighting function $V(\lambda)$:

$$Y = \int_{380nm}^{770nm} L(\lambda)V(\lambda)d\lambda \tag{2.2}$$

The function $V(\lambda)$, which is called the *spectral luminous efficiency curve* [CIE 1986], gives more weight to the wavelengths, to which the human visual system (HVS) is more sensitive. This way luminance is related (though non-nonlinearly) to our perception of brightness. The function V for the daylight vision (photopic) and night vision (scotopic) is plotted in Figure 2.2. Terms scotopic and photopic will be discussed in more detail in Section 3.2. Luminance, Y, is usually given in cd/m^2 or equivalent *nit* units.

Since the most common multi-exposure technique for acquiring HDR images [Reinhard et al. 2005, Chapter 4] can not assess the absolute luminance level but only a relative luminance values, most HDR images do not contain luminance values but rather the values of *luminance factor*. Such luminance factor must be multiplied by a constant number, which depends on a camera and lens, to get actual luminance. Such constant number can be easily found if we can measure the luminance of a photographed surface [Krawczyk et al. 2005a].

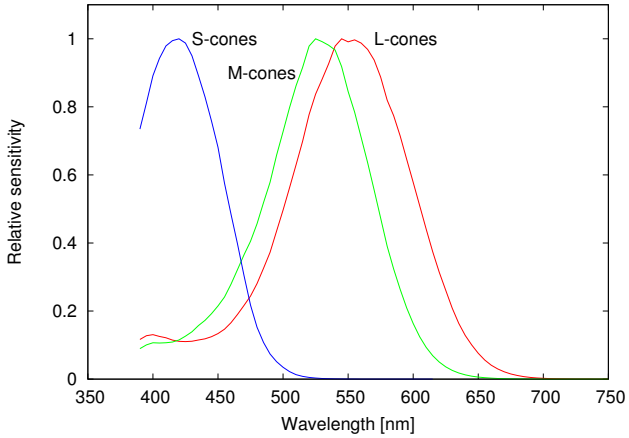


Figure 2.3: Cone photocurrent spectral responsivities. After [Stockman and Sharpe 2000].

2.2 Color

Colors are perceptual rather than physical phonomena. Although we can precisely describe colors using physical units of spectral radiance, such description does not give immediate answer whether the described color is green or red. *Colorimetry* is the field that numerically characterizes colors and provides a link between the human color perception and the physical description of the light. This section introduces the most fundamental aspects of colorimetry and introduces color spaces, which will be used in later chapters. More detailed introduction to colorimetry can be found in [Fairchild 1997] and [Reinhard et al. 2005], while two handbooks, [Wyszecki and Stiles 2000] and [Hunt 1995], are more exhaustive source of information.

The human color perception is determined by three types of cones: L, M and S, and their sensitivity to wavelengths. We will come back to the function of the photoreceptors in Section 3.2. The light in the visible spectrum is in fact multi-dimensional variable, where each dimension is associated with particular wavelength. However, the visible color is a projection of this multi-dimensional variable to three primaries, corresponding to three types of cones. Such projection is mathematically described as a product of the spectral power distribution, $\phi(\lambda)$, and the spectral response of the type of cones, $C_L(\lambda)$, $C_M(\lambda)$ and $C_S(\lambda)$:

$$R = \int_\lambda \phi(\lambda) C_L(\lambda) d\lambda \tag{2.3}$$

$$G = \int_\lambda \phi(\lambda) C_M(\lambda) d\lambda \tag{2.4}$$

$$B = \int_\lambda \phi(\lambda) C_S(\lambda) d\lambda \tag{2.5}$$

The spectral responsivities of cones are shown in Figure 2.3.

As the result of three-dimensional encoding of color in the HVS, the number of distinguishable colors is limited. Also, two stimuli of different spectral power distributions

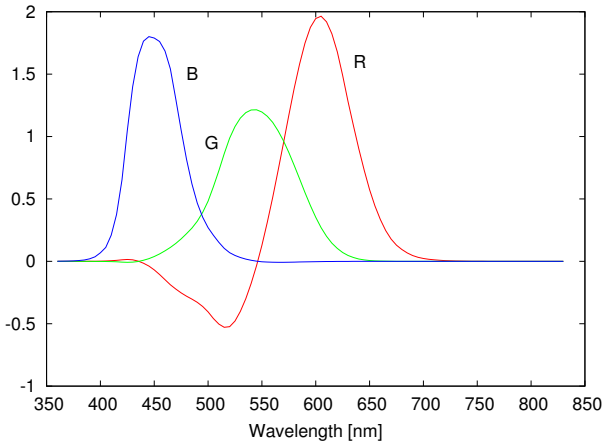


Figure 2.4: Color matching functions for the CIE matching stimuli R, G and B and 2° standard observer. Data downloaded from http://www.cvrl.org/.

can be seen as having the same color if only their R, G, and B projections match. The latter property of the HVS is called *metamerism*.

To uniquely describe visible color gamut, CIE standardized in 1931 a set of primaries for the standard colorimetric observer. Since the cone spectral responsivities were not known at that time, the primaries were based on color matching experiment, in which monochromatic stimuli of particular wavelength was matched with a mixture of the three monochromatic primaries (435.6 nm, 546.1 nm, and 700 nm). The values of color-maching mixture of primaries for each wavelength gave the R, G and B primaries shown in Figure 2.4. The drawback of this procedure was that it resulted in negative value of R primary. The negative part represents out of gamut colors, which are too saturated to be within visible or physically feasible range. To bring those colors into the valid gamut, the colors must be desaturated by adding monochromatic light. Since adding monochromatic light results in increasing the values of all R, G and B components, there is a certain amount of the added light that would make all components positive.

To avoid negative primaries and to connect colorimetric description of the light with photometric measure of luminance (see previous section), CIE introduced XYZ primaries in 1931. The primaries, shown in Figure 2.5, were designed so that primary Y represents luminance and its spectral tristimulus values are equal the luminous efficiency function (see Figure 2.2). Although the standard has been established over 70 years ago, it is still commonly used today, especially as a reference in color conversion formulas.

For a convenient two-dimensional representation of the color, chromacity coordinates are often used:

$$x = \frac{X}{X+Y+Z} \tag{2.6}$$

$$y = \frac{Y}{X+Y+Z} \tag{2.7}$$

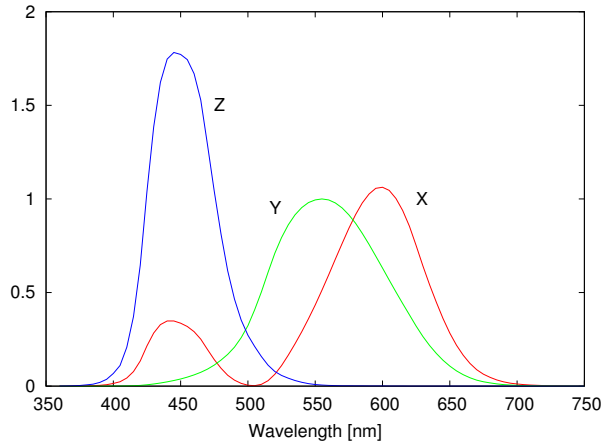


Figure 2.5: Color matching functions for the CIE matching stimuli X, Y and Z and 2° standard observer. Data downloaded from http://www.cvrl.org/.

Such coordinates must be accompanied by the corresponding luminance value, Y, to fully describe the color.

The visible differences between colors are not well described by chromacity coordinates x and y. For better representation of perceptual color differences, CIE defined uniform chromatcity scales (UCS) in 1976, which are known as CIE 1976 Uniform Chromacity Scales:

$$u' = \frac{4X}{X + 15Y + 3Z} \tag{2.8}$$

$$v' = \frac{9Y}{X + 15Y + 3Z} \tag{2.9}$$

Note that u', v' chromacity space only approximates perceptual uniformity and a unit Cartesian distance can denote from 1 JND[1] to 4 JND units.

The Uniform Chromacity Scales do not incorporate luminance level in their description of color. This is a significant limitation, as color difference can strongly depend on actual luminance level. Uniform color spaces have been introduced to address this problem. The first color space, CIE 1976 $L^*a^*b^*$, is defined by:

$$L^* = 116(Y/Y_n)^{1/3} - 16 \tag{2.10}$$

$$a^* = 500\left[(X/X_n)^{1/3} - (Y/Y_n)^{1/3}\right] \tag{2.11}$$

$$b^* = 200\left[(Y/Y_n)^{1/3} - (Z/Z_n)^{1/3}\right] \tag{2.12}$$

and the second color space, CIE 1976 $L^*u^*v^*$, by:

$$L^* = 116(Y/Y_n)^{1/3} - 16 \tag{2.13}$$

[1] JND – Just Noticeable Difference is usually defined as a measure of contrast at which a subject has 75% chance of correctly detecting visual difference in a stimulus.

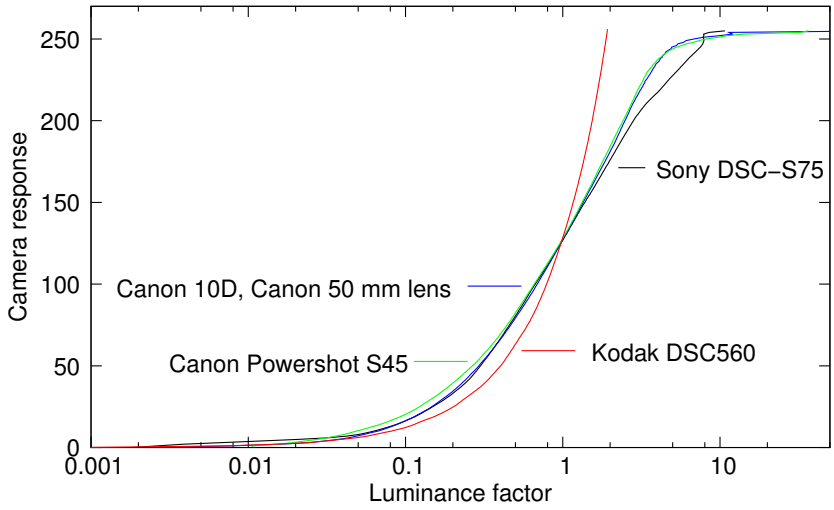


Figure 2.6: Response functions of several digital cameras. The responses of red, green and blue color components are averaged and represented as a single line. The curves were measured using *pfscalibration*[3] software.

$$u^* = 13L^*(u' - u'_n) \tag{2.14}$$

$$v^* = 13L^*(v' - v'_n) \tag{2.15}$$

The coordinates with the n subscript denote the color of the *reference white*, which is the color that appears white in the scene. For color print this is usually the color of a white paper under given illumination. Both color spaces have been standardized as the studies did not show that the one is definitely better over another and each one has its advantages.

Both CIE 1976 $L^*a^*b^*$ and CIE 1976 $L^*u^*v^*$ color spaces have been designed for low dynamic range color range, available on print or typical CRT displays and cannot be used for HDR images. In Section 5.3 we attempt to address this problem by deriving an (approximately) perceptually uniform color space for HDR pixel values.

The uniform color spaces are the simplest incarnations of color appearance models. Color appearance models try to predict not only the colorimetric properties of the light, but also its appearance under given viewing conditions (background color, surround ambient light, color adaptation, etc.). CIECAM02 [CIE 2002] is an example of such a model that has been standardized by CIE. The discussion of color appearance models would go beyond scope of this thesis, therefore reader should refer to [Hunt 1995] and [Fairchild 1997] for more information.

2.3 Sensor Response

Although radiometric of photometric units give probably the most accurate description of light, the output of most imaging systems, including displays, cameras and also

Figure 2.7: A response curve for a typical negative film shows higher dynamic range (up to 12 f-stops) than can be achieved by most film cameras.

photoreceptors, is neither luminance nor spectral radiance. Figure 2.6 illustrates the response function of several digital cameras and Figure 2.7 the response function of an analog film. Such response functions describe the relation between input luminance and output values for several sensors. The response of most imaging systems usually follows an S-shaped curve, which tends to saturate both the highest and the lowest luminance values. Since the middle segment of those curves has either logarithmic or power function characteristic, this non-linear compression is sometimes confusingly called "gamma correction". The gamma correction is in fact a compression of luminance applied to account for non-linear characteristic of CRT displays. This characteristic happens to be a close match to the non-linear characteristic of the eye for a range of luminance that is achievable on the CRT displays (from 1 to about 100 cd/m^2). However, the sensitivity of the eye is in fact quite different from the power function for luminance levels above 1000 cd/m^2 and below 1 cd/m^2. Therefore a gamma correction should never be used for HDR images.

It is often unclear how sensor's output values should be called. The usual term used for digital video is *luma*, which is a word coined by the NTSC to prevent confusion between the video signal and the traditional meaning of luminance. Since each sensor has its own response characteristic, it is impossible to define a single formula for luma. The relations between luminance and luma used in LDR video compression, which are sometimes called transfer functions, usually involve a power function similar to the gamma correction. However, since the gamma correction is a poor match to the characteristic of the eye for the full range of luminance, those formulas are not applicable to HDR data. To address this problem, in Section 5.3.1 we propose luma encoding of luminance suitable for the full range of visible luminance, which is based on the characteristic of the HVS. Such encoding defines luma in terms of sensitivity to light, in a similar way as luminance is defined in terms of spectral radiance. One advantage of such perceptual representation of luminance is that such a measure of light is perceptually linearized, which means that luma values correlate well with our perception of brightness.

The mistake made by many researchers entering the field of HDR imaging is that they evaluate distortions in HDR images using the root mean square (RMS) metric on the values of luminance. But, since luminance badly corresponds to our perception of

1.	Contrast ratio	$1:R=1:\frac{Y_{peak}}{Y_{noise}}$	general, display specifications
2.	log-10 units (orders of magnitude)	$M=\log_{10}\frac{Y_{peak}}{Y_{noise}}$	general
3.	Exposure latitude (f-stops)	$EL=\log_2\frac{Y_{peak}}{Y_{noise}}$	photography
4.	Signal to noise ratio	$DR_{PSNR}=20\cdot\log_{10}\frac{N_{max}}{RMS_{noise}}[db]$	camera specifications
5.	Density range	$DR=D_{max}-D_{min}\approx M$	photography, film scanners

Table 2.1: Measures of dynamic range used in different disciplines. Y_{peak} is the representative peak (maximum) luminance value and Y_{noise} is the level of noise.

brightness, the result of such metric will not tell much about perceptual differences between the two compared images. Much better approach would be to convert luminance values to perceptually linearized luma before computing the RMS metric.

What is the range of luminance and luma values that an imaging pipeline should handle? A reasonable range of luminance is within 10^{-5} cd/m^2 and 10^{10} cd/m^2, which can capture the luminance of both a moonless sky ($3\cdot10^{-5}$ cd/m^2) and the surface of the sun ($2\cdot10^9$ cd/m^2). In Section 5.3.1 we will show that this range of luminance can be encoded as luma using 4096 discrete steps so that the difference between two consecutive steps is not perceivable. This shows that even if the absolute range of luminance the eye can see is impressive, the actual limitation of the HVS does not allow us to see more than about 4000 visually different shades of gray, and those can be seen only if the slow mechanisms of visual adaptation are involved.

2.4 Dynamic Range

Another important and even more confusing quantity used in digital imaging is the dynamic range. The dynamic range is usually understood as a ratio of the highest and the lowest luminance in an image. However, in most imaging systems the lowest luminance is limited by the noise of that system, such as flare in camera lens, ambient light reflected from the screen of a monitor, or noise in a digital photograph. Therefore, the dynamic range is more precisely defined as a ratio of the representative peak signal to the level of noise in an image. For example, if we assume that a computer monitor is almost perfectly black when the pixels are set to zero, which means that luminance of the screen surface is very close to 0 cd/m^2, the dynamic range of such a theoretical monitor is infinitively high (since the peak luminance is divided by a very small number). However, in real-world the minimum luminance of a good quality LCD monitor in a normally lit room is about 1 cd/m^2. If the maximum luminance of a bright LCD display is about 300 cd/m^2, its dynamic range is in fact 1:300. Note that a similar number is often given in the display specifications as the contrast of a display. However, since there are no strict standards how to measure such contrast, those numbers are usually significantly higher than in reality (we found that some displays sold as 400 cd/m^2 peak luminance monitors, can achieve not more than 250 cd/m^2 when they are new and 200 cd/m^2 after two years of operation).

Camera manufactures usually report the dynamic range of a sensor using the ratio of the maximum sensor capacity to the noise level. Such ratio is measured in decibels using formula 4 given in Table 2.1, where N_{max} is the maximum capacity of a well (given in the number of electrons) and RMS_{noise} is the root mean square of noise. RMS_{noise} is sometimes replaced with the capacity (or voltage) at which the Signal to Noise Ratio (SNR) is equal 1, which indicates that the useful signal has the same amplitude as noise. The sensor dynamic range measures are usually only a theoretical maximum dynamic range of a camera, which in practice is limited by other camera's elements, such as lens, an A/D converter, and processing performed before an image is stored. Note that the sensor's SNR values, also commonly reported in decibels, are quite different to the dynamic range measures. SNR tells what is the ratio of signal to noise at the given luminance level and can indicate whether noise is visible at particular illumination conditions.

A different measure of dynamic range is used in the photography. The amount of light that passes through lens and reaches a camera's film or digital sensor is expressed as the *f-number* and written as $f/\#$, where # is the ratio of the focal length and the diameter of the entrance pupil. The sequence of such *f-numbers* that results in halving the amount of light (luminance) reaching the sensor is a sequence of *f-stops*. The *f-stops* form a geometric series of powers of $\sqrt{2}$: f/0.7, f/1, f/1.4, f/2, f/2.8, f/4, f/5.6, f/8, and so on. Therefore, photographers say that a scene has eight *f-stops* instead of saying that a scene has a dynamic range or contrast ratio 1:256. The number of f-stops is called *exposure latitude* and therefore a high dynamic range image is better known in photography as an image of large exposure latitude (refer to item 3 in Table 2.1). The best film stocks offer about 12 f-stops of exposure latitude, which corresponds to about 3.5 log-10 units. This is still lower dynamic range than the one that can be captured with HDR cameras or multi-exposure techniques, but it shows that high dynamic range images are not so new to the photography [Reinhard et al. 2002b]. Yet another measure of dynamic range that can be found in photography is based on the system of *print zones* introduced by Ansel Adams [Adams 1981]. The print zones correspond roughly to f-stop units (they double or halve the amount of captured light), but they are additionally associated with the shades of gray in the resulting print.

The dynamic range measured for analog films is usually expressed as a *density range*. This measure is a difference between the maximum (D-Max) and the minimum (D-Min) tonal values that a film can register (see Figure 2.7 and item 5 in Table 2.1). Since D-Min and D-Max values are measured on a base-10 log scale, the *density range* is equivalent to "orders of magnitude" or log-10 units (see item 2 in Table 2.1). The *density range* of a good quality film is about 3.4D (note the "D" letter indicating density measure).

All measures of dynamic range discussed in this section and summarized in Table 2.1. The last remaining aspect is the dynamic range that can be perceived by the human eye. The light scattering on the optic of the eye can effectively reduce the maximum luminance contrast that can be projected onto to retina to 2–3 log-10 units. However, since the eye is in fact a highly active sensor, which can rapidly change the gaze and locally adapt, people are believed to be able to perceive simultaneously the scenes of 4 or even more log-10 units [Reinhard et al. 2005, Section 6.2] of dynamic range.

Chapter 3

Modelling the Human Visual System

The purpose of this chapter is to briefly introduce the reader to the computational models of the Human Visual System (HVS). Elements of such models are used in the later chapters, for example to build a filter of invisible noise in Section 5.6.5, to design a visual difference predictor for HDR images in Chapter 4, and to derive a transducer function for large contrast magnitudes in Section 6.3.2. The description of the mechanisms of the human vision given in this chapter is neither detailed nor complete, therefore this chapter is more a reference than a complete guide to the computational models of the HVS.

The following sections focus on the quantitative models, rather than the anatomical aspects of vision. An in-depth discussion of the psychophysical and anatomical aspects of vision can be found in several excellent handbooks, such as [Wandell 1995] or [Hood and Finkelstein 1986]. Each section of this chapter gives only short background information on the functionality of particular mechanism, followed by the discussion of models used to predict behavior of that mechanism. Each section gives also several practical applications in which such perceptual models are used.

Figure 3.1 summarizes the content of this chapter by linking each visual mechanism in a complete visual pipeline. The figure contains most of the elements practically used in the computational models of vision. However, the actual models will vary in the selection of elements and in order in which they form a processing pipeline.

3.1 Optics of the Eye

Every optical system found in real world, including the human eye, is imperfect and distorts the light that travel through it. As result of this, the light that passes though the optics of the eye gets scattered and forms a blurred image on the retina. A simulation of such blurring is shown in Figure 3.2. A computer rendered image that exhibits no imperfections of the optics is shown on the left, while the same image but with simulated light scattering in eye's optics is shown on the right. The right image shows

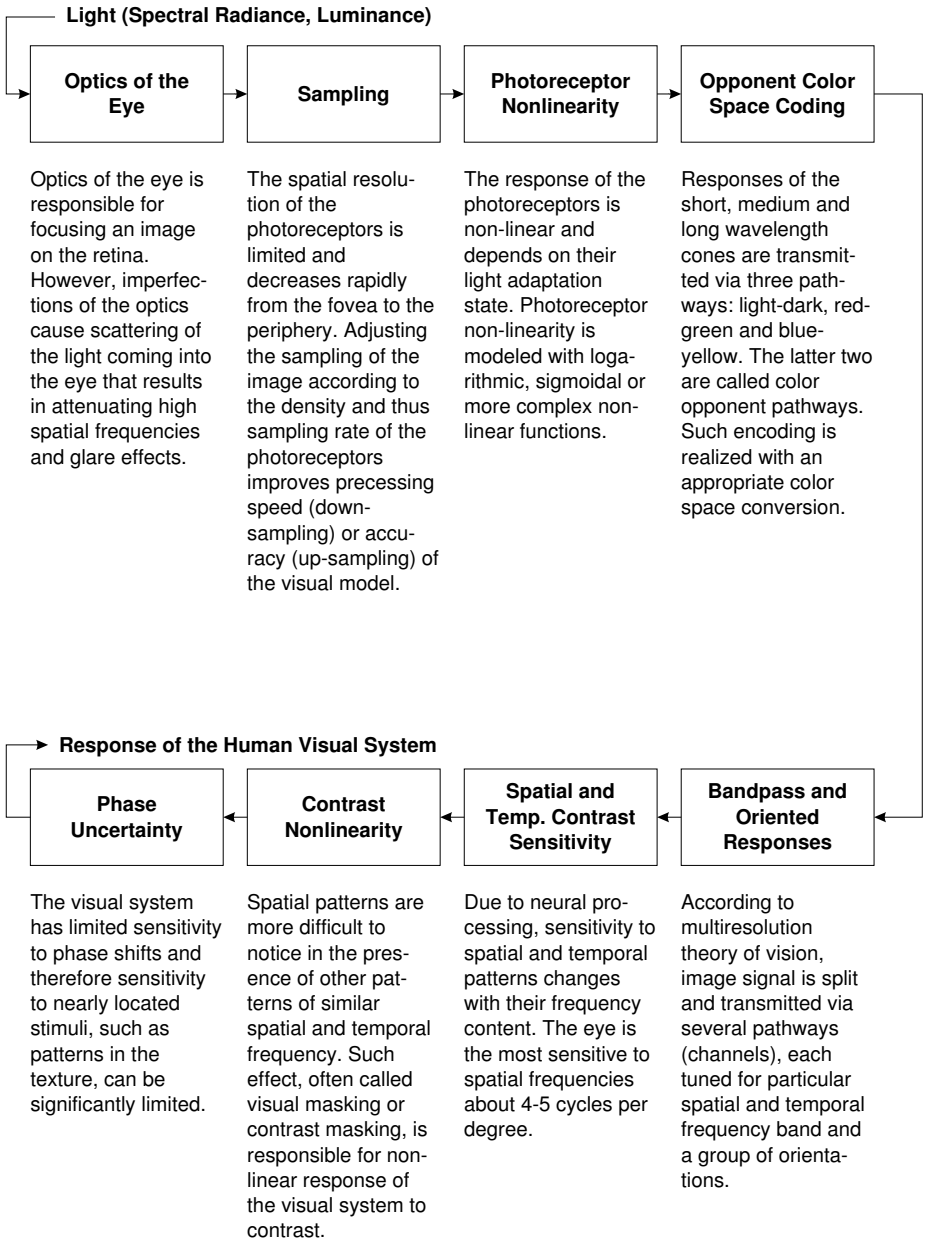


Figure 3.1: A generic data flow of computational models of visual system.

not only glaring effect round the light bulb, but also loss of contrast, especially in the areas near the light.

Figure 3.2: A computer generated image (left) and the same image with simulated blurring due to the optics of the eye (right). Original HDR images have been tone mapped using log-linear mapping. Image courtesy of Jozef Zajac.

Besides the blurring effect, often referred as *blooming*, other effects can be observed. *Flare* is observed as a set of colored, concentric rings (*lenticular halo* surrounding the light sources, and as radial streaks emanating from the center of the light source (*ciliary corona*) [Spencer et al. 1995]. Other effects include diffraction due to a pupil or eyelashes [Nakamae et al. 1990, Kakimoto et al. 2005]. These effects, however, appear when the eye is adapted to dark light conditions and the observed scene includes bright light sources.

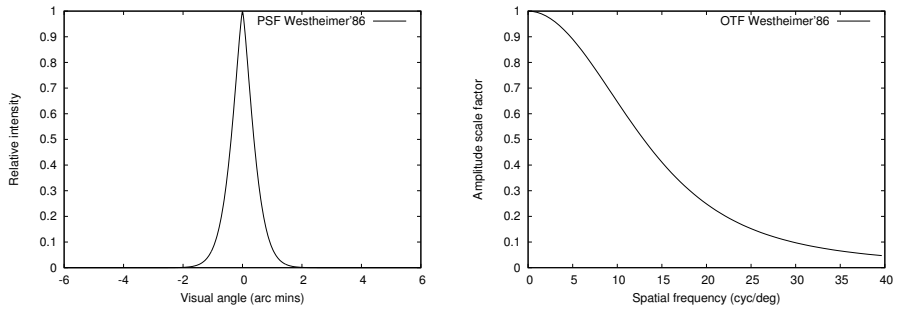


Figure 3.3: An example of the Point Spread Function (left) and the Optical Transfer Function (right) of the human eye. Based on the model from [Westheimer 1986].

The glare effect is usually modelled using a Point Spread Function (PSF) [Westheimer 1986, Spencer et al. 1995] in the spatial domain, or Optical Transfer Function (OTF) in the Fourier domain [Deeley et al. 1991, Marimont and Wandell 1994, Barten 1999]. These are the functions of spatial frequency (OTF) or angular distance (PSF), eccentricity (distance from the foveal region), pupil size, wavelength and defocus of the eye. Usually only few of these attributes are included in the models. The examples of PFS and OTF are shown in Figure 3.3.

Another important limitation of the optics of the eye is *chromatic aberration*. Since the light of different wavelength refracts differently, the eye cannot place all wavelengths at focus at the same time. Usually short wavelengths are projected on the retina out of

focus creating a blurrier image than middle wavelengths. Marimont and Wandell modelled the effect of chromatic aberration in their work on the OTF of the eye [Marimont and Wandell 1994].

A physically plausible simulation of the eye optics distortions became possible with the introduction of HDR imaging. LDR images usually have larger luminance values clipped and are not properly calibrated in luminance units, thus making them unsuitable as a source for glare simulation. Note that glare effects must be simulated in linear units of luminance or radiance, and not in gamma corrected color spaces. Therefore, HDR images, which usually contain linear values of luminance, can be directly used for simulation of the glare effect.

The effects of the optics are simulated in computer graphics to introduce a believable impression of bright light sources on LDR displays [Spencer et al. 1995, Kakimoto et al. 2005]. An HDR display takes advantage of the glare effect to hide the blur of the display [Seetzen et al. 2004]. In Chapter 4 we will show that the optical part of the visual system must be also taken into account in order to predict visible differences in HDR images.

3.2 Sampling

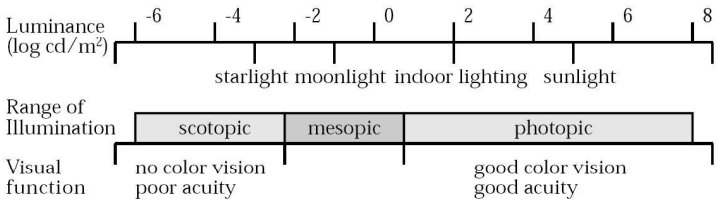


Figure 3.4: The range of luminance in the natural environment and associated visual parameters. From [Ferwerda et al. 1996].

The light that passes through optics of the eye is sampled by two kinds of photoreceptors: *rods*, responsible for low light vision, and *cones* responsible for day-light vision. The range of luminance in which rods operate is called *scotopic*, the range in which cones operate is called *photopic*, and the range in which both rods and cones are active is the *mesopic* range. The mapping of physical luminance to the ranges of photoreceptors' activity is illustrated in Figure 3.4. The day-light and color vision photoreceptors, cones, are further divided into three types, each one sensitive to different wavelengths: L-cones (long wavelengths — red), M-cones (medium wavelengths — green) and S-cones (short wavelengths – blue).

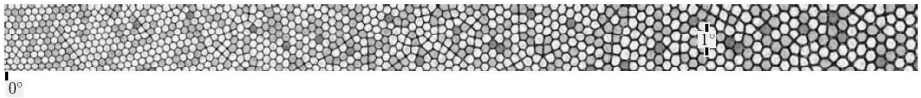


Figure 3.5: A generated pattern of cones, starting from at the foveal center on the left. From [Deering 2005].

Photoreceptors form approximately a hexagonal pattern. The region with the strongest concentration of color-vision receptors (cones) is the *fovea*. As the distance from the fovea (*eccentricity*) increases, the density of cones drops rapidly, and the number of rods increases. The eye achieves the highest acuity, which is the ability to see fine details, in the fovea.

Derring [Deering 2005] has shown that an accurate pattern of the photoreceptors can be synthetized based on a few basic principles (see Figure 3.5). Although such accurate models of the photoreceptor sampling pattern may provide better estimation of images registered by the retina, this is rarely done in practice. The reason for this is that the limitation of the eye's optics are usually stronger that those of the photoreceptor sampling pattern. The effect of sampling pattern can be observed only for very fine patterns of high spatial frequency (≥ 60 cycles per degree) using *visual interferometry* technique, which can project sinusoidal patterns directly on the retina, excluding the effects of optics [Wandell 1995, p. 61]. Sampling is in fact included in some of visual models (e.g. [Lubin 1995]), however mostly for the purpose of limiting the resolution of input images and thus speeding up the computations.

3.3 Photoreceptor Non-linearity

Photoreceptors convert light falling on the retina into neural signals that are relayed to the other parts of the visual system. However, their neural response to light is not linear and strongly depends on their state of adaptation to luminance levels. As a result, the eye is more sensitive to relative luminance levels ($Y/Y_{background}$) than absolute luminance values (Y), and the sensitivity to relative luminance decreases for low luminance conditions. Such effects are sometimes called *luminance masking* (masking by the level of luminance), which should not be confused with the visual masking described in Section 3.7.

The response of the photoreceptors is usually modelled as an S-shaped function (on log-linear plot), known as the Michaelis-Menten or Naka-Rushton equation:

$$\frac{R}{R_{max}} = \frac{Y^n}{Y^n + \sigma^n} \tag{3.1}$$

where R is the photoreceptor response, R_{max} is the maximum response, Y is luminance, σ is the half-saturation constant, and n is the sensitivity control exponent that has value between 0.7 and 1.0.

The half-saturation constant, σ is the value of Y that causes the half-maximum response and it depends on the state of global, local and temporal adaptation. Curves for several values of *sigma* are plotted in Figure 3.6. There are sophisticated models of visual adaptation that can compute the proper value of σ based on an HDR image or HDR video sequence [Pattanaik et al. 2000, Irawan et al. 2005]. If high complexity of those models can not be afforded or lower accuracy of the visual model is acceptable, Equation 3.1 can be replaced with simpler formulas that do not depend on the adaptation state. This is possible by introducing a simplifying assumption that the eye can perfectly adapt to very small patches, such as single pixels. Assuming that $n = 1$, Daly [Daly 1993] proposes a shift invariant model of photoreceptor response:

$$\frac{R}{R_{max}} = \frac{Y}{Y + c_1 Y^b} \tag{3.2}$$

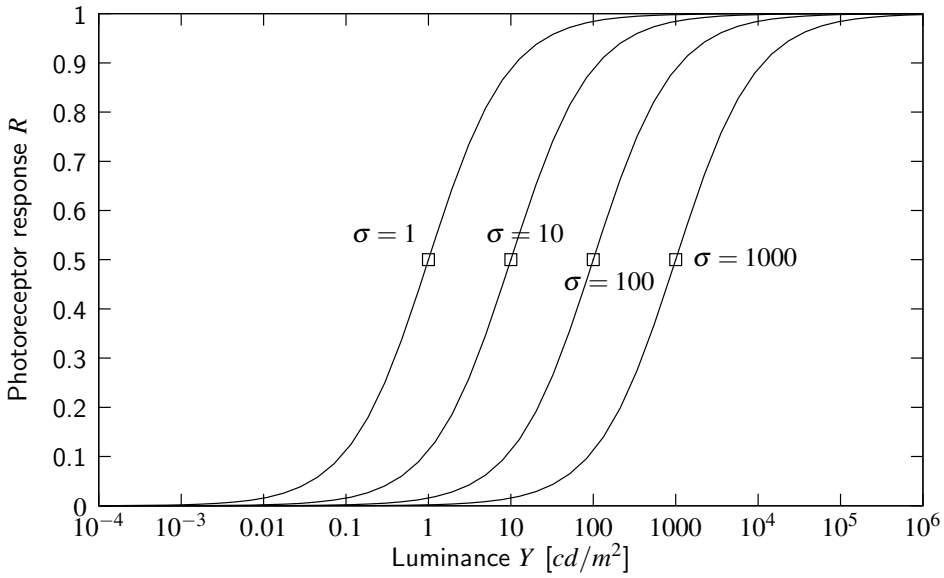


Figure 3.6: Photoreceptor response curves for several values of half-saturation constant *sigma*.

where $b = 0.63$ and $c_1 = 12.6$. Note that the above formula does not depend on the adaptation state. Other authors suggest modelling photoreceptor response as a logarithmic function, which is in agreement with the Weber-Fechner law. The relation between a logarithmic function and the Weber-Fechner law will be discussed in detail in Section 5.3.1. The Weber-Fechner law, however, is a good approximation only for photopic vision and fails for the mesopic light conditions, which cover the working range of most LCD and CRT displays. A better approximation of the response function for luminance range from about 1 cd/m^2 to about 400 cd/m^2 is given by the power function. This is why most formulas for lightness (e.g. *CIE 1976 lightness*) as well as gamma correction used for LDR color spaces [Poynton 2003, Chapter 23] are formulated as power functions with the exponent ranging from $1/3$ to $1/2$. Although these functions estimate response of the entire visual system and not only photoreceptors, their characteristic is mostly affected by the non-linearity of the photoreceptor response.

The logarithmic function can be used to model photoreceptor response in the photopic luminance range and the power function in the mesopic range. However, it would be more convenient to use a single function for the entire range of visible luminance, including scotopic vision. We derive such a response function from psychophysical models in Section 4.2.2. The derived response function is appropriate for the luminance levels that can be found in HDR images.

The non-linearity that comes from photoreceptors' response is used in most *non-linear* (gamma corrected) color spaces, such as sRGB [IEC 61966-2-1:1999 1999] or CIE $L^*u^*v^*$. Such color spaces perceptually linearize physical luminance, so that the resulting values are proportional to our impression of brightness of lightness. This is why compressive non-linearity of luminance is mandatory step in any video/image compression system. It is also employed as the first stage of many visual models, such as

VDP [Daly 1993] or HDR VDP described in Chapter 4.

3.4 Opponent Color Space Coding

There is abundant evidence that the visual system encodes color information as two opponent color pairs: red–green and blue-yellow, instead of directly transmitting the responses of three types of cones (sensitive to red, green, blue hues). This is confirmed by the simple observation that we can perceive mixtures of color coming from both of the opponent color pairs, such as orange and cyan, but we never perceive mixtures of red and green, or blue and yellow [Wandell 1995, p. 318].

The reason for such opponent color encoding is efficiency. The red, green and blue coordinates of colors that can be found in real world are strongly correlated with each other. A standard technique used to decorrelate multidimensional data (in this case three-dimensional) is the Principal Component Analysis (PCA). PCA performed on a large number of natural images results in three principal components: luminance, red–green and blue-yellow color channels. The visual system has evolved to efficiently encode color found in real word, and therefore, not surprisingly, it uses the dimensions close to the principal components to encode color. Such encoding significantly reduces the amount of information that needs to be send to the brain.

All color spaces used for image and video compression use a variant of an opponent color space, such as YC_RC_B. CIE uniform color spaces, CIE 1976 $L^*u^*v^*$ and CIE 1976 $L^*a^*b^*$, have color components oriented along red–green and blue–yellow dimensions. The same opponent encoding is also used in visual models [Jin et al. 1998, Bolin and Meyer 1998, Pattanaik et al. 1998].

3.5 Bandpass, Oriented and Temporal Responses

The widely accepted multiresolution theory claims that the images registered by the retina are transmitted to the brain via several visual channels, each one carying information about different spatial/temporal frequency band, orientation and color. An example of such multi-resolution representation, excluding temporal and color aspects, is illustrated in Figure 3.7. The multiresolution theory explains several aspects of visual system and can also help constructing more accurate models of vision.

There are several computational models of visual channels, which differ from each other by the conformity with the psychophysical measurements and their computational cost. The most psychophysically plausible models are based on *Gabor functions*. These, however, are expensive to compute, non-invertible and suffer from numerous problems discussed in detail in [Daly 1993]. Subband transforms, such as *QMF subband transform* [Simoncelli and Adelson 1989] or *wavelet transforms*, are computationally more efficient. The shortcoming of these representations is that they usually represent the oriented responses for 45° and 135° as a single channel, which may lead to significant failures if the models are used to predict visual masking effects [Zeng et al. 2000]. The *cortex transform* [Watson 1987] is a representation that offers a plausible match with the psychophysical measurements and can be computed efficiently,

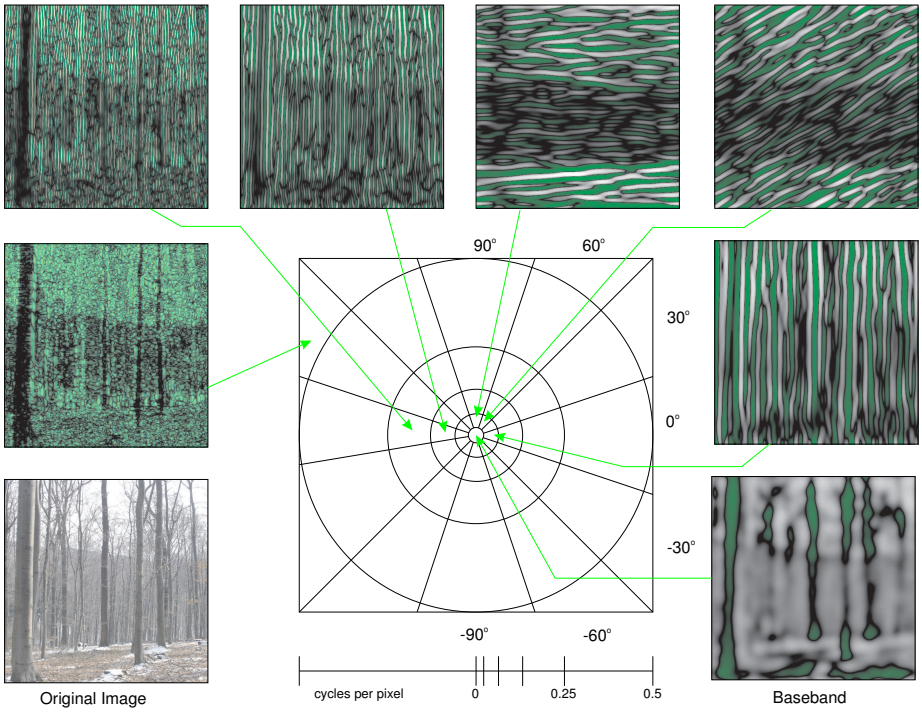


Figure 3.7: Cortex transform decomposition. The diagram in the middle represents and image in the Fourier domain divided into six spatial and six orientational bands. The images around show content of particular bands in the spatial domain.

although it is computationally more expensive than the subband transforms. The transform consist of a set of frequency and orientation selective filters in the Fourier domain, which decompose an image into several channels, as illustrated in Figure 3.7. The cortex transform is used in the Visual Difference predictor and its HDR extension described in Section 4.

Multiresolution representation of images are commonly used in image compression and processing. Pyramids, such as the Gaussian pyramid or the Laplacian pyramid, loosely correspond to multiresolution models and are basic tools of image processing [Gonzalez and Woods 2001]. Also the wavelet representation of images shares many similarities with multiresolution models of vision. In Chapter 6 we explore the problem of multiresolution representation of images in more detail and propose a novel representation that is especially effective for image processing (such as tone-mapping) that is free of artifacts.

3.6 Spatial and Temporal Contrast Sensitivity

The sensitivity of the visual system to contrast varies with its spatial and temporal frequency, orientation, wavelength, adaptation luminance and several other factors. Some of these effects can be explained by the influence of the eye's optics and photoreceptor response non-linearity, discussed in Sections 3.1 and 3.3, but others come from the

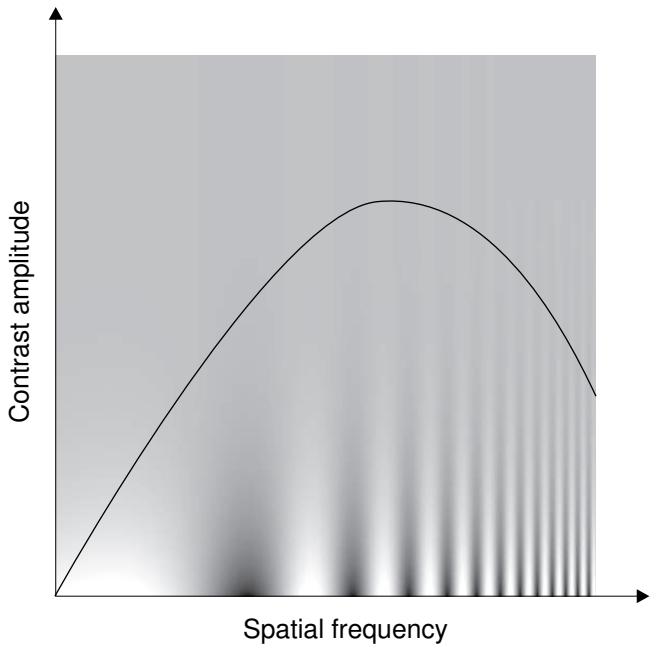


Figure 3.8: An approximate characteristic of contrast sensitivity function shown on the Campbell-Robson contrast sensitivity pattern [Campbell and Robson 1968]. When observed from the proper distance, the sinusoidal pattern reveals the CSF characteristic of the eye.

neural mechanisms of the retina. The sensitivity of the entire visual system, including optical and neural parts, is described by the Contrast Sensitivity Function (CSF). An example of the CSF is shown in Figure 3.8. The sensitivity is defined as an inverse of the threshold given as Michelson's contrast (for definition of the Michelson's contrast, refer to Section 6.2.1, Table 6.2).

The CSF model is almost obligatory part of any visual model or algorithm that takes advantage of the limitations of the visual system. The quantization matrix used in DCT-based image and video compression is largely affected by the CSF. Sub-sampling of chrominance channels used in compression is dictated by lower sensitivity of the eye for the high-frequency color patterns, which is described by the CSF. Computational models of vision usually incorporate the CSF in either of two ways: either they weight each visual channel (see Section 3.5) by a weighting factor coming from the CSF or they use the CSF as a filter in the Fourier domain. The first approach offers limited accuracy in terms of spatial resolution since each visual channels spans a broad range of luminance in which the CSF sensitivity differs. The second approach assumes that the CSF is shift-invariant, which may not be true especially if the mechanism of local adaptation are involved. A method that can compute the influence of the CSF in Fourier domain, including the effects of local adaptation, is described in Section 4.2.3.

The most often cited CSF is the function proposed by Barten [Barten 1999]. Barten built his model by carefully designing each source of noise in the HVS and then fitting the data from several phychophysical measurements to the model. The model is however limited to photopic luminance conditions. We also found that it does not give

accurate predictions for large luminance of adaptation values, exceeding 1000 cd/m^2. We found that the CSF used in Daly's Visual Difference Predictor [Daly 1993], which is based on Meeteren's CSF model [Van Meeteren and Vos 1972] and improved by Kodak, gives more reliable prediction for a broad range of lighting conditions. There is very limited amount of data on the CSF for color data (stimuli different than the modulation of luminance), and most visual models are based on the paper by Mullen [Mullen 1985]. It is important to notice that Mullen measured the color CSF for the corrected effect of *chromatic aberration*, therefore using her model requires adding this effect in earlier stages of the visual model (see Section 3.1). Many models, with the notable exception of [Bolin and Meyer 1998], ignore this fact and use Mullen's measurements improperly, without modelling the chromatic aberration effect.

3.7 Contrast Non-linearity

Figure 3.9: An example of visual masking (contrast masking). The original image (left) has been distorted with random noise (right). The noise is visible mostly in the flat regions of the sky, where it is not masked with high frequencies of the grass and the trees. Image courtesy of Grzegorz Krawczyk.

The effects that are the result of non-linear response of the HVS to contrast are known as *visual masking* or *contrast masking*. Visual masking occurs when the stimuli that is normally visible becomes invisible in the presence of another stimuli. This is illustrated in Figure 3.9, where random noise was added to the image on the right. Although the noise is equally distributed across the entire image, it is only objectionable in the sky, where it is not masked with high frequency pattern of the grass and the trees.

The visual masking effect is the strongest when the masking signal has similar frequency, orientation and color as the masked signal. This is usually modelled using multiresolution representations (see Section 3.5) by applying a masking function to each visual channel separately. Some amount of masking can be also observed between visual channels of different frequency and orientation, however these effects are much weaker than the inter-channel masking and are rarely included in the models.

When the contrast of the masker is close to the contrast of the signal, the detection of the signal can improve. However, this effect, called *facilitation*, can be rarely observed in natural images and therefore is usually not included in the models. The facilitation effect is observed in similar settings as the crispening effect [Whittle 1986].

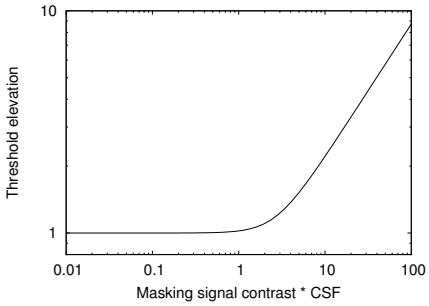


Figure 3.10: Threshold elevation function.

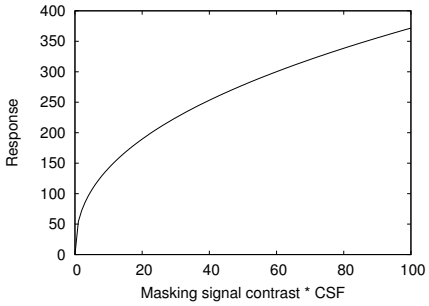


Figure 3.11: Contrast transducer function.

There are two major approaches to modelling of visual masking: the *threshold elevation function* and the *transducer function*. The threshold elevation function, drawn in Figure 3.10, tells how many times the detection threshold will increase in the presence of the masking signal of a given contrast. Although the shape of such threshold elevation function vary across spatial frequencies, a single function can be used if the contrast of masking signal is normalized by the sensitivity predicted by the CSF [Daly 1993]. The threshold elevation function is commonly used in visual difference predictors, such as Daly's VDP [Daly 1993] and HDR VDP described in Chapter 4. Transducer function [Wilson 1991], on the other hand, applies non-linearity to the masked contrast in order to convert it to the response of the HVS. Such response can be scaled in the JND units, so that the difference of one unit in the response corresponds to one Just Noticeable Difference. Transducer function is employed in visual difference predictors, such as Sarnoff's VDP [Lubin 1995], is used for visual optimization of JPEG2000 [Zeng et al. 2000] compression and in the models of visual masking developed for the purpose of computer graphics applications [Ferwerda et al. 1996, Bolin and Meyer 1998]. In Section 6.3.2 we derive a contrast transducer function that is especially suitable for large contrast magnitudes found in HDR images.

3.8 Phase Uncertainty

Most of the models presented in the previous sections are focused on the transmission of amplitudes in the HVS, without paying much attention to the transmission of phase. For example, both OTF and CSF can predict how amplitudes are modulated when passing through the optics of the eye or the retina, but they do not model changes in phase of the signal. Figure 3.12 illustrates the importance of phase in image interpretation. If only amplitude is preserved and phase is discarded, the image contains only noise and is not recognizable. However the major features of the image can be recognized if only phase is preserved, even though the amplitude is set to zero (zero-response MTF).

Although image phase plays important role in image recognition, the sensitivity of the HVS to phase distortions is limited. This is illustrated in Figure 3.13, in which the phase of selected frequency bands has been shifted by angles ranging from 15° to 180°. As can be seen in this image, it is difficult to notice phase shifts smaller than 90°. This is because the HVS has limited phase sensitivity (phase uncertainty), which

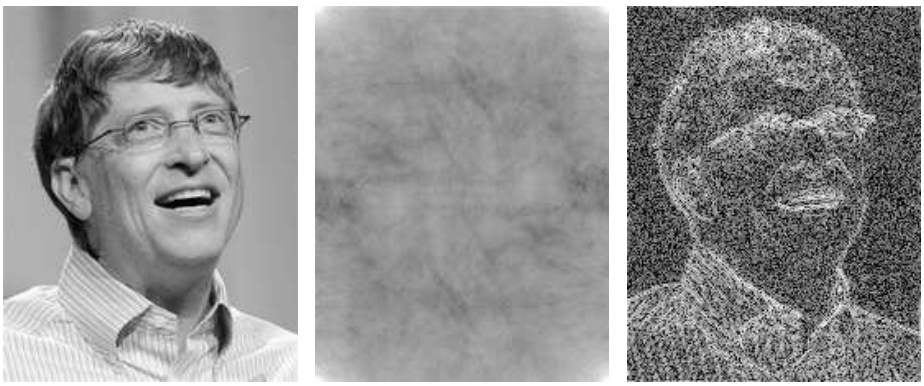


Figure 3.12: The role of phase and amplitude in the perception of images. The original image (left) has been converted to the Fourier domain and then converted back to the spatial domain, but using only amplitude (center) or phase (right) data.

can range from 15° to 90°, depending on the bandwidth of affected frequencies, the power content of the frequency bands and the content of the neighboring frequency bands [Caelli et al. 1985].

Phase sensitivity (or phase uncertainty) plays important role in the accurate prediction of visual masking. It facilitates distinction between textured regions, where masking is high, and regions width edges, where masking is low. This is illustrated in Figure 3.8, which shows a signal that contains both the texture and the edges on a flat surface. The signal is band-pass filtered to produce response of a single visual channel. If such response is directly used to predict visual masking, the masking is the highest at the edges, where the amplitude is the highest. Also, masking would be limited to the peaks of the bandpass filtered signal, while zero-crossings would result in lack of masking. However, because masking due to edges is much lower, and because masking for textures extends over the whole area of the texture and not only the peaks of the signal amplitude, some visual models apply a filter that simulates phase uncertainty of the HVS. To produce the response shown as a green line in Figure 3.8, we used a non-linear filter, similar to that proposed in JPEG 2000 [Zeng et al. 2000]:

$$r_i = \frac{a}{Card(\Omega_i)} \sum_{k \in \Omega_i} |y_k|^{\beta} \tag{3.3}$$

where r is the channel response with the phase uncertainty effect, y is bandpass image value, indices i and k denote pixel location, Ω_i is the neighborhood of the pixel i, $Card(\Omega_i)$ is the number of pixels that belong to the neighborhood and a is a normalization factor. Parameter β is usually set to a small value, such as 0.2. Such nonlinear filter should be applied to a single visual channel. In JPEG 2000 the filter is applied to wavelet coefficients [Zeng et al. 2000]. In Daly's VDP [Daly 1993] a similar filter is applied to band pass and orientation filtered images, which are the result of the cortex transform and which represent visual channels (refer to Section 3.5). The application of such filters greatly improves prediction of masking, especially at the edges of smooth surfaces and in textured regions.

Figure 3.13: Visibility of the phase shift distortion. A 2-octave frequency band has been distorted by shifting its phase by 15°, 45°, 60°, 90°, 135° and 180°. The other frequencies have not been modified. Although phase plays important role in image recognition, the sensitivity of the HVS to phase is limited. The phase shift distortions start to be noticeable for the images above at 90° and larger shifts.

3.9 Threshold and Supra-threshold Effects

When considering models of visual system it is important to distinguish between threshold and supra-threshold effects.

Threshold or *subthreshold* effects are those that can be observed at very small magnitude of the stimuli, usually at the contrast at which the stimuli is barely visible. The CSF, described in Section 3.6, predicts the performance of the visual system only for small contrast, close to the *contrast detection threshold*. The detection threshold, illustrated in Figure 3.15, is the smallest amplitude of contrast that makes the stimuli just noticeable.

Suprathreshold effects, on the other hand, are those that consider contrast magnitudes significantly larger than the detection threshold. The visual masking (contrast masking) is such a supra-threshold effect. To measure visual masking, the smallest increments and decrements for supra-threshold stimuli are identified in so called contrast discrimination experiments, as shown in Figure 3.15. Such experiments measure the smallest difference of amplitude of sinusoidal patterns that is distinguishable for a human observer.

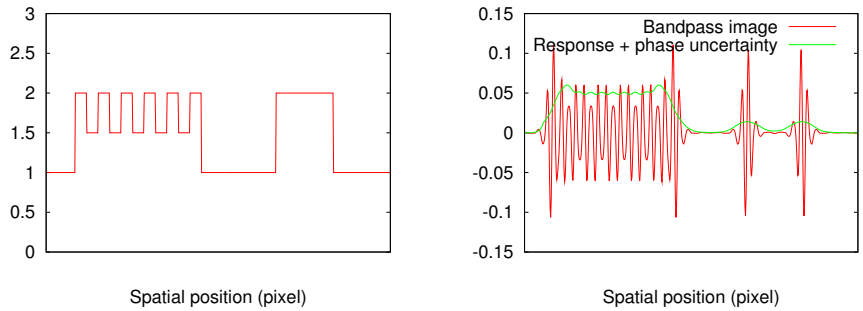


Figure 3.14: The effect of phase sensitivity on the prediction of masking signal. Left pane: an original signal containing the texture on the left and two edges on the right. Right pane: bandpass image of the signal (red) and the response predicted with the phase uncertainty (green). Note that the response due to the edges is much lower than due to the texture, although bandpass amplitude due to edges is higher.

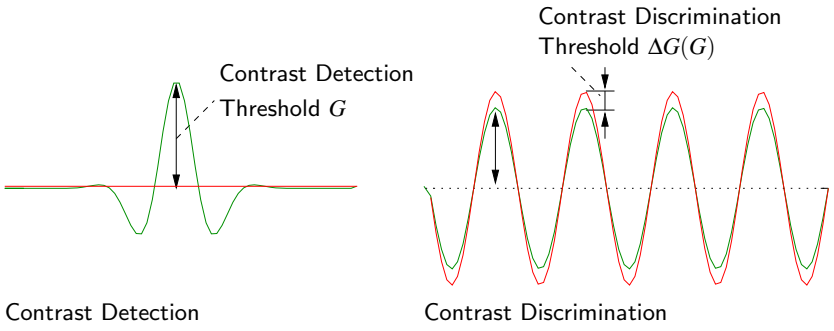


Figure 3.15: The luminance profile of the stimuli used for contrast detection and contrast discrimination measurements. The test (red) and the standard (green) stimulus are displayed one after another. The threshold is the smallest visible contrast (detection) or a difference of contrast (discrimination).

Chapter 4

A Visual Difference Predictor for HDR Images

When designing an image synthesis or processing application, it is desirable to measure the visual quality of the resulting images. To avoid tedious subjective tests, where a group of people has to assess the quality degradation, objective visual quality metrics can be used. The most successful objective metrics are based on models of the Human Visual System (HVS) and can predict such effects as a non-linear response to luminance, limited sensitivity to spatial and temporal frequencies, and visual masking [Nadenau 2000].

Most of the objective quality metrics have been designed to operate on video and images that are to be displayed on CRT or LCD displays. While this assumption seems to be clearly justified in case of low-dynamic range images, it poses problems as new applications that operate on HDR data become more common. A perceptual HDR quality metric could be used for the validation of the aforementioned HDR image and video encodings. Another application may involve steering the computation in a realistic image synthesis algorithm, where the amount of computation devoted to a particular region of the scene would depend on the visibility of potential artifacts.

In this chapter we propose several modifications to the original Visual Difference Predictor [Daly 1993]. The modifications improve a prediction of perceivable differences in the full visible range of luminance. This extends the applicability of the original metric from a comparison of displayed images (compressed luminance) to a comparison of real-world scenes of measured luminance (HDR images). The proposed metric does not rely on the global state of eye adaptation to luminance, but rather assumes local adaptation to each fragment of a scene. Such local adaptation is essential for a good reduction of contrast visibility in High-Dynamic Range (HDR) images, as a single HDR image can contain both dimly illuminated interior and strong sunlight. For such situations, the assumption of global adaptation to luminance does not hold.

The following sections give a brief overview of the objective quality metrics (Section 4.1), describe the modifications to VDP (Section 4.2) and then calibrate the parameters of the proposed metric based on psychophysical data collected in an experiment on an HDR display (Section 4.3). Finally, the predictions of the HDR VDP and the

original Daly's VDP are compared. This chapter is an extended revision of the work published in [Mantiuk et al. 2004b] and [Mantiuk et al. 2005a].

4.1 Previous Work

Several visual difference metrics for digital images have been proposed in the literature [Barten 1990, Daly 1993, Heeger and Teo 1995, Lubin 1995, Taylor et al. 1997, Wang and Bovik 2002, Zetzsche and Hauske 1989, Ramasubramanian et al. 1999]. They vary in complexity and in the visual effects they can predict. However, no metric proposed so far was intended to predict visible differences in High-Dynamic Range images. If a single metric can accurately predict differences for either very dim or bright light conditions, it may fail on images that contain both very dark and very bright areas.

Two of the most popular metrics that are based on models of the HVS are Visual Difference Predictor (VDP) [Daly 1993] and Sarnoff Visual Discrimination Model [Lubin 1995]. Their predictions were shown to be comparable and the results depended on test images, therefore, on average, both metrics performed equally well [Li et al. 1998]. We chose VDP as a base of our HDR quality metric because of its modularity and thus good extensibility.

4.2 Visual Difference Predictor

In this section we describe our modifications to the original VDP, which enable the prediction of visible differences in HDR images. In this chapter we give only a brief overview of the original VDP and focus on the extension to high-dynamic range images. For detailed description of the VDP, refer to [Daly 1993].

The data flow diagram of the VDP for high-dynamic range images (HDR VDP) is shown in Figure 4.1. The HDR VDP receives a pair of images as an input (original and distorted, for example by image compression) and generates a map of probability values, which indicates how likely the differences between those two images are perceived. Both images should be scaled in the units of luminance. In case of low-dynamic range images, pixel values should be inverse gamma corrected and calibrated according to the maximum luminance of the display device. In case of HDR images no such processing is necessary, however luminance should be given in cd/m^2.

The first three stages of HDR VDP model behavior of the optics and retina. The original image is filtered by Optical Transfer Function (OTF), which simulates light scattering in the cornea, lens, and retina. To account for the nonlinear response of photoreceptors to light, the amplitude of the signal is nonlinearly compressed and expressed in the units of Just Noticeable Differences (JND). Because HVS is less sensitive to low and high spatial frequencies, the image is then filtered by Contrast Sensitivity Function (CSF). Those three stages are mostly responsible for contrast reduction in the HVS and are described in detail in the following Sections 4.2.1, 4.2.2, and 4.2.3. The next two computational blocks – the cortex transform and visual masking – decompose the image into spatial and orientational channels and predict perceivable differences in each channel separately. Phase uncertainty further refines the prediction of masking by removing dependence of masking on the phase of the signal. Since the visual masking

does not depend on luminance of a stimuli, this part of the VDP is left unchanged, except for a minor modification in the normalization of units (details in Section 4.2.4). In the final error pooling stage the probabilities of visible differences are summed up for all channels and a map of detection probabilities is generated. This step is the same in both versions of the VDP.

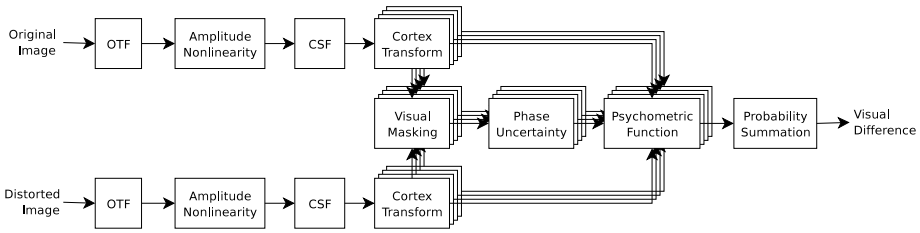


Figure 4.1: Data flow diagram of the High Dynamic Range Visible Difference Predictor (HDR VDP)

4.2.1 Optical Transfer Function

Due to scattering of light in the cornea, lens and retina, the visibility of low contrast details is significantly reduced in the presence of bright light sources. For example, it is very difficult to see the license plate number at night if the head lamps of the car are on. While such dramatic contrast changes are uncommon for typical LCD for CRT displays, they have significant influence on perception of real life scenes or images seen on HDR displays. To account for this effect, the first stage of HDR VDP simulates light scattering in the human eye for given view conditions.

Light scattering in the optics is usually modeled as Optical Transfer Function (OTF) in the Fourier domain or as Point Spread Function (PSF) in the spatial domain. The scattering depends on a number of parameters, such as spatial frequency, wavelength, defocus, pupil size, iris pigmentation, and age of the subject. Because we would like to limit the number of parameters to what is needed for our application, we choose the function of Deeley et al. [1991], which models OTF for monochromatic light and which takes into account a luminance adaptation level. The OTF of this model is given by:

$$OTF(\rho,d) = \exp[-(\frac{\rho}{20.9-2.1d})^{1.3-0.07d}] \tag{4.1}$$

where d is a pupil diameter in mm and ρ is spatial frequency in cycles per degree. Specifically, the luminance level is taken into account via its effect on the pupil diameter, calculated for particular adaptation luminance using the formula of Moon and Spencer[Moon and Spencer 1944]:

$$d = 4.9 - 3\tanh[0.4(\log y_{adapt} + 1.0)] \tag{4.2}$$

where Y_{adapt} is a global adaptation level in cd/m^2. Figure 4.2 shows OTFs for several levels of adaptation. The global adaptation level can be calculated as an average luminance of an image in log domain or supplied to the VDP as an external parameter.

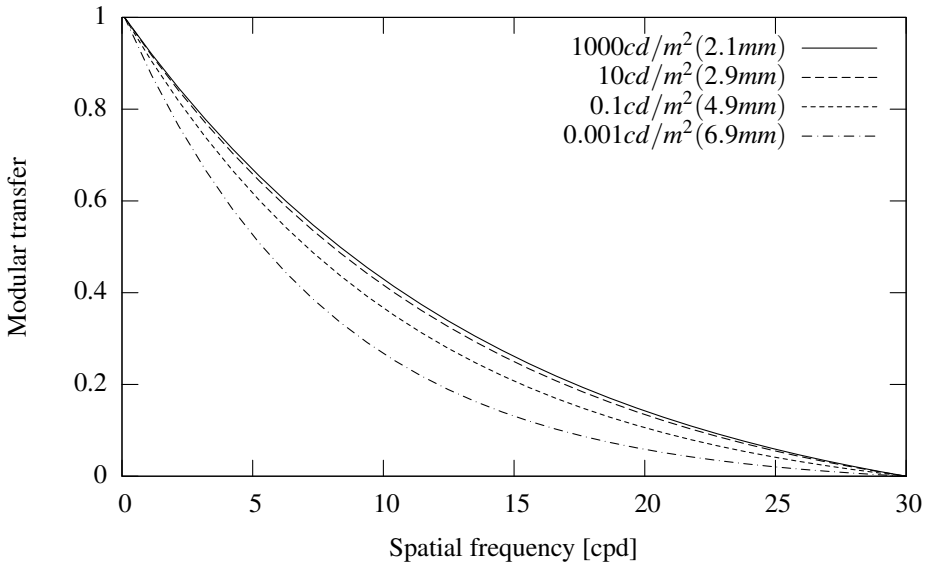


Figure 4.2: Optical MTFs from the model of Deeley et al. [Deeley et al. 1991] for different levels of adaptation to luminance and pupil diameters (given in parenthesis).

4.2.2 Amplitude Nonlinearity

The original VDP utilizes a model of the photoreceptor to account for non-linear response of the HVS to luminance, as illustrated in Figure 4.3. Such non-linear response to luminance is responsible for the effect called luminance masking (see Section 3.3). Perceivable differences in bright regions of a scene would be overestimated without taking into account this non-linearity. The drawback of using the model of the photoreceptor is that it gives arbitrary units of response, which are loosely related to the threshold values of contrast sensitivity studies. The Contrast Sensitivity Function (CSF), which is responsible for the normalization of contrast values to JND units in the original VDP, is scaled in physical units of luminance contrast. Therefore using a physical threshold contrast to normalize response values of the photoreceptor may give an inaccurate estimate of the visibility threshold. Note that the response values are non-linearly related to luminance. Moreover, the model of photoreceptor, which is a sigmoidal response function (see Figure 4.3), assumes equal loss of sensitivity for low and high luminance levels, while it is known that the loss of sensitivity is generally observed only for low luminance levels[1] (see Figure 4.4). Even if the above simplifications are acceptable for low-dynamic range images, they may lead to significant inaccuracies in case of HDR content.

Instead of modeling the photoreceptor response, we propose converting luminance values to a non-linear space that is scaled in JND units. Such space should guarantee the property that adding or subtracting a value of 1 in this space results in a just perceivable

[1]The loss of sensitivity is generally not observed for higher levels of luminance if the eye is adapted to those levels. However, drop of sensitivity can be expected if the eye is adapted to significantly lower luminance than the stimuli. For example there is significant loss of sensitivity for specular highlights in natural images, as the eye is usually adapted to the luminance of an object instead of highlight.

change of brightness. If $y = \psi(l)$ is a function that converts values in JND-scaled space to luminance, we can rewrite the required property as:

$$\psi(l+1) - \psi(l) = tvi(y_{adapt}) \tag{4.3}$$

where *tvi* is a *threshold versus intensity* function and y_{adapt} is adaptation luminance. The function *tvi* predicts a minimum difference of luminance that is visible to a human observer. As we show later in Section 5.3.1, the above formulation makes this problem very similar to the derivation of the luminance encoding for HDR image and video compression, although the required properties are different. We give a short derivation for completeness below and more a detailed description in Section 5.3.1.

We use the Taylor series expansion:

$$\psi(l+1) = \psi(l) + \frac{d\psi(l)}{dl} + \ldots \tag{4.4}$$

to replace the left side of Equation 4.3 with its first-order approximation:

$$\frac{d\psi(l)}{dl} = tvi(y_{adapt}) \tag{4.5}$$

Assuming that the eye can adapt to a single pixel of luminance y as in [Daly 1993] (see also Section 5.3.1), that is $y_{adapt} = y = \psi(l)$, the equation can be rewritten as:

$$\frac{d\psi(l)}{dl} = tvi(\psi(l)) \tag{4.6}$$

Finally, the function $\psi(l)$ can be found by solving the above differential equation. In the VDP for HDR images we have to find a value of l for each pixel of luminance y, thus we do not need function ψ, but its inverse ψ^{-1}. This can be easily found since the function ψ is strictly monotonic.

The inverse function $l = \psi^{-1}(y)$ is plotted in Figure 4.3 together with the original model of photoreceptor. The function properly simulates the loss of sensitivity for scotopic levels of luminance (compare with Figure 4.4). For the photopic luminance, the function has logarithmic response, which corresponds to Weber's law.

The actual shape of the *threshold versus intensity* (*tvi*) function has been extensively studied and several models have been proposed [Ferwerda et al. 1996, CIE 1981]. To be consistent with the original VDP, we derive a *tvi* function from the CSF used there. We find values of the *tvi* function for each adaptation luminance y_{adapt} by looking for the peak sensitivity of the CSF at each y_{adapt}:

$$tvi(y_{adapt}) = P \cdot \frac{y_{adapt}}{\max_\rho CSF(\rho, y_{adapt})} \tag{4.7}$$

where ρ denotes spatial frequency. Similarly as in the the original VDP, parameter P is used to adjust the absolute peak contrast threshold. The optimal value of the parameter P for HDR VDP is calibrated to psychophysical data in Section 4.3. A function of relative contrast – *contrast versus intensity* ($cvi = tvi/y_{adapt}$) – is often used instead of *tvi* for a better data presentation. The *cvi* function for *tvi* derived by us is plotted in Figure 4.4.

In our HDR VDP we use a numerical solution of Equation 4.6 and a binary search on this discrete solution to convert luminance values y to l in JND-scaled space. The subsequent parts of the HDR VDP operate on l values.

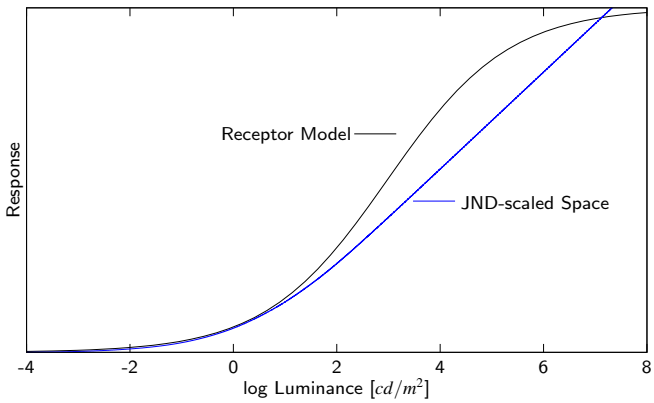


Figure 4.3: Response curve of the receptor model used in the original VDP (continuous line) and mapping to JND-scaled space used in our HDR extension of the VDP (dashed line). The sigmoidal response of the original receptor model (adaptation to a single pixel) overestimates contrast at luminance levels above 10 cd/m^2 and compresses contrast above 10 000 cd/m^2. Psychophysical findings do not confirm such luminance compression at high levels of luminance. Another drawback of the receptor model is that the response is not scaled in JND units, so that CSF must be responsible for proper scaling of luminance contrast.

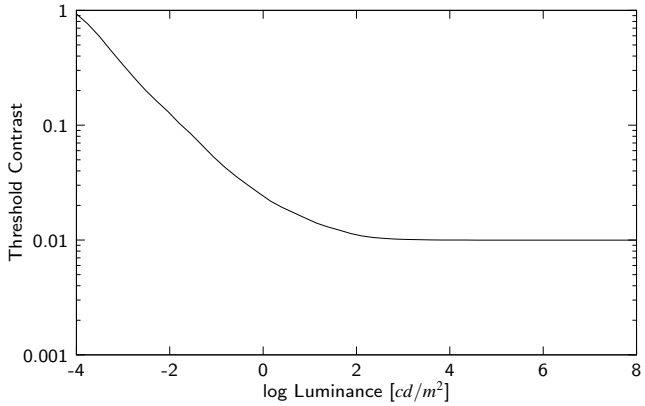


Figure 4.4: *Contrast versus intensity cvi* function predicts the minimum distinguishable contrast at a particular adaptation level. It is also a conservative estimate of a contrast that introduces a Just Noticeable Difference (JND). The higher values of the *cvi* function at low luminance levels indicate the loss of sensitivity of the human eye for low light conditions. The *cvi* curve shown in this figure was used to derive a function that maps luminance to JND-scaled space.

4.2.3 Contrast Sensitivity Function

The Contrast Sensitivity Function (CSF) describes the loss of sensitivity of the eye as a function of spatial frequency and adaptation luminance. It was used in the previous section to derive the *tvi* function. In the original VDP, the CSF is responsible for both modeling the loss of sensitivity and normalizing contrast to JND units. In

our HDR VDP, normalization to units of JND at the CSF filtering stage is no longer necessary as the non-linearity step has already scaled an image in JND units (refer to the previous section). Therefore the CSF should predict only the loss of sensitivity for low and high spatial frequencies. The loss of sensitivity in JND-scaled space can be modeled by a CSF that is normalized by peak sensitivity for particular adaptation luminance:

$$CSF_{norm}(\rho, y_{adapt}) = \frac{CSF(\rho, y_{adapt})}{max_{\rho} CSF(\rho, y_{adapt})} \tag{4.8}$$

Unfortunately, in case of HDR images, a single CSF can not be used for filtering an entire image since the shape of the CSF significantly changes with adaptation luminance. As can be seen in Figure 4.5, the peak sensitivity shifts from about 2 $cycles/degree$ to 7 $cycles/degree$ as adaptation luminance changes from scotopic to photopic. To normalize an image by CSF function taking into account different shapes of CSF for different adaptation levels, a separate convolution kernel should be used for each pixel. Because the support of such convolution kernel can be rather large, we use a computationally more effective approach: we filter an image in the Fourier domain several times, each time using CSF for different adaptation luminance. Then, we convert all of the filtered images to the spatial domain and use them to linearly interpolate pixel values. We use luminance values from the original image to determine the adaptation luminance for each pixel (assuming adaptation to a single pixel) and thus to choose filtered images that should be used for interpolation. A more accurate approach would be to compute the adaptation map [Yee and Pattanaik 2003], which would consider the fact that the eye can not adapt to a single pixel. A similar approach to non-linear filtering, in case of a bilateral filter, was proposed in [Durand and Dorsey 2002a]. The process of filtering using multiple CSFs is shown in Figure 4.6.

As can be seen in Figure 4.5, the CSF changes its shape significantly for scotopic and mesopic adaptation luminance and remains constant above 1 000 cd/m^2. Therefore it is usually enough to filter an image using a CSF for $y_{adapt} = \{0.0001, 0.01, 1, 100, 1000\}$ cd/m^2. The number of filters can be further limited if the image has a lower dynamic range.

CSF predicts the behavior of the complete visual system, including the optical and neuronal parts. The optical part is however already simulated in the HDR VDP pipeline as OTF filtering (see Section 4.2.1). Therefore, only the neural part should play role at this stage of HDR VDP. To extract neural part from the overall CSF, the CSF used in HDR VDP is divided by the OTF.

4.2.4 Other Modifications

An important difference between the original VDP and the proposed extension for HDR images is that the first one operates on CSF normalized values and the latter one represents channel data in JND-scaled space. Therefore, in case of the VDP for HDR images, original and distorted images can be compared without any additional normalization and scaling. This is possible because a difference between the images that equals one unit in JND-scaled space gives a probability of detection equal to one

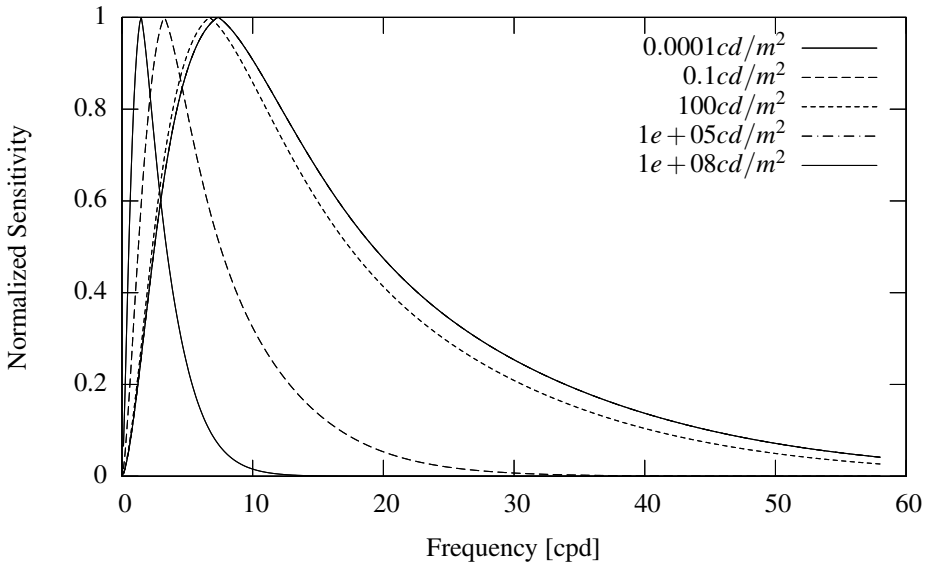


Figure 4.5: Family of normalized Contrast Sensitivity Functions (CSF) for different adaptation levels. The peak sensitivity shifts towards lower frequencies as the luminance of adaptation decreases. Shape of the CSF does not change significantly for adaptation luminance above 1 000 cd/m^2.

JND, which is exactly what this step of the VDP assumes. Therefore the local contrast difference in the original VDP:

$$\Delta C_{k,l}(i,j) = \frac{B1_{k,l}(i,j)}{\overline{B_K}} - \frac{B2_{k,l}(i,j)}{\overline{B_K}} \tag{4.9}$$

in case of the VDP for HDR images becomes:

$$\Delta C_{k,l}(i,j) = B1_{k,l}(i,j) - B2_{k,l}(i,j) \tag{4.10}$$

where k,l are channel indices, i,j pixel coordinates and $B1,B2$ are corresponding local contrast values of the channel for the target and mask images.

4.2.5 Implementation

The source code of HDR VDP is available under the GPL license and can be downloaded from the web page http://hdrvdp.sourceforge.net/. It is integrated with *pfstools* package, which can read most of the HDR file formats. The software provides a ready-to-use metric that can be used in a broad range of digital imaging applications, ranging from validation of computer graphics algorithms to detection of atrifacts in compressed images.

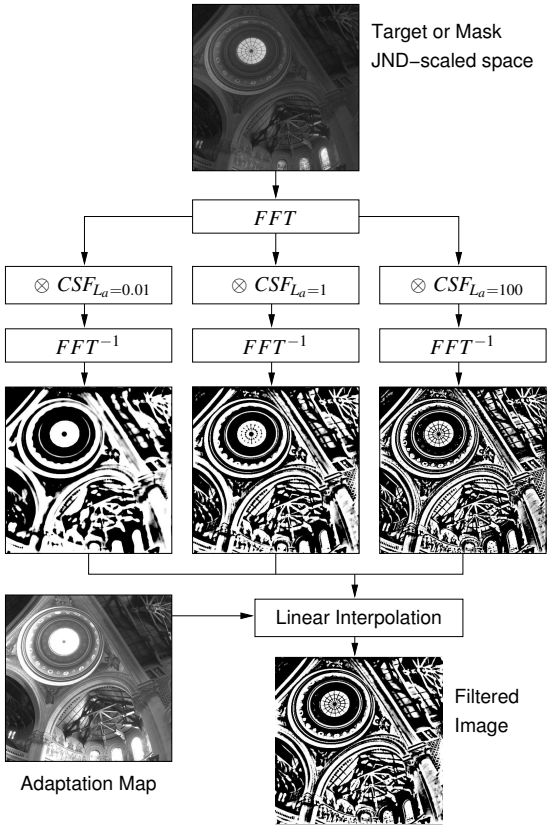


Figure 4.6: To account for a changing shape of the Contrast Sensitivity Function (CSF) with luminance of adaptation, an image is filtered using several shapes of CSF and then the filtered images are linearly interpolated. The adaptation map is used to decide which pair of filtered images should be chosen for the interpolation.

The detailed documentation of the HDR VDP software can be found on the web page. To give an impression how the software operates, the box below shows a typical usage scenario:

`vdp␣original.exr␣distorted.exr␣prediction.png`

Predict differences between an original `original.exr` and distorted `distorted.exr` images and create the visualization of the prediction in `prediction.png`.

4.3 Calibration

Both original and HDR VDP contain several adjustable parameters that can significantly improve predictions. To optimize HDR VDP predictions for complex images, its parameters have been optimized to find the best match between VDP predictions and differences found in a subjective experiment.

A psychophysical experiment that assessed the detection of differences in complex images was conducted. Then we used the collected data to find the best set of HDR VDP parameters that would give its response that is the closest to the result of the subjective tests.

Eight subjects took part in the experiment, which involved detecting visible differences in images shown on a projector based HDR display [Seetzen et al. 2004]. The luminance of the HDR images was reproduced on HDR display without any tone compression and was clamped between 0.05 and 2 700cd/m^2 (the minimum and maximum luminance that could be achieved on the display). The images were observed from 0.5 m and each image spanned about 20 visual degrees. All participants had normal or corrected to normal vision and were experienced in digital imaging.

For each pair of images (original and distorted image), a subject was to mark areas where differences between the images were visible. The marking was done using square blocks, each one of the size of one visual degree along its edge. Figure 4.7 shows the screen capture of the application used for the experiment. The result of each test was a matrix of 1 and 0 values, where value 1 denoted visible differences in a block and 0 no visible differences. Each subject was to mark eleven image pairs, which contained natural scenes (HDR photographs), computer graphics rendering, and one simple stimuli image (luminance ramp). The second image of each pair was distorted with a simple pattern noise, like a narrow band sinusoidal grating, blur, or random noise.

For the data collected from all subjects and for all images, we try to find the best set of HDR VDP parameters, that would give the VDP response, which is the closest to the subjective data. Because the resolution of VDP probability map is one pixel and the resolution of subjective response is a square block of about 30×30 pixels, we have to integrate VDP response, so that the data can be compared (see Figure 4.8). The natural choice of operator for integration is a maximum probability value (a subject marks the block if any distortion is visible). The VDP probability map however may contain single stray pixels of high probability value, which would cause the high probability of detection for the whole surrounding area. Since it is quite unlikely that a subject will notice the differences in single pixels, we choose percentile, rather than maximum, for integrating over the square block areas. Because we don't know which percentile is the best for integration, we leave it as one of the parameters of the optimization procedure.

The objective function of the optimization procedure has three parameters: a percentile used for integration k, peak contrast sensitivity P, and slope of the masking threshold elevation function s. The peak contrast sensitivity P is the minimum contrast that is visible to a human observer (the inverse of the maximum value of the CSF) and was discussed in Section 4.2.3. Refer to [Daly 1993] for the discussion on the slope of the masking function. The objective function is therefore given as:

$$f(k,P,s) = \sum_{images} \sum_{blocks} (prctile[VDP(p,s),k] - M)^2 \cdot w \tag{4.11}$$

where the first sum denotes summation over all images, the second over all rectangular blocks, $prctile$ the k'th percentile of the probability values in a block, VDP is the probability map produced by VDP, M is an averaged subjective response and w is the weighting factor for each block. The weighting factor w was introduced to account for variability of the subjective data. The average subjective response M can be any value between 0 and 1 because the subject did not mark the distorted regions in the same way. For the same reason, the importance of each block is weighted by factor

w, which denotes how much trust we can put in subjective data. If some subjects reported distortions in a particular block visible and the other subjects not visible, we can not make solid statement what should be the correct answer. Therefore we use the weighting factor:

$$w = exp(-\frac{D^2}{0.04}) \tag{4.12}$$

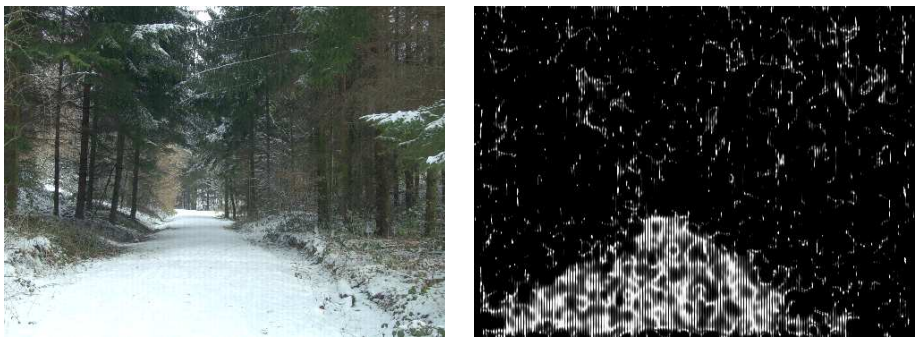
where D is a standard deviation of subjective responses across the subjects. This way the blocks that have standard deviation greater than 0.5 are practically not taken into account in the optimization procedure.

We numerically minimize the objective function f using the gradient descent method. To find a global minimum and to avoid stopping at a local minimum, we use several randomly selected starting points. Several runs of the optimization procedure gave the lowest value of the objective function for the parameters: $k = 82$, $P = 0.006$, $s = 1$. The value of 0.6% for the peak contrast sensitivity P is more conservative than 1% commonly presumed in video and image processing applications, but it also assumes lower sensitivity than the original VDP (0.25%). The slope of the masking threshold elevation function s may vary between 0.65 and 1.0 and can be explained by the learning effect [Daly 1993] (subjects are more likely to notice differences when the mask is a pattern that is predictable or they are familiar with). Although we let the slope in the optimization procedure be any value in the range of 0.5–1.5, the best fitting was found for the value 1.0, which indicated low learning level. This result was according to our expectations, since complex images form complex masking patterns, which are difficult to learn.

Figure 4.7: Screen capture of the program used in the experiment. Visible differences between two simultaneously displayed images (original on the left and distorted on the right) were marked with semi-transparent blue square blocks.

4.4 Comparison with LDR Visual Difference Predictor

To test how our modifications for HDR images affected a prediction of the visible differences, we compared the results of Daly's VDP and our modified HDR VDP.



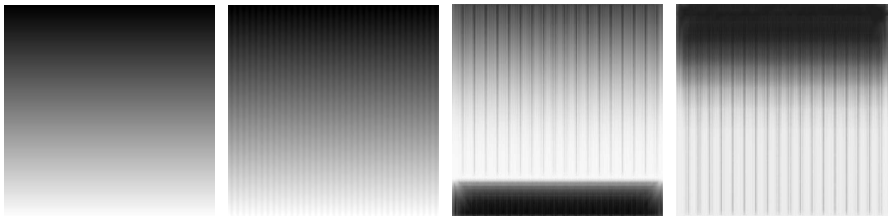
(a) Distorted image (b) VDP probability map

(c) integrated prob. map (d) averaged subjective response

Figure 4.8: Given the distorted image (a) and its not distorted version, HDR VDP produces a probability map (b). The probability map must be integrated in rectangular blocks (c) before it can be compared with the subjective response (d).

The first pair of images contained a luminance ramp and the same ramp distorted by a sinusoidal grating (see Figure 4.9). The probability map of Daly's VDP (Figure 4.9(c)) shows lack of visible differences for high luminance area (bottom of the image). This is due to the luminance compression of the photoreceptor model (compare with Figure 4.3). HDR VDP does not predict loss of visibility for high luminance (Figure 4.9(d)), but it does for lower luminance levels, which is in agreement with the *contrast versus intensity* characteristic of the HVS. The visibility threshold for average and low luminance is also lowered by the CSF, which suppresses the grating of 5 $cycles/degree$ for luminance lower than 1 cd/m^2 (see Figure 4.5). Because Daly's VDP filters images using the CSF for a single adaptation level, there is no difference in the grating suppression for both low and high luminance regions of the image.

The next set of experiments was performed on a set of HDR images that are commonly used for testing tone mapping operators. The first row of Figure 4.10 shows a prediction of contouring artifacts in the *Memorial Church* image. Both VDPs predicted properly visibility of the artifacts in the non-masked areas (floor and columns). However, Daly's VDP failed to predict distortions in the bright highlight on the floor



(a) Mask (b) Target (c) Daly's VDP (d) HDR VDP

Figure 4.9: A logarithmic luminance ramp (a) from 10^{-4} cd/m^2 (top of the image) to 10^6 cd/m^2 (bottom of the image) was distorted with a sinusoidal grating of contrast 10% and frequency 5 $cycles/degree$ (b). The original and the distorted image was compared using both versions of the VDP and the resulting probability map was shown in subfigures (c) and (d), where brighter gray-levels denote higher probability.

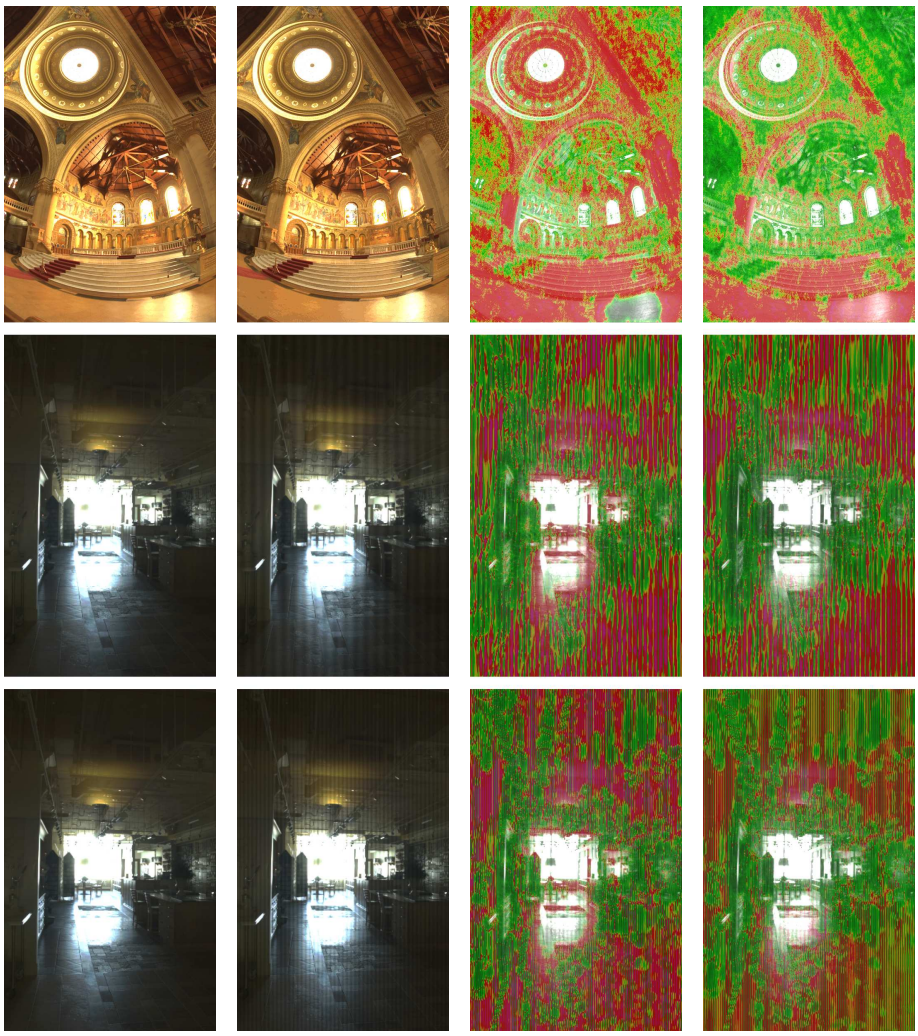
(bottom right of the image), which can be caused by excessive luminance compression at high luminance levels. Daly's metric also overestimated visible differences in dark regions of the scene. Similar results were obtained for the *Design Center* image distorted by a sinusoidal grating of different frequencies (the second and third row of Figure 4.10). High frequency noise (the third row) was suppressed for the low luminance region of the image (the right bottom corner) only in case of the HDR VDP. Such noise is mostly visible in the brighter parts of the image (the ceiling lamp and the areas near the window), for which the CSF predicts higher sensitivity at high frequencies.

This short validation confirmed a better prediction of HDR VDP at high luminance levels (in accordance with the *cvi*) and at low luminance levels in the presence of high frequency patterns (in accordance with the CSF). However, more tests should be performed in the future to test the prediction of contrast masking.

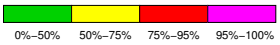
4.5 Conclusions and Future Work

In this chapter we derive several extensions to the original Visual Difference Predictor. The extensions enable the comparison of High-Dynamic Range images. Local contrast reduction is modeled in the extended HDR VDP using three-tier processing: linear shift invariant OTF for light scattering, nonlinear shift invariant conversion to JND-scaled space for the response of the photoreceptor, and the last linear and shift variant CSF for lower sensitivity to low and high spatial frequencies. Such model allows separate processing of high and low contrast information in HDR images. The predictor is then calibrated to the psychophysical data collected in the detection experiment on the HDR display.

In future work we would like to further extend the VDP to handle color images in a similar way as it was done in [Jin et al. 1998], but also take into consideration extended color gamut and the influence of chromatic aberration on the OTF [Marimont and Wandell 1994]. A more extensive validation of HDR VDP predictions is necessary to confirm a good correlation between the predicted distortions and the actual quality



(a) Mask (b) Target (c) Daly's VDP (d) HDR VDP



(e) Color-coded scale of detection probability for VDP output

Figure 4.10: Several test images (a) were distorted by quantization in log domain (first row), 5 $cycles/degree$ 10% contrast sinusoidal noise (second row), and 2 $cycles/degree$ 10% contrast sinusoidal noise (third row). The last two columns show results of both Daly's VDP (c) and HDR VDP (d) using color-coded probability scale (e).

degradation as perceived by a human observer.

Chapter 5

Compression of HDR Images and Video

The bit-depth precision of majority of image and video formats can soon become insufficient for the new generation of displays. The traditional image and video formats, such as JPEG, PNG or MPEG, employ color spaces that fail to represent scenes of dynamic range over 2 or 3 orders of magnitude and extended color gamut. The 8 bits per color channel were more than sufficient when these formats were designed and the best CRT displays could achieve contrast ratio of 1:200 and they peak luminance did not exceed 100 cd/m^2. Now, commercially available displays can show contrast of 1:1000. The prototypes of HDR displays are capable of showing contrast 1:50 000 and have the peak luminance of 3000 cd/m^2 [Seetzen et al. 2004]. Moreover, the color gamut of typical displays also becomes much larger. These new advances in display technology makes the transition to new image and video encoding formats, capable of supporting new displays, essential.

One of the weakest points of the existing image and video file formats is that they are device dependent. The gamma correction non-linearity, still used in most color spaces used for compression, was originally designed for the first CRT TV sets [Poynton 2003]. When technology changes rapidly, building standards based on the characteristics of the particular devices does not seem to be appropriate. This chapter describes image and video encoding that is device independent and is solely based on the capabilities of the human visual system. The basic concepts of such device independent encoding are introduced in Section 5.1.

Higher precision of visual data does not only mean better reproduction of images and video, but also new possibilities of reproduction. The display that is provided with high accuracy device independent images, can render them using an optimal tone and gamut mapping algorithm, and even adjust for the viewing conditions. HDR information is already exploited in video games to accurately simulate a range of perceptual effects, such as visual glare, night vision and motion blur, which enhance realism of the displayed images. Given HDR video input, the perceptual effects, as shown in Figure 5.1, could be rendered in real-time by the display [Krawczyk et al. 2005b].

This chapter presents several of algorithms for compression of HDR images and video,

Figure 5.1: A range of perceptual effects that can be simulated based on HDR data. From left to right: visual glare (see light scattering at the edges of the objects); motion blur can be correctly simulated in linear luminance domain; given absolute luminance values, color deficiency of night (scotopic) vision can be simulated.

that can represent all information that is visible to the human eye. In particular, a color space for efficient encoding of HDR pixels is derived in Section 5.3. Section 5.4 describes extensions required to encode HDR video using MPEG-4 compression. As HDR formats have just started gaining popularity, it is important to provide a backward compatibility with the existing LDR formats. The schemes for backward compatible compression of HDR images and video are described in Sections 5.5 and 5.6.

This chapter consolidates previous work on HDR image and video compression published in [Mantiuk et al. 2004a], [Mantiuk et al. 2006c] and [Mantiuk et al. 2006a].

5.1 Device- and Scene-referred Representation

Capturing of HDR video and images has become easier with the development of HDR cameras. On the other end of the pipeline, display of HDR data has become possible with the availability of new generation of HDR displays. However, in order to make those two ends of the pipeline work together, there is a need for a common format of data. This can be achieved with so called *scene-referred* representation of images and video.

Commonly used image formats (JPEG, PNG, TIFF, etc.) contain data that is tailored to particular display devices: cameras, CRT or LCD monitors. For example, two JPEG images shown using two different LCD monitors may be significantly different due to dissimilar image processing, color filters, gamma correction etc. Obviously, such representation of images vaguely relates to the actual photometric properties of the scene it depicts, but it is dependent on a display device. Therefore those formats can be considered as *device-referred* (also known as *output-referred*), since they are tightly coupled with the capabilities and characteristic of a particular imaging device.

ICC color profiles can be used to convert visual data from one device-referred format to another. Such profiles define the colorimetric properties of a device for which the image is intended for. Problems arise if the two devices have different color gamuts or dynamic ranges, in which case a conversion from one format to another usually involves the loss of some visual information. The algorithms for the best reproduction of LDR images on the output media of different color gamut have been thoroughly

studied [Morovic and Luo 2001] and CIE technical committee (CIE Division 8: TC8-03) have been started to choose the best algorithm. However, as for now, the committee has not been able to select a single algorithm that would give reliable results in all cases. The problem is even more difficult when an image captured with an HDR camera is converted to the color space of a low-dynamic range monitor (see a multitude of tone mapping algorithms [Reinhard et al. 2005, Chapter 7]). Obviously, the ICC profiles cannot be easily used to facilitate interchange of data between LDR and HDR devices.

Scene-referred representation of images offers a much simpler solution to this problem. The scene-referred image encodes the actual photometric characteristic of a scene it depicts [Reinhard et al. 2005, p.85]. Conversion from such common representation, which directly corresponds to physical luminance or spectral radiance values, to a format suitable for a particular device is the responsibility of that device. HDR file formats are examples of scene-referred encoding, as they usually represent either luminance or spectral radiance, rather than gamma corrected "pixel values".

5.2 HDR Image Formats

There are several existing formats that are capable of encoding higher dynamic range images. They can be classified into three groups:

- Formats originally designed for high dynamic range images. The quantities they store are usually floating points values of a linear radiance or luminance factor[1]. There are several high-precission formats, such as Radiance's RGBE, logLuv TIFF and OpenEXR. These formats are lossless up to the precission of their pixel representation. The backward compatible JPEG HDR format can be also classified to this group, though it is a lossy format. The high-precission formats are described in detail in the following sections and the JPEG HDR format is described in Section 5.5.2.
- Formats designed to store a higher dynamic range because of their application. This group includes: Digital Picture Exchange *DPX* format used in the movie industry to store scanned negatives, *DICOM* format for medical images, and a variety of so called *RAW* formats used in digital cameras. All these formats use more than 8 bits to store luminance, but they are not capable of storing such an extended dynamic range as the HDR formats.
- Formats that store larger number of bits but are not necessary intended for HDR images. Twelve or more bits can be stored in JPEG-2000, MPEG-4 (ISO/IEC 14496-2 or ISO/IEC 14496-10) and TIFF files. All these formats can easily encode HDR if they take advantage of a color space that can represent full visible range of luminance and color gamut.

Variety of formats and lack of standards definitely hinders transition from traditional output-referred LDR formats to scene-referred HDR formats. The HDR formats (Radiance's RGBE, logLuv TIFF and OpenEXR) have not gained widespread acceptance mainly because they offer only lossless compression resulting in huge files sizes. The most successful OpenEXR format has been however integrated with several Open Source and commercial applications, such as Adobe® Photoshop® CS2. The JPEG HDR

[1] For the explanation of luminance factor, refer to Section 2.1

format can gain large popularity if it is adapted by a number of image processing applications. Another reason for small popularity of HDR formats is lack of standards, which comes from little interest from the image and video format community in encoding HDR images. Other specialized formats, such as DPX, DICOM and cameras' RAW formats, usually do not allow storing as high dynamic range as the HDR formats. Since they are designed to be used for a specific application, it is unlikely that they will evolve into general purpose image formats.

The recent video compression standards offer an extended bit-depth of up to 12 bits for ISO/IEC 14496-2 and ISO/IEC 14496-10 AVC/H.264 with high profiles defined in the Fidelity Range Extensions (FREXt), and 16 bits for JPEG-2000 format. This unfortunately does not imply that these extensions were designed to store higher dynamic range. Despite the higher bit-depth, the specified transfer functions allow encoding only up to 2.5 log-10 units of dynamic range. The obvious step would be to extend the specifications of these format to allow encoding HDR images and video. The following sections of this chapter propose several such extensions, including color space for HDR pixels (Section 5.3), an efficient encoding of sharp contrast edges (Section 5.4.2), and finally backward compatible encoding of video (MPEG) and images (Section 5.6). However, before these extensions are introduced, the following subsections give and overview of the most popular scene-referred HDR image formats. More detailed review of the existing HDR image formats can be found in [Reinhard et al. 2005, Section 3].

5.2.1 Radiance's HDR Format

One of the first HDR image formats, which gained much popularity, was introduced with the Radiance rendering package[2]. Therefore it is known as the *Radiance* picture format [Ward 1991] and can be recognized by the file extensions .hdr or .pic. The file consist of a short text header, followed by run-length encoded pixels. The pixels are encoded using so called *RGBE* or *XYZE* representations, which differ only by a color space that is used. *XYZE* color format can encode full visible color gamut, while *RGBE* is limited to the chromacities that lie within the triangle formed by the red, green and blue color primaries. Since both representations are very similar, we only describe the *RGBE* encoding.

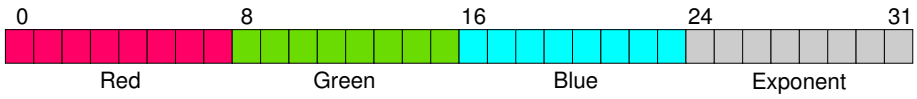


Figure 5.2: 32-bit per pixel RGBE encoding

RGBE pixel encoding represents colors using four bytes: the first three bytes encode red, green and blue color channels, and the last byte is a common exponent for all channels (see Figure 5.2). RGBE is essentially a custom floating point representation of pixel values. RGBE encoding takes advantage of the fact that all color channels are strongly correlated in RGB color space and their values are at least of the same order of magnitude. Therefore there is no need to store a separate exponent for each color channel.

[2]Radiance is an open source light simulation and realistic rendering package. Home page: http://radsite.lbl.gov/radiance/

5.2.2 logLuv TIFF

The major drawback of floating point representation of pixel values is that floating point numbers do not compress well. This is mainly because additional bits are required to encode mantissa and exponent separately, instead of a single integer value. Such representation, although flexible, is not really required for visual data. Furthermore, precision error of floating point numbers varies across the full range of possible values and is different than the "precision" of our visual system, as illustrated in Figure 5.9. Therefore, better compression can be achieved when integer numbers are used to encode HDR pixels.

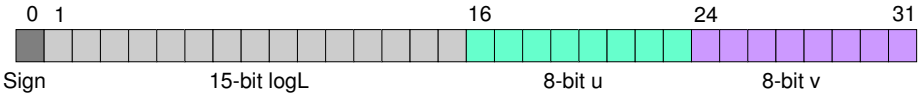


Figure 5.3: 32-bit per pixel LogLuv encoding

The *LogLuv* encoding [Ward Larson 1998] requires only integer numbers to encode full range of luminance and color gamut that is visible to the human eye. It is an optional encoding in the TIFF library. This encoding benefits from the fact that the human eye is not equally sensitive to all luminance ranges. In the dark we can see a luminance difference of a few hundredths of cd/m^2, while in the sunlight we need a difference of tens of cd/m^2 to see a difference. This effect is often called luminance masking and is discussed in Section 3.3. But if, instead of luminance, a logarithm of luminance is considered, the detectable threshold values do not vary so much and a constant value can be a conservative approximation of the visible threshold. Therefore if a logarithm of luminance is encoded using integer numbers, quantization errors roughly correspond to the visibility thresholds of the human visual system, which is a desirable property for pixel encoding. 32-bit LogLuv encoding uses two bytes to encode luminance and another two bytes to represent chrominance (see Figure 5.3). Chrominance is encoded using a perceptually uniform chromacity scale u' v' (see Section 5.3.2 for details). There is also 24-bit LogLuv encoding, which needs fewer bits to encode pixels with the precision that is below the visibility thresholds. However, this format is rather ineffective to encode, due to discontinuities resulting from encoding two chrominance channels with a single lookup value.

5.2.3 OpenEXR

An OpenEXR format or (the EXtended Range format), recognized by the file name extension .exr, was made available with an open source C++ library in 2002 by Industrial Light and Magic (see `http://www.openexr.org/` and [Bogart et al. 2003]). Before that date the format was used internally by Industrial Light and Magic for the purpose of a special effect production. The format is currently promoted as a special-effect industry standard and many software packages already support it. Some features of this format include:

- Support for 16-bit floating-point, 32-bit floating-point, and 32-bit integer pixels. The 16-bit floating-point format, called "half", is compatible with the *HALF* data type in NVIDIA's Cg graphics language and is supported natively on their new GeForce FX and Quadro FX 3D graphics solutions.

- Multiple lossless image compression algorithms. Some of the included codecs can achieve 2:1 lossless compression ratios on images with film grain.
- Extensibility. New compression codecs and image types can easily be added by extending the C++ classes included in the OpenEXR software distribution. New image attributes (strings, vectors, integers, etc.) can be added to OpenEXR image headers without affecting backward compatibility with existing OpenEXR applications.

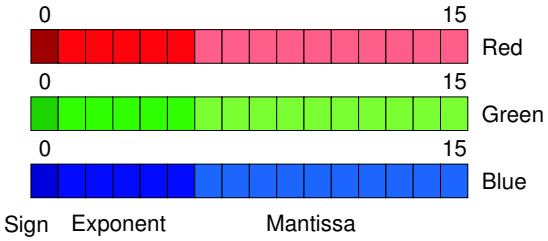


Figure 5.4: 48-bit per pixel OpenEXR half-precision floating point encoding

Although OpenEXR file format offers several data types to encode channels, color data is usually encoded with 16-bit floating point numbers, known as half-precision floating point. Such two byte floating point number consist of one bit of sign, 5-bit exponent, and 10-bit mantissa, as shown in Figure 5.4 (thus the format is known also as S5E10).

5.2.4 Formats Used in Cinematography

While the dynamic range employed in digital photography is usually limited to 2–3 orders of magnitude, a much broader dynamic range of 4–5 orders of magnitude can be achieved with analog film. The problem of digital encoding, which emulates the dynamic range and *S*-shaped response curve of film has been addressed in patent literature [Lucian et al. 2005]. A hardware solution using two 14-bit analog-to-digital converters to separately digitize the log-linear and shoulder/toe portions of the response curve is proposed to recover more details in dark and bright scene regions. Custom wavelet encoders, such as layered wavelet encoders, have been designed especially for the purpose of storing wide dynamic range scans of film negatives used in cinematography [Demos 2004]. Such compression method however require substantial bit-rates and are not suitable for on-DVD storage or real-time playback. The dynamic range level achieved with analog film and its digital emulation is also too low to meet HDR standards. Besides, it can be argued that the video encoding format should be designed for the capabilities of the human eye rather than analog film or camera characteristics.

5.3 Color Space for HDR Pixels

The recent advances in digital camera and display technologies make standard 8-bit per color channel representation of visual data insufficient. This is mostly due to the extended dynamic range of new capture and display devices: high dynamic range cameras can capture dynamic range over 150dB (compared to 65dB for a typical camera) and

new HDR displays can show contrast ratio of 30 000:1 (compared to 400:1 for a typical LCD). Furthermore, these devices can cover much wider range of absolute luminance levels, ranging from 0.1 cd/m^2 to 3 000 cd/m^2 for a HDR display. Since the typical color spaces, such as YC_rC_b, *sRGB* or CIE $L^*u^*v^*$ cannot encode the full luminance range of HDR data, a new representation of the visual data that can accommodate the extended dynamic range is needed.

High dynamic range (HDR) imaging is a very attractive way of capturing real world appearance, since it assumes the preservation of complete and accurate luminance (or spectral radiance) values that can be found in a scene. Each pixel is represented as a triple of floating point values, which can range from 10^{-5} to 10^{10}. Such a huge range of values is dictated by both real world luminance levels and the capabilities of the human visual system (HVS), which can adapt to a broad range of luminance levels, ranging from scotopic (10^{-5} – 10 cd/m^2) to photopic (10 – 10^6 cd/m^2) conditions. Obviously, floating point representation results in huge memory and storage requirements and is impractical for storage and transmission of images and video. In this section we derive a color space that can efficiently encode HDR pixel values.

Choice of the color space used for image or video compression has a great impact on the compression performance and capabilities of the encoding format. To offer the best trade-off between compression efficiency and visual quality without imposing any assumptions on the display technology, we propose that the color space used for compression has the following properties:

1. The color space can encode the full color gamut and the full range of luminance that is visible to the human eye. This way the human eye, instead of the current imaging technology, defines the limits of such encoding.

2. A unit distance in the color space correlates with the Just Noticeable Difference (JND). This offers a more uniform distribution of distortions across an image and simplifies control over distortions for lossy compression algorithms.

3. Only positive integer values are used to encode luminance and color. Integer representation simplifies and improves image and video compression.

4. A half-unit distance in the color space is below 1 JND. If this condition is met, the quantization errors due to rounding to integer numbers are not visible.

5. The correlation between color channels should be minimal. If color channels are correlated, the same information is encoded twice, which worsens the compression performance.

6. There is a direct relation between the encoded integer values and the photometrically calibrated XYZ color values.

There are several color spaces that already meet some of the above requirements, but there is no color space that accommodates them all. For example, the Euclidean distance in $L^*u^*v^*$ color space correlates with the JND (Property 2), but this color space does not generalize to the full range of visible luminance levels, ranging from scotopic light levels, to very bright photopic conditions. Several perceptually uniform quantization strategies have been proposed [Sezan et al. 1987, Lubin and Pica 1991], including the grayscale standard display function from the DICOM standard [DICOM PS 3-2004 2004]. However, none of these take into account broad dynamic range and diversified luminance conditions as required by Property 1. Property 3 would suggest that the

derivation of such color space can be formulated as color quantization problem [Brun and Tremeau 2003]. However, color quantization techniques focus mostly on clustering problem in the three-dimensional color space assuming that a perceptually uniform color space, such as CIE $L^*u^*v^*$ or CIE $L^*a^*b^*$, is already given.

5.3.1 Luminance and Luma

We begin the derivation of the color space that incorporates all of the above listed properties with the luminance channel. Real-world physical luminance, given in cd/m^2, should be converted into integer numbers (Properties 3 and 6), so that the error due to rounding to the nearest integer is not visible (Property 4). Additionally, it is desirable that the integer values representing luminance closely correspond to the sensory response of the HVS (Property 2). For example, intensity of sound is usually measured using non-linear decibel (dB) units since such a measure well corresponds to the perceived loudness of sound. We would like to find a similar measure of luminance for all possible light conditions. Our derivation is similar to other methods that model sensory output for a physical signal based on its threshold characteristic, such as transducer functions [Wilson 1980, Barten 1999], the grayscale standard display function [DICOM PS 3-2004 2004], or the capacity function in tone mapping [Ashikhmin 2002]. Such luminance conversion is also called a *transfer function* in image compression literature.

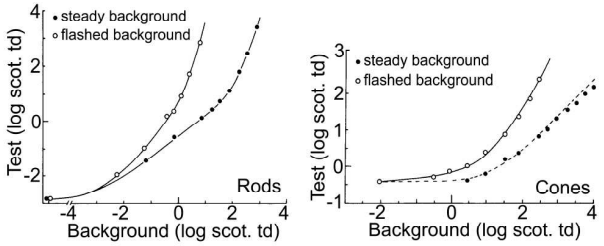


Figure 5.5: Threshold versus intensity characteristic for fully adapted (filled) and maladapted eye (open), for rods (left) and cones (right). The threshold detection performance drops when the eye is not fully adapted. From [Walraven et al. 1990].

Let us assume that the function $t(y_{adapt})$ gives a conservative estimate of the smallest difference of luminance that is visible to the human eye (the detection threshold) at a particular adaptation level, y_{adapt}. We are looking for a function $l \rightarrow y : y(l)$ that converts sensory units l (e.g. response of the photoreceptor), which we will call *luma* (refer to Section 2.1), into physical luminance y. Because the luma values l will be encoded as integer numbers (Property 3), we have to make sure that rounding to integers does not introduce visible distortions (Property 4). The maximum quantization error due to rounding of luma values, l, is ± 0.5. Since the detection thresholds are given in luminance, we have to convert this rounding error from luma, l, to luminance, y. This can be done by the Taylor series expansion of the function $y(l)$:

$$y(l+0.5) - y(l) \approx 0.5 \cdot \frac{dy}{dl} \tag{5.1}$$

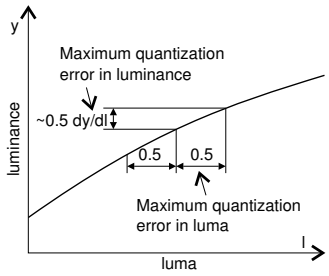


Figure 5.6: Maximum quantization error in sensory values, l, must be expressed in luminance, y, before it can be compared with the detection threshold, $t(y_{adapt})$.

This step is illustrated in Figure 5.6. We then make sure that the maximum rounding error is below or equal the detection threshold, $t(y_{adapt})$:

$$0.5 \cdot \frac{dy}{dl} < t(y_{adapt}) \tag{5.2}$$

To simplify our problem, we assume that the eye is adapted to the luminance of a single pixel, $y_{adapt} = y$. Although such an assumption is not true in real-world situations, it gives a conservative estimate of the detection threshold: the detection threshold is higher when the eye is not fully adapted [Walraven et al. 1990, Irawan et al. 2005]. This is illustrated in the Figure 5.5, which shows how thresholds increase when the eye is not fully adapted to background illuminance. We can rewrite the above inequality as the following equality:

$$\frac{dy}{dl} = 2 \cdot \frac{t(y)}{k} \tag{5.3}$$

where k is a constant greater than 1. The larger the value of k, the more conservative the encoding (the lower is the quantization error for luminance), but also the more bits are needed to encode l. An important consequence of rewriting Inequality 5.2 as Equality 5.3 is that the differencial change in y per differential change in l now directly relates to the sensory threshold $t(y)$, therefore the equation meets Property 2. The above equation can be solved in either of two ways:

- by solving a differential equation:

$$\frac{dy}{dl} = 2 \cdot \frac{t(y(l))}{k} \tag{5.4}$$

- or an integral:

$$\frac{dl}{dy} = 0.5 \cdot \frac{k}{t(y)} \;\Rightarrow\; l(y) = 0.5 \int \frac{k}{t(y)} dy \tag{5.5}$$

The solution of Equation 5.5 gives a function $y \rightarrow l : l(y)$, which converts physical luminance y into sensory units l, and the solution of Equation 5.4 gives the inverse function $l \rightarrow y : y(l)$. Note that y from Equation 5.3 has been replaced in Equation 5.4 with $y(l)$ to make the right side of the equation the function of l.

Finally, we must decide on the boundary conditions and find the value of the constant k. The boundary conditions will define the range of physical luminance that should be represented by the sensory units l. A reasonable range of luminance is within

$10^{-5}\ cd/m^2$ and $10^{10}\ cd/m^2$, which can capture the luminance of both a moonless sky ($3 \cdot 10^{-5}\ cd/m^2$) and the surface of the sun ($2 \cdot 10^{9}\ cd/m^2$). Therefore we can write the boundary conditions:

$$\begin{aligned} y(0) &= 10^{-5}\ cd/m^2 \\ y(l_{max}) &= 10^{10}\ cd/m^2 \end{aligned} \tag{5.6}$$

where l_{max} is the maximum value of l we want to encode and is usually equal $l_{max} = 2^{bits} - 1$. This gives us two point boundary problem, which can be solved using the *shooting method* [Press et al. 2002, Chapter 17][3]. The solution will give us the value of k. If the value of k is greater than 1, the sensory units, l, can represent luminance with sufficient precision, and that we have chosen an adequate number of bits.

So far we have not made any assumptions about the actual shape of the contrast detection threshold function $t(y_{adapt})$. We can start with a simplistic case, where this function equals 1% of the Weber fraction[4], that is $t(y_{adapt}) = 0.01\ y_{adapt}$. This is a very imprecise, but unfortunately still commonly used assumption in computer vision and image compression, which is also referred as the Weber-Fechner law[5]. We make further simplification and consider the case where $k = 1$, where the maximum quantization error is exactly equal $t(y_{adapt})$, rather than being greater than the threshold. From Equation 5.5 and our assumptions, we get:

$$l(y) = 0.5 \int \frac{1}{0.01y} dy = 50 \cdot \ln|y| + c \tag{5.7}$$

From the lower boundary condition (Equation 5.6), we have $c = -50 \cdot \ln|10^{-5}| = 575.65$. This way we derive a logarithmic compression function, which is commonly used for processing HDR images. Additionally, the derived function has the useful property that the unit difference corresponds to 1% contrast. We insert the upper boundary condition into Equation 5.7, we get $l(10^{10}) = 1726.9$, which means that we need at least 11 bits to represent the full visible range of luminance with a 1% step. Although such precision is usually regarded sufficient for video displayed on CRT displays, skilled observers are reported to notice contrast as low as 0.25%. Moreover, the contrast detection threshold is decreased with increased luminance of adaptation. Since new LCD and plasma displays are much brighter than their CRT counterparts, they eye is adapted to higher luminance levels when viewing such displays. Therefore, it is not certain whether 1% contrast is still a conservative assumption. To accurately predict visibility of distortions under a broad range of viewing conditions, more accurate models of detection threshold should be employed.

The detection threshold of the HVS is usually modelled in psychophysics with either a threshold versus intensity function (t.v.i.) or a more complex Contrast Sensitivity Function (CSF). The difference between them is that the t.v.i. function is measured for a fixed pattern, such as a circular patch on a uniform background, and the CSF is measured for a sinusoidal patterns or Gabor patches of different spatial frequencies. In our analysis we consider the most popular models of t.v.i. and CSF, which include:

[3] Briefly, a shooting method is an iterative procedure that performs a binary search for the k value until the differential equation meets the boundary conditions.

[4] Weber fraction is usually defined as $W = (y_{max} - y_{min})/y_{min}$.

[5] It was shown over 40 years ago that the Weber-Fechner law does not match the experimental data for luminance [Stevens and Stevens 1960]. The discrepancy between the Weber-Fechner law and the real measurements is even higher for high dynamic range images.

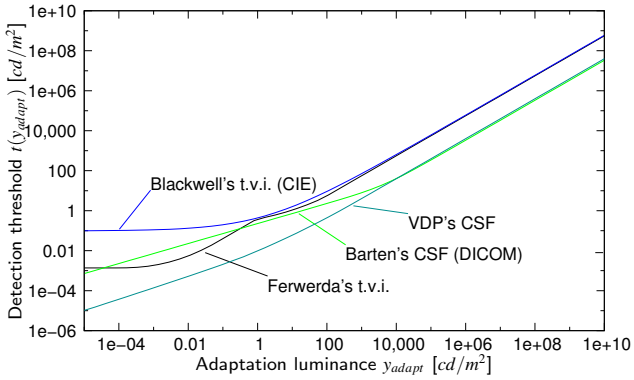


Figure 5.7: Comparison of the detection threshold models based on different CSF and t.v.i. functions.

- Ferwerda's t.v.i. [Ferwerda et al. 1996], which is commonly used in computer graphics;
- The t.v.i. model suggested by Bodmann [1973] based on Blackwell's data [Blackwell and Blackwell 1971] for 20–30 year old observers and adopted by the CIE standard [CIE 1981];
- Barten's CSF model [Barten 1999] adopted by the DICOM standard [DICOM PS 3-2004 2004];
- and Meeteren's CSF model [Van Meeteren and Vos 1972], improved by Kodak and used in the Visual Difference Predictor (VDP) [Daly 1993].

While t.v.i. functions can be used directly to replace the function $t(y_{adapt})$, some assumptions must be made before the thresholds can be found from a CSF. Sensitivity modelled by a CSF can depend on the stimuli size, viewing conditions, spatial and temporal frequency, eccentricity and orientation. To make a conservative choice, we assume the worst case scenario and always choose the point on the CSF where sensitivity is the highest. Since sensitivity is modelled as an inverse of Weber's fraction, we get:

$$t(y_{adapt}) = \frac{y_{adapt}}{\max_{\rho} CSF(\rho, y_{adapt})} \tag{5.8}$$

assuming a simplified CSF, which is a function of spatial frequency ρ and luminance of adaptation y_{adapt}. This is the same approach that we used to derive the t.v.i. function from the contrast sensitivity function in Section 4.2.2, Equation 4.7.

For comparison, the $t(y_{adapt})$ functions based on the above listed t.v.i. and CSF models are plotted in Figure 5.7. Note that all functions follow a similar shape, but they are also shifted along t-axis between each other. This comes from the difference in measuring methods and also from the differences in the peak sensitivity between individuals. In general, the CSF models show lower thresholds than the t.v.i. models.

Using each of the four detection threshold models, we found the coefficient k by solving the two point boundary problem, as described above, for the visible range of luminance and for 12-bit encoding. The resulting curves ($l(y)$ functions) are plotted in Figure 5.8. The constant k was above 1 for all functions (the smallest $k = 1.4481$ was found for

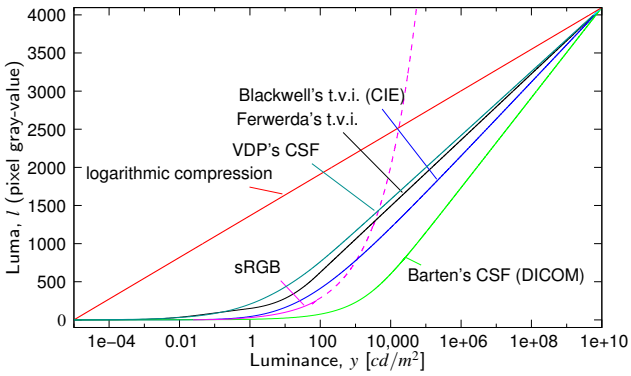


Figure 5.8: Luminance to luma mappings, derived from different threshold models. A logarithmic function and the sRGB color space are included for comparison.

VDP's CSF), therefore rounding the values of those functions to integer numbers does not introduce errors above the detection threshold. All four curves based on the t.v.i. or CSF data have slightly different shapes, resulting in different sensitivity for different luminance ranges. Additionally, Figure 5.8 contains two additional curves depicting the nonlinearity (gamma correction) used in the sRGB standard [IEC 61966-2-1:1999 1999] and a logarithmic compression. The sRGB nonlinearity is plotted as a continuous line up to 80 cd/m^2, which is the display white luminance level assumed by the standard. The dashed line illustrates how the sRGB nonlinearity accelerates for high luminance levels, making it practically unsuitable for HDR data. The sRGB color space has not been designed to encode luminance levels above a few hundreds cd/m^2. Also, the logarithmic compression curve has been fit into 12-bit luma range. This curve is significantly different than the other functions, which model the human perception more accurately. The 12-bit logarithmic encoding resulted in a relative quantization error about 0.42%.

At this point, we removed both the curve derived from the Ferwerda's t.v.i. and the curve based on the Barten's CSF from further consideration. The Ferwerda's t.v.i. is based on data from very few subjects and is measured for cone and rod vision separately, therefore it is less plausible than the other curves. The curve derived from Barten's CSF results in too coarse quantization for luminance below 1 000 cd/m^2 and too conservative quantization for luminance above that point (a steeper curve means that a luminance range is projected on a larger number of discrete sensory values, l, thus lowering quantization errors). Since we would like the quantization error to be at least as conservative as the quantization of the sRGB color space, this curve is not suitable for our application. The remaining two curves are equally suitable for encoding HDR and the choice between them may depend on the application. VDP's CSF is more conservative for low luminance. The curve derived from the CIE data is close to the gamma correction used in the sRGB color space, which gives better compatibility with low-dynamic range images, for which sRGB is de facto a standard.

We use a numerical method to derive the functions shown in Figure 5.8. However, for many applications, it is desirable to have an analytical formula, which could facilitate conversion between HDR luminance and 12-bit luma. We propose an analytical model that is both simple and resembles similar formulas used for the same purpose but for

low dynamic range. We define a conversion from luminance to luma as:

$$l(y) = \begin{cases} a \cdot y & \text{if } y < y_l \\ b \cdot y^c + d & \text{if } y_l \leq y < y_h \\ e \cdot \log(y) + f & \text{if } y \geq y_h \end{cases} \tag{5.9}$$

The above model is similar to the sRGB non-linearity, which also consists of linear and power function segments. The difference is that the above model additionally includes a logarithmic segment for high luminance.

To fit the model to the numerical solution of $l(y)$ for both the CIE and VDP's detection models, we use the Levenberg-Marquardt nonlinear regression. Additionally, we enforce C^1 continuity in y_l and y_h in order to achieve a smooth function. We get the best fit to the data for the constants listed in the table below:

Model	a	b	c	d	e	f	y_l	y_h
CIE t.v.i.	17.554	826.81	0.10013	-884.17	209.16	-731.28	5.6046	10469
VDP's CSF	769.18	449.12	0.16999	-232.25	181.7	-90.160	0.061843	164.10

An inverse mapping, from luma to luminance, can be found using the formula:

$$y(l) = \begin{cases} a' \cdot l & \text{if } l < l_l \\ b'(l + d')^{c'} & \text{if } l_l \leq l < l_h \\ e' \cdot \exp(f' \cdot l) & \text{if } l \geq l_h \end{cases} \tag{5.10}$$

where the coefficients are given in the table below:

Model	a'	b'	c'	d'	e'	f'	l_l	l_h
CIE t.v.i.	0.056968	7.3014e-30	9.9872	884.17	32.994	0.0047811	98.381	1204.7
VDP's CSF	0.0013001	2.4969e-16	5.8825	232.25	1.6425	0.0055036	47.568	836.59

It is important to note that the model from Equation 5.9 is only an approximation of the accurate mapping function, derived by a numerical or analytical solution of Equations 5.4 or 5.5. The applications that require high accuracy of the predicted quantization errors should use the accurate solution rather than the approximate model. Although it is possible to design a more accurate model, it would be too complex to be practical. It is also important to note that a pre-computed lookup table for luma to luminance mapping can often give much better performance than an analytical formula that involves computationally expensive power and logarithmic functions. However, as we recognize that the lack of simple formulas often discourage the application of a method, we propose this simplified model as a better alternative to the logarithmic compression,

The problem of perception-based image data quantization that minimizes contouring artifacts has been extensively studied in the literature [Sezan et al. 1987, Lubin and Pica 1991] but mostly for LDR imaging. A simpler mapping function for HDR images than the one derived above is used in the LogLuv format [Ward Larson 1998]. LogLuv uses a logarithmic function to map from luminance values to 15-bit integers. The quantization error of such mapping against a range of visible luminance is shown in Figure 5.9. LogLuv mapping function is well aligned to the c.v.i. curve at high luminance values. However, the logarithmic mapping is too conservative for scotopic and mesopic conditions. As a result, a significant amount of bits is wasted to encode small contrast changes at low luminance, which are not visible to the human observer. We propose a more effective mapping from luminance to discrete values, which is in a better agreement with human perception.

5.3.2 Chrominance and Chroma

Having derived the luminance component of the color space for HDR, we now focus on encoding chrominance as two 8-bit chroma channels. Using eight bits per channel to encode color is motivated by existing image formats, which often offer twelve or more bits for luminance channel, but rarely encode chrominance with higher precision than eight bits per channel.

Although an obvious choice for image and video compression would be a variant of YC_rC_b color space, we rejected it because of its limited color gamut. HDR frames should preserve the full visible color gamut (recall Property 1 from Section 5.3), even though it cannot be displayed on the existing displays. We have experimented with several color spaces, including a variant of RGB with an extended gamut (more saturated primaries), but finally we achieved the best results with the CIE 1976 Uniform Chromacity Scales u', v' (refer to Section 2.2). Similarly as in [Ward Larson 1998], we compute the values for chrominance channels using the equations:

$$\begin{aligned} u' &= \frac{4X}{X+15Y+3Z} \\ v' &= \frac{9Y}{X+15Y+3Z} \end{aligned} \tag{5.11}$$

Then we encode u' and v' using 8-bits:

$$\begin{aligned} u_{8bit} &= u' \cdot 410 \\ v_{8bit} &= v' \cdot 410 \end{aligned} \tag{5.12}$$

Note that we use u' and v' chromatices rather than u^* and v^* of the $L^*u^*v^*$ color space. Although u^* and v^* give better perceptual uniformity and predict loss of color sensitivity at low light (Property 2), they are strongly correlated with luminance. Such correlation is undesired in image or video compression (Property 5). Besides, u^* and v^* could reach high values for high luminance, which would be difficult to encode using only eight bits.

The remaining question is whether u_{8bit} and v_{8bit} lead to visible quantization errors and thus contouring artifacts (Property 4). It has been reported that skilled observers can see differences in u', v' of only about 0.002 (0.82 for u_{8bit} and v_{8bit}) (see [Hunt 1995], p. 154), which is still below the maximum quantization error $u_{8bit} \pm 0.5$ and $v_{8bit} \pm 0.5$. For validation, we displayed a chromacity diagram for quantized u_{8bit} and v_{8bit} for several luminance levels on a calibrated monitor. We could see contouring artifacts for blue and purple colors for the highest luminance levels, which would suggest that 8-bit encoding does not give sufficient precision. This is alleviated by either limiting the color gamut or using perceptually more uniform color space (the u', v' chromacity diagram is only approximately uniform and the ratio between the smallest and the largest color difference can exceed four to one). However, such artifacts are not expected to be noticeable in complex images.

5.3.3 Application to Image and Video Compression

The proposed color space for HDR pixels has been successfully used in three image and video compression algorithms: an HDR extension to MPEG compression described in

Section 5.4; a backward compatible HDR video compression described in Section 5.6, and HDR image compression outlined below.

The algorithm for encoding static HDR images is mostly based on the JPEG image encoding with a few extensions added to accommodate HDR data. Instead of YCrCb we use the color space derived in Section 5.3. Since luminance in this color space is encoded with 12 bits, both DCT transformation and variable-length coding are extended to support larger values. The results show that our DCT-based image compression for HDR images is both efficient and fast. The algorithm is specifically developed to be included in the Open Source OpenEXR library (http://www.openexr.org/) as a freely available and efficient lossy compression format for HDR images. More information on OpenEXR format can be found in Section 5.2.3.

Although the HDR image video encoding has not been well established so far, many practical applications would benefit greatly by providing more precise, possibly calibrated streams of temporally coherent data. The proposed color space for HDR pixels relies on insensitivities of the HVS in terms of luminance and contrast perception, and therefore it is appropriate for all those applications whose goal is to reproduce the appearance of images as perceived by the human observer in the real world. This assumption matches well to such applications as realistic image synthesis in computer graphics, digital cinematography, documenting reality, tele-medicine, and some aspects of surveillance.

Linear HDR data encoding is required by many applications, such as re-lighting using dynamic HDR environment maps. Linear or logarithmic HDR encoding might be desirable in remote sensing, space research, and typical computer vision applications such as monitoring, tracking, recognition, and navigation. For other applications, custom quantization algorithms can be required, for example to match sensor characteristics used to acquire HDR data in medical applications. In such a case the luminance quantization approach (Section 5.3.1) can be easily adapted.

5.3.4 Discussion

The luminance encoding proposed in this section can be considered an extension of typical gamma correction for the full range of luminance values visible to the human eye. Obviously, "gamma correction" is not the correct term for the proposed nonlinearity since we do not correct voltage of cathode ray tubes. Nevertheless, it is worth pointing out that both the gamma correction and the proposed nonlinearity are consistent in the luminance range from about 1 to 500 cd/m^2 (i.e. the luminance range in which typical CRT and LCD displays operate). In this range both nonlinearities are modelled as a power function with the exponent being less than one (see Equation 5.9).

Interestingly, there is also an analogy between the derived $l(y)$ function and the response of a typical film negative. The film response, as shown in Figure 2.7, consists of five segments: D_{min} (minimum density), the toe, the straight-line segment, the shoulder, and D_{max} (maximum density). Such a characteristic is also known as the *D-logE* curve. If we compare the film response from Figure 2.7 with the $l(y)$ function from Figure 5.8, we notice that our visual system has a minimum response (below 0.01 cd/m^2) followed by the segment of gradually increasing slope, which corresponds to the toe in the film response. The visual system shows a logarithmic response above $1,000$ – $10,000$ cd/m^2, similar to the straight-line segment (on log-linear plot) for a film. The

difference is that the visual system, unlike a film, does not saturate for high luminance when it is adapted to these luminance levels. In addition, the eye can perceive simultaneously much larger dynamic range of luminance than a film can capture.

One difficulty that arises from our color encoding is that the source HDR images must be calibrated in absolute units of cd/m^2. The pixel values must be also given using absolute XYZ values, where Y represents an absolute luminance value (traditional XYZ coordinates are normalized to be within the range 0-100 and Y represent luminance factor instead of luminance). This is necessary since the performance of the HVS is significantly affected by the absolute luminance levels. For instance, the detection thresholds are significantly higher for low light conditions. The major source of this problem are the existing HDR capture techniques, such as multi-exposure methods, which give an accurate measurement of relative luminance (luminance factor), but give no information on absolute luminance levels. The conversion from relative to absolute luminance units is however very simple and requires multiplication of all XYZ color coordinates by a single constant. Such a constant needs to be measured only once for a camera. The measurement can be done by capturing a scene containing a uniform light source of known illuminance or a surface of measured luminance [Krawczyk et al. 2005a]. If such a measurement is not possible, an approximate calibration of an image to absolute units, by assuming typical luminance levels of some objects (e.g. the sky or a daylight illuminated wall), is usually sufficient.

Although we strongly support scene-referred encoding of image and video we also see some problems related to this approach. A substantial part of the visual material created today is not an exact replica of the real world, but rather stems from human or computer-created or enhanced images, which are only intended to look like the real scenes. For instance, night scenes in movies are often shot at daylight and then post-processed to give them a nocturnal look. How should such scenes be encoded if they intend to represent low light conditions but are displayed at much higher luminance levels? In such cases, scene-referred encoding of images may not be appropriate and images should represent the intended appearance of a scene. Nevertheless, such images should be stored in an HDR "appearance-referred" format, which would encode the optimal luminance levels at which particular scene should be displayed. If a display device is not capable of displaying such an image, it would apply a tone mapping algorithm [Reinhard et al. 2005] to deliver the best image for its capabilities.

Figure 5.9 shows comparison of the proposed luminance encoding to other popular encoding methods. Although 12-bit encoding derived from CIE t.v.i. function results in quite high relative errors, especially in the low luminance range, these errors are below perceivable threshold. This suggests that all other encodings are conservative in this matter and should not lead to perceivable distortions. However, the proposed encoding can achieve this goal using only 12 bits while other encodings need 16 bits to encode luminance. Note that although Half floating point numbers (S5E10) used on OpenEXR format give the highest precision for values between 10^{-4} – $10^{4.8}$, they cannot represent numbers above 65,504, which makes this format less suitable for storing real-world absolute luminance values.

To enrich visual information stored in image or video files, we postulate scene-referred encoding in favor of device-referred representation commonly used today. HDR images are an example of such scene-referred encoding, which unlike plain images can represent the whole visual information visible to the human eye. We show that HDR scene-referred images and video can be efficiently encoded. We derive a color space for

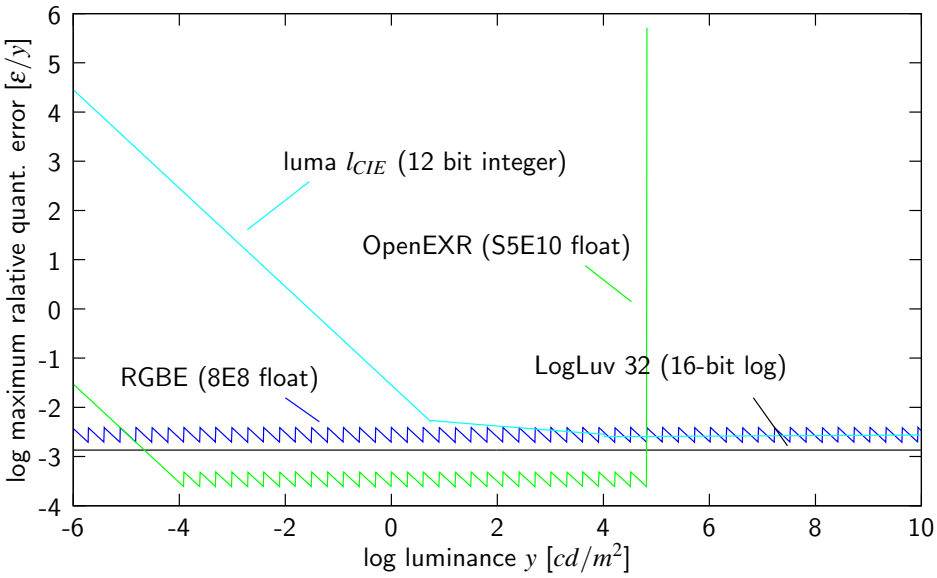


Figure 5.9: Quantization error of popular luminance encoding formats. *luma* l_{CIE} — the perceptual encoding derived from CIE t.v.i. (refer to Section 5.3.1); *OpenEXR* — 16-bit floating point encoding used in OpenEXR format (Section 5.2.3); *LogLuv* — 16 bit logarithmic encoding used in 32-bit version of LogLuv format (Section 5.2.2); *RGBE* — shared mantissa floating point encoding used in Radiance's RGBE format (Section 5.2.1). The error is computed as the maximum distortion in luminance due to rounding error of particular representation and given in relative units (ε/y, where ε is the absolute error given in luminance values). The edgy shape of both *RGBE* and *OpenEXR* is caused by rounding the mantissa. Note that S5E10 float format used in OpenEXR cannot store values larger than 65,504 and therefore its plot is cut a this point.

efficient encoding of HDR data from the detection thresholds of the HVS. To test and demonstrate efficiency of our approach, we implemented a complete HDR MPEG-4 encoder, discussed in the next section.

5.4 HDR Extension of MPEG-4

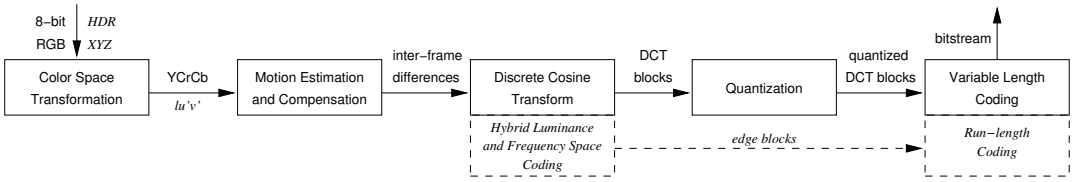


Figure 5.10: Simplified pipeline for the standard MPEG video encoding (black, solid) and proposed extensions (italic, dashed) for encoding High Dynamic Range video. Note that *edge blocks* are encoded together with DCT data in the HDR flow.

This section explains how the MPEG encoding standard, both the Advanced Simple Profile (ISO/IEC 14496-2) [ISO-IEC 14496-2 1999] and the Advanced Video Coding (ISO/IEC 14496-10) [ISO/IEC 14496-10 2005], can be extended to handle HDR data. As a framework for HDR video encoding we selected the MPEG-4 standard, which is state-of-the-art in general video encoding for low dynamic range (LDR) video. Recent studies demonstrate that wavelet transforms extended into the temporal domain and coupled with motion prediction can also be successfully applied for LDR video compression (e.g. [Shen and Delp 1999]), but no wavelet-based standard utilizing inter-frame compression has been established so far.

The scope of required changes to MPEG-4 encoding is surprisingly modest. Figure 5.10 shows a simplified pipeline of MPEG-4 encoding, together with proposed extensions. While a standard MPEG-4 encoder takes as an input three 8-bit RGB color channels, the HDR encoder must be provided with pixel values in the absolute XYZ color space [CIE 1986]. Such color space can represent the full color gamut and the complete range of luminance the eye can adapt to. Next pixel values are transformed to the color space that improves the efficiency of encoding. MPEG-4 converts pixel values to one of the family of YC_BC_R color spaces, which exhibit low correlation between color channels for a natural images. As illustrated in Figure 5.10, the proposed extension to MPEG stores color information using a perceptually linearized $lu'v'$ introduced in Section 5.3. We choose an 11-bit representation of luminance as it turns out to be both conservative and easy to introduce to the existing MPEG-4 architecture.

The next stage of MPEG-4 encoding involves motion estimation and compensation (refer to Figure 5.10). Such inter-frame compression results in significant savings in bit-stream size and can be easily adapted to HDR data. After the motion compensation stage, inter-frame differences are transformed to a frequency space by the Discrete Cosine Transform (DCT). The frequency space offers a more compact representation of video and allows perceptual processing.

A perceptually motivated quantization of DCT frequency coefficients is the lossy part of the MPEG-4 encoding and the source of the most significant bit-stream size saving. Although the MPEG-4 standard assumes only the quantization of LDR data of a display device, in Section 5.4.1 we generalize the quantization method to the full range of visible luminance in HDR video.

Due to quantization of DCT coefficients, noisy artifacts may appear near edges of high-contrast objects. While this problem can be neglected for LDR data, it poses a significant problem for HDR video, especially for synthetic sequences. To alleviate this, in Section 5.4.2 we propose a hybrid frequency and luminance space encoding, where sharp edges are encoded separately from smoothed DCT data.

In the following sections we describe our extensions to the MPEG-4 format, which are required for efficient HDR video encoding. For detailed information on the MPEG-4 encoding refer to the standard specification [ISO-IEC 14496-2 1999].

Additional examples and the demonstration video can be found on the project web page: http://www.mpi-inf.mpg.de/resources/hdrvideo/index.html.

5.4.1 Quantization of Frequency Components

The color space $lu'v'$ derived in Section 5.3 takes account for a non-linear response of the visual system to light (luminance masking) at a broad range of luminance adaptation levels. However, the loss of information in the human eye is limited not only by the thresholds of luminance contrast but also by the spatial configuration of image patterns (spatial and temporal contrast sensitivity and contrast masking). To take full advantage of those HVS characteristics, MPEG encoders apply the Discrete Cosine Transform DCT to each 8×8 pixel block of an image. Then each DCT frequency coefficient is quantized separately with the precision that depends on the spatial frequency it represents. As we are less sensitive to high frequencies (refer to Section 3.6), larger loss of information for high frequency coefficients is allowed. In this section we show that the MPEG-4 quantization strategy for frequency coefficients can be applied to HDR data.

In MPEG encoders, the quantization of frequency coefficients is determined by a quantization scale q_{scale} and a weighting matrix W. Frequency coefficients F are changed into quantized coefficients $\hat{F}$ using the formula:

$$\hat{F}_{ij} = \left[\frac{F_{i,j}}{W_{i,j} \cdot q_{scale}} \right] \text{ where } i,j = 1..8 \tag{5.13}$$

The brackets denote rounding to the nearest integer and i,j are indices of the DCT frequency band coefficients. The weighting matrix W usually remains unchanged for whole video or a group of frames, and only the coefficient q_{scale} is used to control quality and bit-rate. Note that the above quantization can introduce noise in the signal that is less than half of the denominator $W_{i,j} \cdot q_{scale}$.

Both the HDR perceptually quantized space $lu'v'$ and the gamma corrected YC_BC_R space of LDR pixel values are approximately perceptually uniform [Nadenau 2000, Section 7.2.2]. In other words, the same amount of noise results in the same visible artifacts regardless of the background luminance. If quantization adds noise to the signal that is less than half of the denominator of equation 5.13, quantizing frequency coefficients using the same weighting matrix W in both spaces introduces artifacts, which differ between those spaces by a roughly constant factor. Therefore to achieve the same visibility of noise in the HDR space as in LDR space, the weighting matrix W should be multiplied by a constant value. This can be achieved by setting a proper value of the coefficient q_{scale}.

The default weighting matrices currently used in MPEG-4 for quantization [ISO-IEC 14496-2 1999, Section 6.3.3] are tuned for typical CRT/LCD display conditions and luminance adaptation levels around 30–100 cd/m^2. Contrast sensitivity studies [Van Nes and Bouman 1967] demonstrate that the HVS is the most sensitive when adapted to the luminance of several hundred cd/m^2 and the corresponding threshold values essentially remain unchanged across all higher luminance adaptation values. On the other hand, the threshold values significantly increase for the lower luminance adaption levels. This means that MPEG-4 weighting matrices are conservative for HDR data. More effective and still conservative quantization can be expected if separate weighting matrices are used for lower luminance levels. However, this requires additional storage overhead, as updated matrices have to be encoded within the stream. Moreover, such adaptive quantization requires multi-pass encoding, which restricts possible applications. Another solution is prefiltering of input images to remove imperceptible spatio-temporal

frequencies [Border and Guillotel 2000]. Pre-filtering of HDR video will be discussed in detail in Section 5.6.5.

5.4.2 Encoding of Sharp Contrast Edges

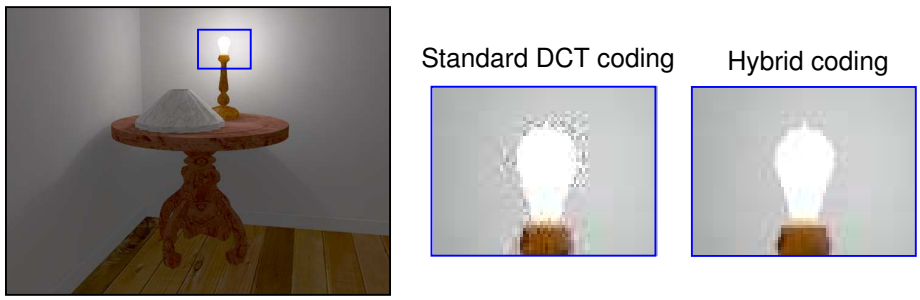


Figure 5.11: Quality comparison of the standard DCT coding of the block and our hybrid frequency and luminance space coding. Quantized DCT blocks show artifacts at sharp edges, which are not visible for the hybrid encoding. The hybrid encoding increased size of the bit-stream by 7%.

In the previous section we showed that the quantization of DCT coefficients can be safely applied to the perceptually quantized HDR space thus greatly reducing the size of the video stream. Unfortunately, the DCT is not always an optimal representation for HDR data. HDR images can contain sharp transitions from low to extremely high luminance values, for example at the edges of light sources. Information about sharp edges is encoded into high frequency DCT coefficients, which are coarsely quantized. This results in visible noisy artifacts around edges, as can be seen in Figure 5.11. This is especially pronounced in the case of synthetic images, which often contain sharp luminance transitions between neighboring pixels. To solve this problem we propose a hybrid encoding, which stores separately low-frequency data in DCT blocks and elevation of sharp edges in "edge blocks".

Figure 5.12 illustrates how, in case of 1D data, input luminance that contains a sharp edge can be split into two signals: One piece-wise constant that contains the sharp edge alone and another that holds slowly changing values. The original signal can be reconstructed from those two signals. Due to the fact that sharp edges occur in sequences relatively infrequently, the signal that stores them can be effectively encoded. The second signal no longer contains large values of high frequency coefficients and can be transformed into a compact DCT representation.

A process of hybrid encoding of a single 8×8 block is shown in Figure 5.13. The original block (5.13a) contains a part of a stained glass from the "Memorial Church" HDR image. To isolate sharp edges from the rows of this block, we use a simple local criterion: If two consecutive pixels in a row differ by more than a certain threshold (discussed in the next paragraph), they are considered to form a sharp edge. In such case the difference between those pixels is subtracted from all pixels in the row, starting from the second pixel of that pair up to the right border of the block. The difference itself is stored in the edge block at the position of the second pixel of that pair. The algorithm is repeated for all 8 rows of the block. This step is shown in Figure 5.13b. After

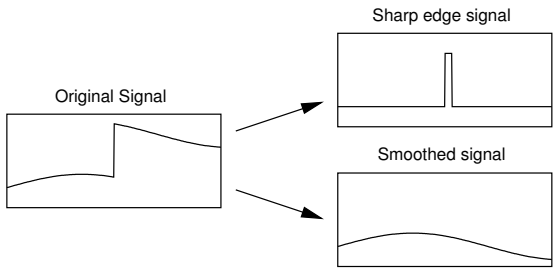


Figure 5.12: Decomposition of a signal into sharp edge and smoothed signals.

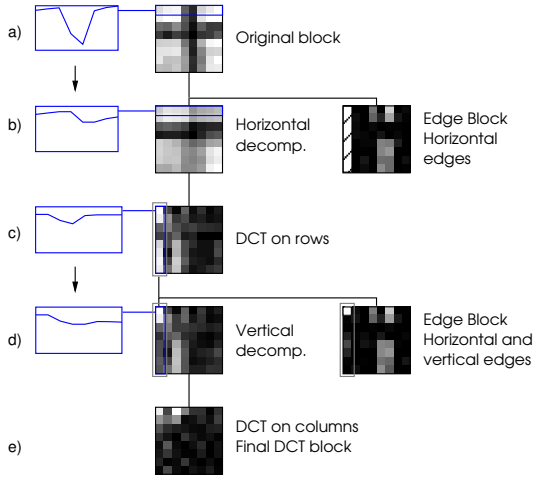


Figure 5.13: Steps of a hybrid frequency and luminance space coding of a single 8×8 block. Blue insets on the left show a cross-section of the first row (a and b) and the first column (c and d) of the block values. Note how the curves are smoothed as edges are removed from the block, resulting in lower values for the high frequency DCT coefficients.

the rows have been smoothed, they can be transformed to DCT space (Figure 5.13c). Due to the fact that the smoothed and transformed rows contain large values only for the DC frequency coefficients, only the first column containing those coefficients has to be smoothed in order to eliminate sharp edges along the vertical direction. We process that column in the same way as the rows and place resulting "edges" in the first column of the edge block (Figure 5.13d). Finally, we can apply a vertical DCT (Figure 5.13e).

Most of the values of the resulting edge blocks are equal to zero and can be compressed using run-length encoding. However, because this is still more expensive in terms of bit-rate than encoding DCT blocks alone, only the edges that are the source of visible artifacts should be coded separately in edge blocks. The threshold contrast value that an edge must exceed to cause visible artifacts depends on the maximum error of the quantization (refer to Section 5.4.1) and can be estimated. Table 5.1 shows such thresholds for MPEG-4 standard quantization matrices and 11-bit encoded luminance in the l space (refer to Section 5.3). The thresholds were found for an estimated quantization error greater than 1 Just Noticeable Difference (JND), where 1 JND equals 13.26 units

q_{scale}	1–5	6	7	8	9–31
Threshold inter	n/a	936	794	531	186
Threshold intra	n/a	n/a	919	531	186

Table 5.1: Threshold contrast values of a sharp edge above which artifacts caused by DCT quantization can be seen. The values can be used to decide whether a sharp edge should be coded in a separate edge block. The thresholds are given for different compression quality factors q_{scale} and for both intra- and inter-encoded frames (since MPEG-4 uses different weighting matrices to quantize intra- and inter-encoded frames). Note that for $q_{scale} \leq 5$ noisy artifacts are not visible and no hybrid encoding is necessary.

of the l space. Note that the lowest threshold equals 186, which corresponds to the local luminance contrast 1:30 for mesopic and 1:5 for photopic range (see Figure 5.8). Because such high contrast between neighbouring pixels rarely occurs in low dynamic range images, hybrid coding shows visible improvement of quality for high contrast HDR video.

The proposed hybrid block coding improved quality of encoded sequences at the cost of a larger bit-stream (see Figure 5.11). The artifacts that the hybrid coding can eliminate are mostly visible in synthetic and non-photorealistic images, since those often contain smooth surfaces that do not mask noise. Such artifacts can not be eliminated in post-processing, like blocky artifacts of the DCT. The hybrid coding gives additionally more localized control over the quality than q_{scale} factor. This way, it is possible to remove salient high frequency artifacts while the overall quality is kept the same. Although the hybrid encoding is not strictly necessary to encode HDR video, it solves the problem of encoding high values of frequency coefficients, which would otherwise require extended variable-length coding tables. We noticed that using the standard MPEG-4 variable-length coding of AC coefficients is sufficient for HDR video when the hybrid block coding is used.

5.4.3 Implementation Details

In this section we outline technical details of our implementation of HDR compression and playback.

Our HDR encoder / decoder is based on the XviD library[6], which is an open source implementation of the Simple Profile ISO MPEG-4 standard [ISO-IEC 14496-2 1999]. We extended this implementation to support an encoding of DCT coefficients using more than 8-bits per color channel (NOT_8_BIT). This let us encode perceptually quantized luminance (l, refer to Section 5.3) represented as 11-bit integers. The two color channels $u'v'$ are sub-sampled to half of the resolution of the original image and encoded with 8-bit precision. The special treatment of sharp contrast edges (refer to Section 5.4.2) is applied only to the luminance channel. The edge blocks are encoded in the video stream together with DCT blocks. To reduce impact on the stream size, only those edge blocks are encoded that are not empty (less than 7% for our test sequences). The non-empty blocks are compressed using a run-length encoding. More effective coding of the edge blocks could further improve compression.

[6] XviD project home page: http://www.xvid.org

To playback an HDR video we created a player capable of decoding, tone mapping, and applying post-processing effects in real-time. To achieve such performance we had to overcome the bottleneck problem of CPU-to-GPU memory transfer. A naive approach would be transferring HDR frames to the GPU as 16- or 32-bit floating point RGB textures. Instead, we send data in the $lu'v'$ format (11-,8-,8-bit, refer to the previous paragraph). The $lu'v'$ format gives a gain of 20-40% of a texture size without any visible degradation of quality. Color conversion from the $lu'v'$ to RGB format is implemented effectively using fragment shaders and thus lowering CPU load on MPEG decoding.

To apply real-time global tone mapping to the video, we employed a simple lookup table approach. Because the number of possible values of the quantized luminance (luma), l, is small (2048 for 11 bits), we use tone mapping function only for the 2048 corresponding real-world luminance values and send the resulting values them to the graphics card as a 1D texture. We later use dependent texture lookups to find the values of tone mapped pixels. Tone mapping parameters and computationally expensive variables, such as logarithmic mean luminance of a frame, are provided within the bitstream as an annotation script. This way any global tone mapping operator can be implemented with a marginal effect on performance. On a Pentium IV 2.4GHz processor and an ATI Fire GL X1 graphic card we were able to decode and display about 30 frames per second for a sequence of the resolution 640×480.

5.4.4 Results

Computer graphics animations, panoramic images, and video captured using specialized HDR cameras were used for testing the proposed HDR extension of MPEG. The OFFICE sequence is an example of indoor architectural walk-through rendered using global illumination software with significant changes of illumination levels between rooms (Figure 5.14). The camera panning was simulated for the CAFETERIA panorama obtained using the *SpheronVR PanoCam* camera. The scene contains both a dim cafeteria interior and a window view on a sunny day (Figure 5.15). To capture natural grayscale sequences we used a Silicon Vision Lars III HDR video camera, which returned linear radiance values. The LIGHT sequence shows a direct view of halogen lamp which illuminates objects with different reflectance characteristics (Figure 5.16).

As we discussed in Section 5.3.4, our perceptual quantization strategy for luminance values performs the best for HDR video calibrated in terms of luminance values. Such calibrated data are immediately available for our computer animations resulting from the global illumination computation. We also performed a calibration procedure for the Lars III HDR video camera, using a Kodak GrayCard with 80% reflectance. For the remaining video material we assigned a common sense luminance level for selected scene regions and then rescaled all pixel intensities accordingly.

To give an overview of the capabilities of the proposed HDR video encoding, we compared its compression ratio with state-of-the art LDR video compression and existing intra-frame (static image) HDR encoding.

Although LDR and HDR video compression store a different amount of information and their performance cannot be matched, such comparison can give a general notion of the additional overhead required to store HDR data. To compare the performance of LDR and HDR encoding, each test sequence was compressed using our HDR encoder,

Figure 5.14: OFFICE sequence with simulated low-level lightning, dynamic range $-4.0 \div 0.2[\log cd/m^2]$. The main frame is tone mapped using the Pattanaik et al. [Pattanaik et al. 2000] algorithm. Lack of colors and the bluish cast are due to the night vision post-processing as proposed by Thompson et al. [Thomspon et al. 2002]. The exploration window reveals color and details in the $-2.2 \div -1.2[\log cd/m^2]$ range. The scene model courtesy of VRA, GmbH.

decompressed, and tone mapped to LDR format. Then the same source HDR sequence was tone mapped, encoded to MPEG-4 using the FFMPEG[7] encoder (LDR MPEG-4 ISO/IEC 14496-2), and decoded. The quality of the resulting frames from both LDR and HDR encoding was measured using the Universal Quality Index [Wang and Bovik 2002], which gives more reliable quality measure than PSNR and at the same time is less computationally expensive as VDP. Next, we matched pairs of LDR and HDR streams that had a similar quality index, and compared their sizes. The results are shown in Table 5.2.

Only after some time we noticed that the employed method of comparing HDR and LDR video compression was very disadvantageous for the HDR compression. The tone mapping that was used to reduce the dynamic range before LDR MPEG-4 compression was very effective at eliminating noise in the original sequence, which had to be encoded in HDR video stream. However, the same tone mapping was not so effective at eliminating compression artifacts after decoding the HDR content. Therefore, the compression performance of HDR video was strongly affected by the low-amplitude noise in the source sequences. More adequate quality / bit-rate comparison, for larger number of quality settings and with the application of several quality metrics, will be

[7] FFMPEG project home page: http://ffmpeg.sourceforge.net/

Figure 5.15: CAFETERIA sequence, dynamic range $-1.9 \div 3.6[\log cd/m^2]$. The background frame is clamped to a displayable range. Our dynamic range exploration tool, visible as two windows, shows a luminance range $-1.0 \div 1.0[\log cd/m^2]$ (lower right) and a high luminance range $1.0 \div 3.0[\log cd/m^2]$ (upper left). Details in these windows are not visible in LDR video. The source panorama courtesy of Spheron, Inc.

presented in Section 5.6.7.

The OpenEXR format, which we described in Section 5.2.3, offers nearly lossless encoding (up to quantization precision of 16-bit floating point numbers) and intra-frame compression, i.e., each frame is compressed separately. The performance of such compression can be expected to be below that of inter-frame DCT based encoding used in our encoder. However, the OpenEXR format is commonly used for storing animation frames and we decided to include it in the performance summary in Table 5.2.

5.4.5 Summary

This section presents a technique for encoding high-dynamic range (HDR) video, which requires only modest extensions of the MPEG-4 compression standard. The first component of our technique is a color space for HDR pixels derived from contrast detection characteristic of the human eye. Such color space requires only 11–12 bits to encode the full perceivable luminance range (15 orders of magnitude) and ensures that the quantization error is always below visibility thresholds. The second component is an efficient scheme for handling the DCT blocks with high contrast information by decomposing them into two layers of LDR details and HDR edges, which are separately encoded. The size of a HDR video stream encoded by our technique increases less than

Figure 5.16: LIGHT sequence captured with the HDR video camera, dynamic range $0.3 \div 4.9[\log cd/m^2]$. Details of the halogen bulb are well preserved despite high luminances. The visible range in exploration tool window is $2.9 \div 4.9[\log cd/m^2]$.

two times with respect to its LDR version.

The strengths of the HDR video encoding method can be fully exploited for HDR displays, but the method can be beneficial for LDR displays as well. HDR information makes it possible to adjust tone mapping parameters for any display device and surround lighting conditions, which improves the quality of video reproduction.

5.5 Backward Compatible Compression

Since LDR file formats for images and video, such as JPEG or MPEG, have become widely adapted standards supported by almost all software and hardware equipment dealing with digital imaging, it cannot be expected that these formats will be immediately replaced with their HDR counterparts. To facilitate transition from output-referred LDR to scene-referred HDR imaging, there is a need for backward compatible HDR formats, that would be fully compatible with existing LDR formats and at the same time would support enhanced dynamic range and color gamut. Moreover, if such a format is to be successful and adopted by large part of the market, the overhead of HDR information must be very low, preferably below 30% of the LDR file size. This is because at the beginning very few consumers will have access to HDR technology, such as HDR displays, and the rest of the consumers will not accept doubling the size of the file for the sake of the data they cannot take advantage of. Such backward compatible

	MPEG-4		HDR Enc.		OpenEXR	
Video Clip	ratio	bpp	ratio	bpp	ratio	bpp
OFFICE hq	0.54	0.27	1.00	0.51	32.17	16.27
OFFICE lq	0.51	0.05	1.00	0.10		
LIGHT hq	0.56	0.71	1.00	1.25	22.56	28.25
LIGHT lq	0.57	0.10	1.00	0.18		
CAFETERIA hq	0.63	0.12	1.00	0.19	142.58	27.40
CAFETERIA lq	0.54	0.05	1.00	0.09		

Table 5.2: Comparison of compression performance of LDR MPEG-4, the proposed HDR encoding, and the OpenEXR format. "ratio" is a relative bit-stream size increase or decrease compared to our encoding. "bpp" denotes bits per pixel. "hq" and "lq" next to the video clip name means high quality and low quality respectively. There are empty entries for low quality OpenEXR because this format does not support lossy compression. The proposed HDR encoding gives about half of the compression ratio of MPEG-4 (see also a note in the text). High compression gain of MPEG-4 and HDR encoding for the CAFETERIA video clip can be explained by efficient motion compensation in camera panning.

encoding would also require that the original LDR content is not modified. Although the compression of HDR can be improved if an LDR image can be slightly altered, this would also be unacceptable for majority of customers who do not want to have their LDR content modified.

The following subsections present an overview of both existing and possible solutions for backward compatible image and video encoding. This state of the art summary is a starting point for Section 5.6, which introduces a novel backward compatible format for images and video that offers several efficiency improvements over existing solutions.

5.5.1 Bit-depth Expansion Techniques

The problem of dynamic range compression and expansion arises in many imaging pipelines with constrained bit depth at certain processing stages (only 6 bits per color channel is often used for DVD movies while displays can handle 8 bits/color channel). This may result in the loss of low amplitude signals and false contouring. Bit depth expansion (BDE) techniques are designed specifically to combat those effects and achieve higher perceived bit depth quality than are physically available. For example imperceptible spatio-temporal noise is added to an image prior to the quantization step in dither techniques [Daly and Feng 2003]. Intensity averaging in the optics of display and human eye leads to recovering information below the quantization step. Modern BDE techniques tune a micro-dither amplitude taking into account the interaction of display nonlinearities to obtain a low-spatial frequency flicker from mutually high-pass spatial and temporal noise and achieve 10-bit perceived quality on 8 bit-driver LCDs. When higher bit depth information is not available, low-amplitude details cannot be reconstructed, and processing is focused on removing false contours using adaptive filtering, predictive cancellation, spatial frequency channel coring techniques [Daly and Feng 2004]. All existing BDE and de-contouring techniques are optimized for much lower bit depth expansion than required to accommodate HDR image and video content.

Furthermore, storing HDR video using 8-bit encoding with additional spatio-temporal dither is impractical because dither patterns do not compress well. Also, to make BDE techniques working possible, lossless encoding of the dither pattern into a video stream is required, which may significantly affect the compression performance.

5.5.2 JPEG HDR

Spaulding et al. [Spaulding et al. 2003] showed that the dynamic range and color gamut of typical sRGB images can be extended using residual images. Their method is backward compatible with the JPEG standard, but only considers images of moderate dynamic range. Ward and Simmons [Ward and Simmons 2004] have proposed a backward-compatible extension of JPEG which enables compression of images of much higher dynamic range (JPEG HDR). JPEG HDR is an extension to the JPEG format for storing HDR images that is backward compatible with an ordinary 8-bit JPEG [Ward and Simmons 2004]. A JPEG HDR file contains a tone mapped version of an HDR image and additionally a ratio (subband) image, which contains information needed to restore HDR image from the tone mapped image. The ratio image is stored in user-data JPEG markers, which are normally ignored by applications. This way a naive application will always open a tone mapped version of an image, whereas an HDR-aware application can retrieve the HDR image.

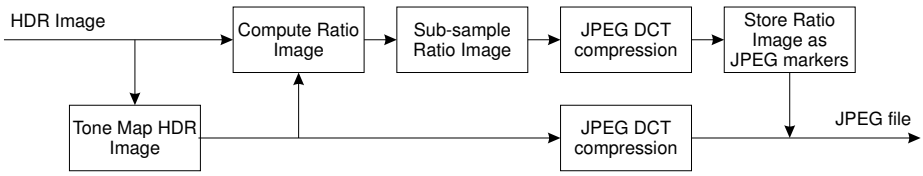


Figure 5.17: Data flow of subband encoding in JPEG HDR format.

A data flow of the subband encoding is shown in Figure 5.17. An HDR image is first tone mapped and compressed as an ordinary JPEG file. The same image is also used to compute ratio image, which stores a ratio between HDR and tone mapped image luminance for each pixel. To improve encoding efficiency, the ratio image is sub-sampled and encoded at lower resolution using the ordinary JPEG compression. After the compression, the ratio image is stored in JPEG markers together with the tone mapped image. To reduce the loss of information due to sub-sampling of the ratio image, two correction methods have been proposed: enhancing edges in a tone mapped image (so called *pre-correction*) and synthesizing high frequencies in the ratio image during up-sampling (so called *post-correction*). Further details on the JPEG HDR compression can be found in [Ward and Simmons 2004] and [Ward and Simmons 2005].

5.5.3 Wavelet Compander

Li et al. [Li et al. 2005] propose that HDR images can be encoded using only 8-bits, if they are undergo a reversible companing operation. They propose a multiscale wavelet architecture, which can compress an HDR image to a lower bit-depth and later expand it to obtain a result that is close to the original HDR image (the so-called compander).

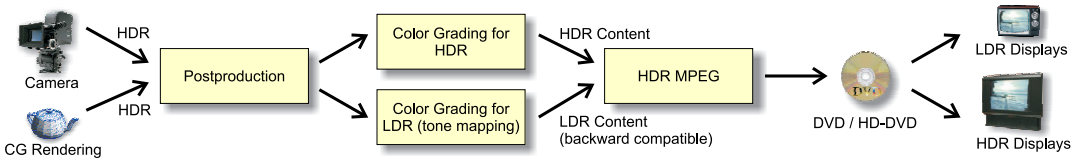


Figure 5.18: The proposed backward compatible HDR DVD movie processing pipeline. The high dynamic range content, provided by advanced cameras and CG rendering, is encoded in addition to the low dynamic range (LDR) content in the video stream. The files compressed with the proposed HDR MPEG method can play on existing and future HDR displays.

The information loss is reduced by amplifying low amplitudes and high frequencies at the compression stage, so that they survive the quantization step to the 8-bit LDR image. Such technique is conceptually similar to *pre-correction* in JPEG HDR. Since the expansion is a fully symmetric inverted process, the amplified signals are properly suppressed to their initial level in the companded HDR image. To further reduce the information loss, the compressed image is iteratively modified to improve the correlation of its subbands with respect to the original HDR image. The authors observe a good visual quality of both the compressed and companded images, but they admit that any guarantee concerning their fidelity to tone mapped (i.e. undergoing just one compression iteration) and original HDR images cannot be given. Moreover, they could clearly see the differences between the corresponding image pairs but they describe them as "visually not disturbing". The obtained PNSR for the companded HDR image is even worse than for ordinary LUT (Look-Up-Table) companding. Given the requirements for a backward compatible image and video compression, the lack of fidelity of tone mapped images is not acceptable, since the original material quality cannot be compromised. Also, the multiscale wavelet framework as proposed by Li et al. severely limits the choice of tone mapping operator. The emphasis on high frequencies at the compression step makes the proposed framework less suitable for standard JPEG and MPEG techniques, which use the quantization matrices that are perceptually tuned to discard a great deal of visually non-important high frequencies. This is confirmed by relatively poor compression rates reported the authors when they attempted to combine JPEG with their companding. Moreover, many existing quantization schemes, which incorporate a visual masking model [Watson et al. 1994, Nadenau 2000], assume that due to a better visibility of the contouring artifacts in smooth image regions than in textured ones a finer quantization is required in those regions. For such image compression approaches superficially high quality coefficients should be set to preserve the high frequency details as required by the compander. It is not clear, how the compander approach can be adopted for lossy HDR video compression, in which apart from just raised quality concerns the issues of temporal coherence and computation efficiency arise (the authors recommend oversampling, i.e., handling the subband computation in the full image resolution to avoid aliasing and do not report any timings for their compander).

5.6 Backward Compatible HDR MPEG

Encoding movies in HDR format is very attractive for cinematography, especially that movies are already shoot with high-end cameras, both analog and digital, that can capture much higher dynamic range than typical MPEG compression can store. To encode cinema movies using traditional MPEG compression, the movie must undergo processing called color grading. Part of this process is the adjustment of tones (tone-mapping) and colors (gamut-mapping), so that they can be displayed on majority of TV sets (refer to Figure 5.18). Although such processing can produce high quality content for typical CRT and LCD displays, the high quality information, from which advanced HDR displays could benefit, is lost. To address this problem, the proposed HDR-MPEG encoding can compress both LDR and HDR into the same backward compatible movie file (see Figure 5.18). Depending on the capabilities of the display and playback hardware or software, either LDR or HDR content is displayed. This way HDR content can be added to the video stream at the moderate cost of about 30% of the LDR stream size. Because of such small overhead, both standard resolution and High-Definition movies can fit in their original storage medium when encoded with HDR information.

The backward compatibility is achieved by encoding the HDR and LDR video frames in an LDR stream that is compatible with MPEG decoders, and a residual stream that enables the restoration of the original HDR stream. To minimize redundancy of information, the residual and LDR streams are decorrelated. Such decorrelation requires perceptually meaningful comparison of the LDR and HDR pixels, which is achieved by introducing a pair of corresponding color spaces that are scaled in terms of the human visual system (HVS) response to luminance and chrominance stimuli. These color spaces are used to build a frame-dependent reconstruction function that approximates values of HDR pixels based on their LDR counterparts. Since the proposed HDR MPEG encoding does not impose any restrictions on LDR or HDR content, both videos can be independently tuned and tone/gamut mapped to achieve the best look on different classes of displays. This tuning flexibility is required for current practices of the DVD industry. To reduce the production costs of HDR DVD players, the compression algorithm is designed so that standard 8-bit MPEG decoding chipsets can be used to decode the HDR stream.

A second major mechanism employed in the outlined compression algorithm is a perception-based HDR filter that predicts the visibility thresholds for HDR frames. The wavelet-based filtering approach, presented in Subsection 5.6.5, is fast as required by video applications, but still models important characteristics of the HVS such as luminance masking, contrast sensitivity, and visual masking for the full visible dynamic range of luminance. We apply our HDR filter to remove invisible noise in the residual video stream taking into account the adaptation conditions and visual masking imposed by the original HDR stream. This leads to even more effective HDR video compression since details that cannot be seen are removed from the residual stream prior to encoding.

More information on this project as well as the demonstration video can be found on the project web page: http://www.mpii.mpg.de/resources/hdr/hdrmpeg/.

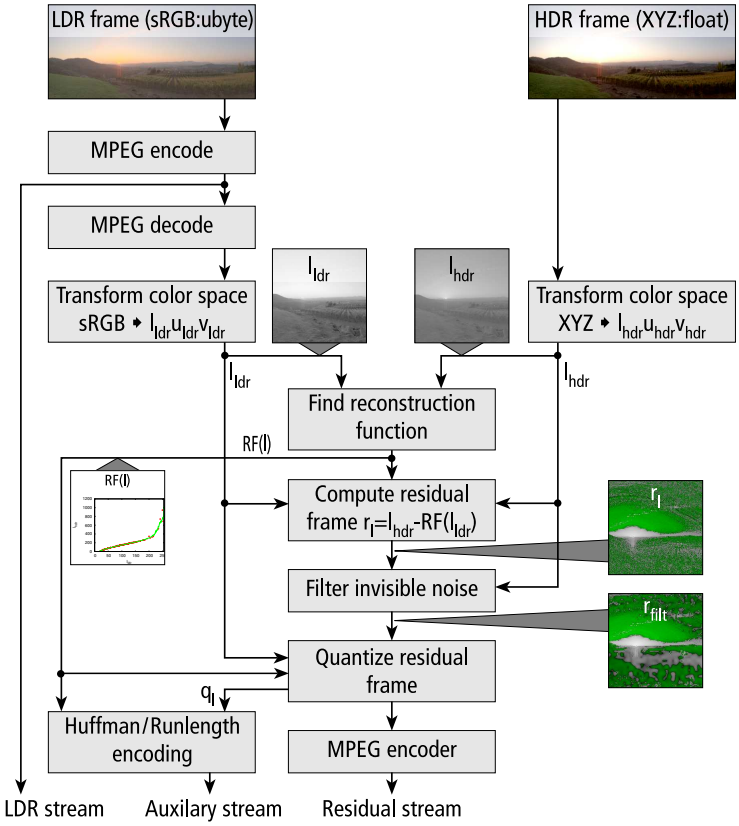


Figure 5.19: A data flow of the backward compatible HDR MPEG encoding. See text for details.

5.6.1 Overview of the Algorithm

The complete data flow of the proposed backward compatible HDR video compression algorithm is shown in Figure 5.19. The encoder takes two sequences of HDR and LDR frames as input. The LDR frames, intended for LDR devices, usually contain a tone mapped or gamut mapped version of the original HDR sequence. The LDR frames are compressed using a standard MPEG encoder (**MPEG encode** in Figure 5.19) to produce a backward compatible LDR stream. The LDR frames are then decoded to obtain a distorted (due to lossy compression) LDR sequence, which is later used as a reference for the HDR frames (see **MPEG decode** in Figure 5.19). Both the LDR and HDR frames are then converted to compatible color spaces, which minimize differences between LDR and HDR colors. The reconstruction function (see **Find reconstruction function** in Figure 5.19) reduces the correlation between LDR and HDR pixels by giving the best prediction of HDR pixels based on the values of LDR pixels. The residual frame is introduced to store a difference between the original HDR values and the values predicted by the reconstruction function. To improve compression, invisible luminance and chrominance variations are removed from the residual frame (see **Filter invisible noise** in Figure 5.19). Finally, the pixel values of a residual frame are quantized (see **Quantize residual frame** in Figure 5.19) and compressed using a

standard MPEG encoder into a residual stream. Both the reconstruction function and the quantization factors are compressed using a lossless arithmetic encoding and stored in an auxiliary stream. The most important steps of the compression algorithm are described in detail in the following subsections while the details, which are sufficient to reimplement the algorithm, are given in the technical report [Mantiuk et al. 2006b].

5.6.2 Color Space Transformations

Both LDR and HDR frames must be transformed to *compatible* and *perceptually uniform* color spaces to enable any comparison between LDR and HDR pixel values and to assess their correlation. The "compatible" color spaces mean here that color channels of both LDR and HDR pixels represent approximately the same information. Perceptual uniformity is needed to estimate color differences according to perceivable, rather than arithmetic, differences. Furthermore, an HDR color space must represent the full color gamut visible to the human eye. To achieve all these goals, we have derived two color spaces: (i) A color space for LDR pixels that encodes chroma using CIE 1976 Uniform Chromacity Scales (u', v', similar to *logLuv* encoding [Ward Larson 1998]) and luma using sRGB nonlinearity, which consist of a linear and power function segments; (ii) A color space for the HDR pixels uses the same u', v' encoding for chroma as the color space for LDR pixels, and a perceptually uniform luminance encoding. The sRGB nonlinearity cannot be used for luminance values ranging from 10^{-5} to 10^{10} cd/m^2, which can be found in real world scenes. Therefore we apply the luminance encoding that has been derived in Section 5.3 from the contrast detection measurements for the full visible range of luminance. This encoding was shown to have similar properties to gamma correction for LDR, but can encode luminance values found in HDR images using 11–12 bits and ensures that the quantization error is below the threshold of visibility. A similar encoding was used in the context of HDR extension to MPEG-4, described in Section 5.4.

5.6.3 Reconstruction Function

Both LDR and HDR frames contain similar information and are therefore strongly correlated. This is illustrated in Figure 5.20, which shows how the luma values of an LDR frame relate to the luma values of an HDR frame. The relation is different for each tone mapping algorithm, but in general it follows an approximately linear function with more variance at high values. Uncorrelated pixels at the right end of the l_{ldr} axis are the result of luminance clamping that is applied in many tone mapping algorithms. Local tone mapping usually results in higher variance and therefore a more "noisy" shape of this relation, while global tone mapping results in a direct one-to-one relationship unless some pixel values are clamped.

The goal of most compression methods is to decorrelate data, so that the same information is not encoded twice. To decorrelate LDR and HDR frames, we find a *reconstruction function*, which predicts the value of an HDR pixel based on the value of the corresponding LDR pixel. Having such a function we need only to encode the differences between values predicted by the reconstruction function and the actual values from an HDR frame. Such differences are usually close to zero and therefore can be efficiently encoded in *residual frames*. The reconstruction function needs to be defined

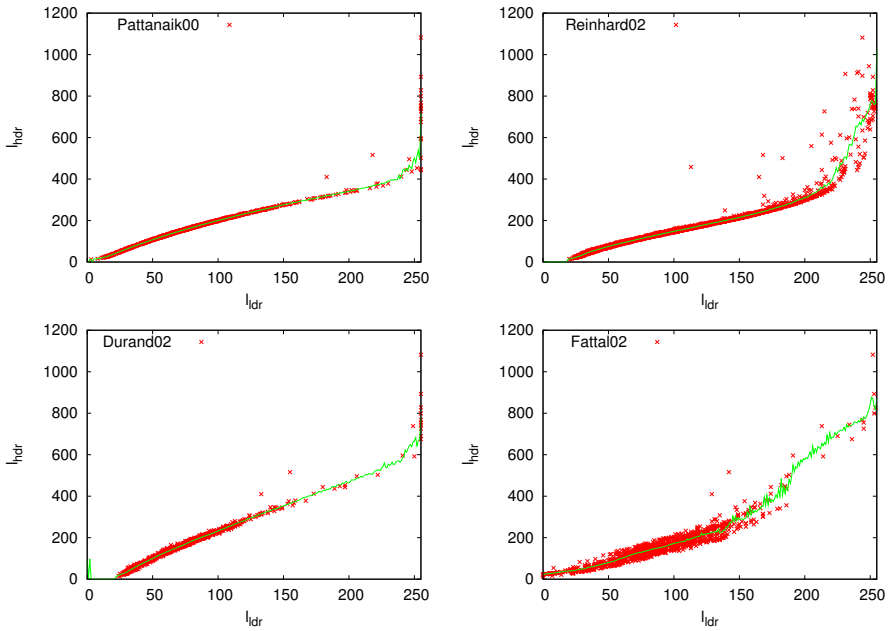


Figure 5.20: The relation between LDR (l_{ldr}) and HDR (l_{hdr}) luma values for various tone mapping algorithms (marked in red) and the corresponding reconstruction functions (marked in green). Tone mapping algorithms (left to right, top to bottom): [Pattanaik et al. 2000], [Reinhard et al. 2002b], [Durand and Dorsey 2002b] and [Fattal et al. 2002]. The relations are plotted for the *Memorial Church* image.

for only 256 values (bins) for 8-bit per channel LDR encoding. The function does not need to be continuous since its major role is to make the values of the residual frame as small as possible. Some examples of reconstruction functions for different tone mapping algorithms are plotted in Figure 5.20 as continuous green lines.

A mapping from LDR values to HDR values is, in the general case, a one-to-many relationship – there are many HDR pixels values that fall in one of 256 bins of the reconstruction function (LDR pixel values). The questions is how to find a value for each bin that would lead to the best compression performance. We experimented with an arithmetic mean, a median and a midrange [8]. While the midrange gave the worst compression ratio, the arithmetic mean and the median exhibited similar performance. We have decided to use an arithmetic mean because of its lower computational cost.

To summarize, we define the reconstruction function as the arithmetic mean of all pixels falling in a corresponding bin Ω_l:

$$RF(l) = \frac{1}{Card(\Omega_l)} \sum_{i \in \Omega_l} l_{hdr}(i) \text{ where } \Omega_l = \{i = 1..N : \; l_{ldr}(i) = l\} \tag{5.14}$$

$l = 0..255$ is an index of a bin, N is the number of pixels in a frame, $l_{ldr}(i)$ and $l_{hdr}(i)$ are luma values of the i-th LDR and HDR pixel respectively.

[8] Midrange is defined as an arithmetic mean of the maximum and minimum value in a set.

We executed a set of tests on video sequences to decide how often a reconstruction function should be updated: each frame, only at each intra-encoded frame (I-frame), or if the update should depend on a difference between consecutive frames. We achieved the best compression ratio when the reconstruction function was updated each frame, while updating it for each I-frame resulted in severe artifacts.

The relation between LDR and HDR frames is complex only for luminance, and color channels can be quite accurately predicted with simple relations $u_{hdr}(i) = u_{ldr}(i)$ and $v_{hdr}(i) = v_{ldr}(i)$. Although this may not be true for some sophisticated gamut mapping cases, we did not find it necessary to compute a reconstruction function for chroma channels for any of the tone mapping operators we tested.

Since the reconstruction function tends to be slowly changing with an increasing slope, we apply an adaptive Huffman algorithm on the differences between the values in consecutive bins to significantly reduce the size of the stored data. The size of the auxiliary data stream, which stores a reconstruction function, is below 1% of the total stream size, therefore its storage overhead is almost insignificant.

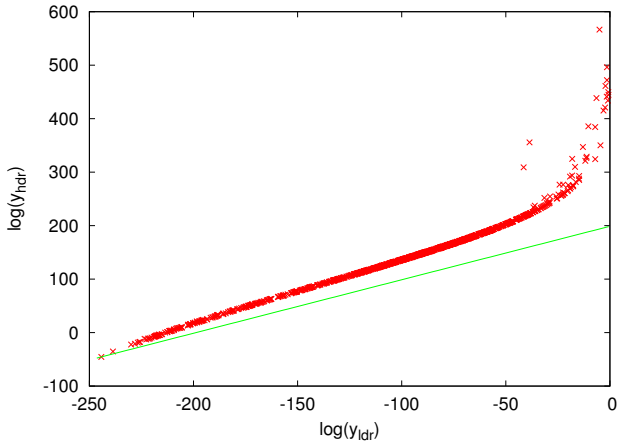


Figure 5.21: A potential reconstruction function for the approach employed in JPEG HDR [Ward and Simmons 2004] compression (marked in green) and a relation between LDR and HDR pixel values (marked in red). The ratio used in JPEG HDR is equivalent to a linear reconstruction function in the logarithmic domain. Such a function does not decorrelate HDR and LDR luma well and therefore reduces compression savings.

We briefly compare our approach with the JPEG HDR compression [Ward and Simmons 2004]. A more detailed comparison will be given in Section 5.6.8. The JPEG HDR compression encodes a ratio between HDR and LDR luminance values, rather than a difference between HDR values and the reconstruction function. However, it can be easily shown that such a ratio is meant to achieve the same goal, which is to decorrelate HDR and LDR pixels. Since a ratio of HDR and LDR luminance corresponds to a difference in the logarithmic domain, and our luminance to luma mapping from Equation 5.9 has roughly logarithmic properties, the ratio encoding of JPEG HDR corresponds to a linear reconstruction function $l_{hdr} = a \cdot l_{ldr}$. As we experimented with such simple reconstruction functions, we found that they give inferior results compared to better fitted ones, like those computed from Equation 5.14. In Figure 5.21 we plot

the reconstruction function used in JPEG HDR. Obviously, it does not follow the data well and some luma information is therefore encoded twice in an LDR and an HDR (subband) stream, which leads to worse compression performance.

5.6.4 Residual Frame Quantization

Although the magnitudes of the differences encoded in residual frames are usually small, they can in fact take values from -4095 to 4095 (for 12-bit HDR luma encoding). Such values cannot be encoded using 8-bit MPEG encoder. Although MPEG standards provide an extension for encoding luma values on 12 bits, such an extension is rarely implemented, especially in hardware. Instead, we would like to reduce the magnitude of residual values so that they can be encoded using a standard 8-bit MPEG encoder.

We have experimented with a non-linear quantization, where large absolute values of residuals were heavily quantized, while small values were preserved with maximum accuracy. Since very few pixels contain a large magnitude of residual, most pixels are not affected by the strong quantization. Such a solution, although giving the best SNR, resulted in poor visual quality for some images. This was because the very few pixels that were heavily quantized attracted attention due to large quantization errors. Therefore the final judgement of quality was mostly based on those few distorted pixels.

A simple clamping of residual values to 8-bit range produced visually better results, but at the cost of losing some details in bright or dark regions. Additionally, to reduce clamping at the cost of a stronger quantization, the residual values can be divided by a constant quantization factor. Such a factor would decide on the trade-off between errors due to clamping and errors due to quantization. Furthermore, we observed that very few bins of a reconstruction function contain residual values that exceed 8-bit range. Therefore the quantization factor can be set separately for each bin, based on the maximum magnitude of the residual that belongs to that bin. Therefore, the residual values after quantization can be computed as:

$$\hat{r_l}(i) = [r_l(i)/q(m)]^{-127 \div 127}, \text{ where } m = k \Leftrightarrow i \subset \Omega_k \tag{5.15}$$

and quantization factor, $q(m)$, is selected separately for each bin Ω_k:

$$q(m) = \max(q_{min}, \frac{\max_{i \in \Omega_l}(|r_l(i)|)}{127}) \tag{5.16}$$

q_{min} is a minimum quantization factor, which is usually set to 1 or 2. $[\cdot]^{-127 \div 127}$ is an operator that rounds the values to the closest integer and then clamps them if they are smaller than -127 or larger than 127. The l subscript in r_l denotes a luma channel.

The quantization factors $q(m)$, where $m = 0..255$, need to be stored in an MPEG stream to later restore non-quantized residual values on the decoding stage. We store quantization factors together with the reconstruction function in the auxiliary data stream. Since quantization factors are usually equal to q_{min} except for a few bins, we found that a run-length encoding followed by the Huffman encoding can effectively reduce the size of this data.

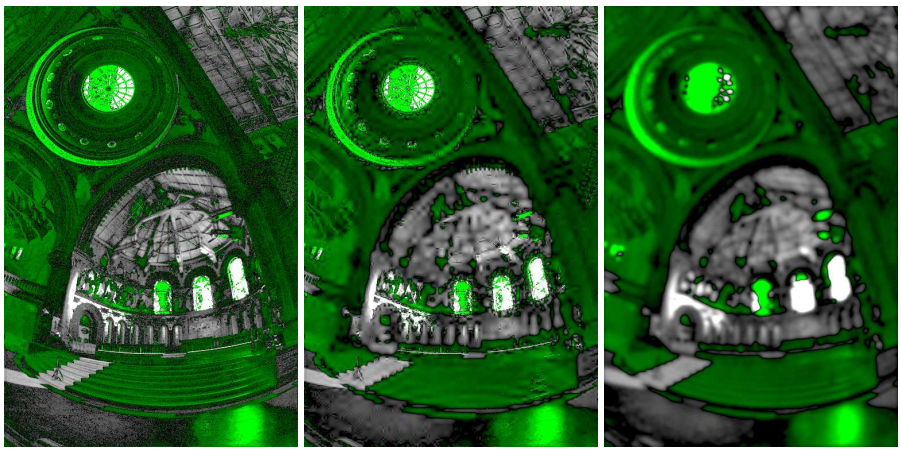


Figure 5.22: Residual frame before (left) and after (center) filtering invisible noise. Details, such as window frame, are lost when low-pass filtering (or downsampling) is used (right). Green color denotes negative values. The *Memorial Church* image courtesy of Paul Debevec.

5.6.5 Filtering of Invisible Noise

Residual frames do not compress well mainly because they contain a large amount of high frequencies. These high frequencies come from three sources: noise in the source HDR images, rounding errors from the tone mapping algorithm, and the DCT quantization errors due to MPEG encoding of LDR frames (refer to Figure 5.19). However, much of this high frequency information does not need to be preserved in the residual stream since it is not visible to the human eye. To remove such invisible noise and thus improve compression efficiency, we introduce a filtering algorithm based on a simplified model of the human visual system (HVS). Although models of the HVS have been used before in CG to control rendering [Ferwerda et al. 1997, Bolin and Meyer 1998, Ramasubramanian et al. 1999], the proposed filtering algorithm has been specially designed to handle HDR data and it has been optimized for speed, so that it can efficiently process video sequences. It is also different from a typical denoising algorithms, e.g. [Bennett and McMillan 2005], since it operates on imperceivable, rather than perceivable noise. It can be used as a standard tool which guarantees that all the visual information that cannot be discerned due to imperfections of the human eye and early vision processing will be filtered out from the image.

The standard MPEG encoding already incorporates many aspects of human vision in order to improve compression efficiency. The gamma corrected color space (or transfer function) accounts for luminance masking (sometimes wrongly named the Weber-Fechner law [Mantiuk et al. 2006c]). The limited spatial contrast sensitivity of the HVS is utilized by the DCT quantization matrix. Two different quantization matrices are used for inter- and intra-frames to take advantage of lower sensitivity to high temporal frequencies. However, contrast masking (or visual masking) is very poorly predicted by the mechanism of MPEG encoding. Since contrast masking is primarily responsible for masking invisible high frequency noise, we focus on modeling this aspect of the HVS to filter residual frames.

There are several methods that incorporate visual masking in image encoding algorithms, such as optimized DCT quantization matrices [Ahumada and Peterson 1993, Watson et al. 1994], the *prequantization* scheme [Safranek 1993], or the point-wise extended masking in the JPEG-2000 standard [Zeng et al. 2000]. However, since all these approaches are either not suitable for video or require significant changes in MPEG encoder/decoder, we decided to use yet another approach, which involves the *prefiltering* of residual frames before they are passed to the MPEG encoder. Prefiltering methods have been shown to improve video compression [Border and Guillotel 2000]. They do not depend on a compression algorithm and therefore do not require any changes to the encoder. The proposed prefiltering algorithm precisely models contrast masking in the wavelet domain, which is quite difficult and inaccurate in the DCT domain. The prefiltering is especially well suited for the residual frames, since they contain mostly low magnitude contrast, while prefiltering involves thresholding of wavelet coefficients that are below the predicted visibility level. If the wavelet coefficients are low, most of them are set to zero and therefore compression efficiency is improved. The prefiltering affects only encoding speed while decoding speed is usually improved due to the reduced stream size.

The input to our residual filtering algorithm consists of two frames: a residual frame (Figure 5.22 left) and an original HDR frame, which is a masker for the residual. Both frames should be stored in the perceptually uniform luma / chroma color space. Output of the filtering is a residual frame with high frequencies attenuated in those regions where they are not visible (Figure 5.22 center). The data flow of the algorithm is shown in Figure 5.23. Though we describe processing that is done on a luma channel, the same processing is performed for two chroma channels, which are subsampled to half of their original resolution. This approximately accounts for the differences between Contrast Sensitivity Function (CSF) for luminance and chrominance.

In the first step we apply the Discrete Wavelet Transform to split a residual frame into several frequency and orientation selective channels. We have experimented with the cortex decomposition [Watson 1987] performed in the Fourier domain, which can better approximate visual channels, but we rejected this approach due to prohibitively long execution times (up to 1 minute per frame). Wavelets, on the other hand, lead to computationally more efficient algorithms and were shown to be useful for modeling many aspects of the HVS [Bradley 1999, Zeng et al. 2000]. We employ CDF 9/7 discrete wavelet basis which is also used for the lossy compression of JPEG-2000. This wavelet basis gives a good trade-off between smoothness and computational efficiency. We use only the three finest scales of the wavelet decomposition since filtering of lower spatial frequencies at coarser scales could lead to noticeable artifacts.

In the next step we account for lower sensitivity of the HVS for high frequencies, which is usually modelled with the Contrast Sensitivity Function (denoted as **CSF** in Figure 5.23). CSF models are described in detail in Section 3.6. We weight each band of wavelet coefficients by a constant value in the same way as is done in JPEG-2000. The weighting factors for a viewing distance of 1 700 pixels ($\approx 1.5 \times$ screen height) are given in the table below.

Scale	LH	HL	HH
1	0.275783	0.275783	0.090078
2	0.837755	0.837755	0.701837
3	0.999994	0.999994	0.999988

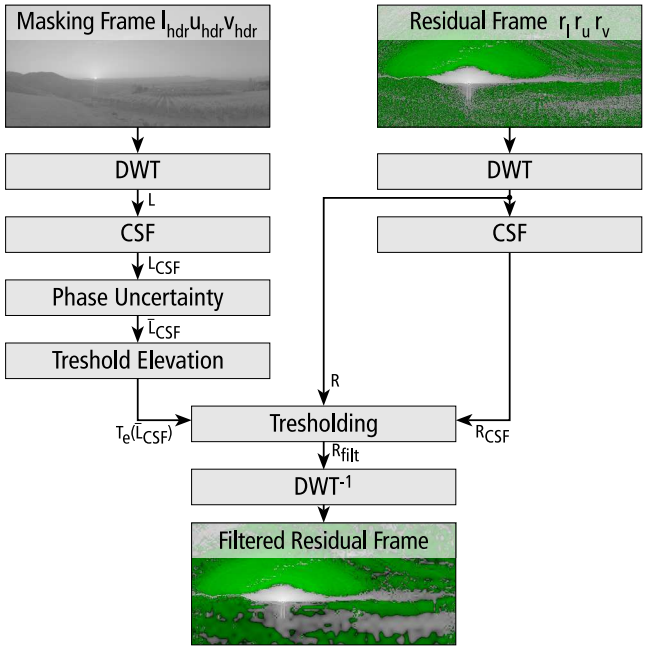


Figure 5.23: A data flow of the residual frame filtering, which removes imperceptible noise for better compression performance.

The visual channels have limited phase sensitivity, ranging from 45° to more than 90°. Because of this, the masking signal affects not only regions where the values of wavelet coefficients are the highest, but may also affect neighboring regions. Phase uncertainty reduces the effect of masking at edges, as opposed to textures which show a high amount of masking. Refer to Section 3.8 for more details on the models of phase uncertainty. Following the point-wise extended masking in JPEG-2000 [Zeng et al. 2000], we model phase uncertainty with the $L_{0.2}$-norm:

$$\bar{L}_{CSF} = \frac{1}{Card(\Theta)} \left(\sum_{\Theta} |L_{CSF}|^{0.2} \right)^{\frac{1}{0.2}} \tag{5.17}$$

where Θ denotes a neighborhood of a wavelet coefficient (we use a box 13×13 kernel in our implementation).

In the following step we predict how contrast thresholds change in the presence of a masking signal, which is an original HDR frame in our case. To model contrast masking (refer to Section 3.7), we employ a threshold elevation function, which we derive from the model proposed by Daly [Daly 1993] (also used in [Ramasubramanian et al. 1999]). We assume a masking slope of 1.0, which was shown to be appropriate for natural images (refer to Section 4.3). We modify the original threshold elevation function to make it applicable to the perceptually uniform luma space, which we introduced in Section 5.6.2. Threshold elevation for this space can be approximated by the function:

$$T_e(\bar{L}_{CSF}) = \begin{cases} 1 & \text{if } \bar{L}_{CSF} \leq a \\ c \cdot (\bar{L}_{CSF}))^b & \text{otherwise} \end{cases} \tag{5.18}$$

where $\bar{L}_{CSF}$ is a wavelet coefficient, $a = 0.093071$, $b = 1.0299$ and $c = 11.535$. The function with original data points is plotted in Figure 5.24.

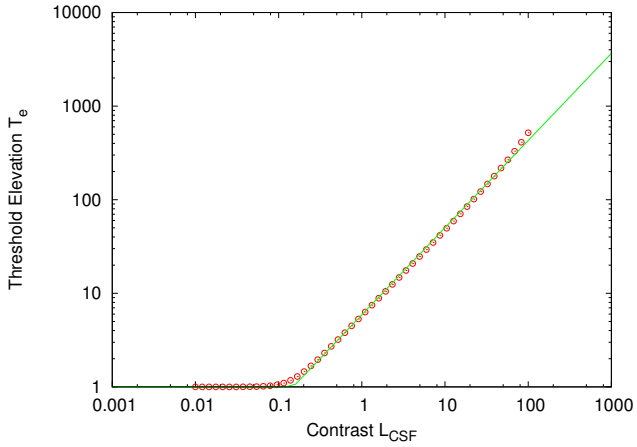


Figure 5.24: The threshold elevation function for contrast represented as wavelet coefficients. Data points were found from the model proposed by Daly [Daly 1993] after conversion to new units of contrast. The solid line is a model from Equation 5.18, which approximates these data points.

Next, we compare each CSF weighted coefficient of a residual frame, R_{CSF}, with the corresponding value of the threshold elevation T_e. If the residual is smaller than the visibility threshold predicted by the threshold elevation function from Equation 5.18, we can safely set this coefficient to zero without introducing visually noticeable changes. Formally, it can be written as:

$$R_{filt} = \begin{cases} 0 & \text{if } T_e(\bar{L}_{CSF}) < R_{CSF} \\ R & \text{otherwise} \end{cases} \tag{5.19}$$

Finally, we transform the filtered wavelets coefficients, R_{filt} back to the image domain ($\mathbf{DWT}^{-1}$ block in Figure 5.23).

The effect of invisible noise filtering on a test image is shown in Figure 5.25. The input image (1) is split into two images, one has luma values of 30% percent of the original (2) and simulates the residual in the compression scheme, the other is 70% of the original (3) and simulates the LDR part. The 30% image (residual) is processed with the invisible noise filter to produce image (4). You can notice that some higher frequencies, especially for the grating patterns bottom left, were completely removed. Despite this, the resulting image, which is the result of summation (3) + (4), does not show any visible artifacts.

The prefiltering method presented above can substantially reduce the size of a residual stream and is a reasonable trade-off between computational efficiency and accuracy of the visual model. The encoding time is affected by no more than 80% when filtering is used and it can only reduce decoding times because of a smaller resulting bit-stream. We have resolved to simplify some aspects of the visual model in order to bring the performance to an acceptable level. For example, we do not model the Optical Transfer Function (OTF) since we found that its local effect is negligible (close or below the

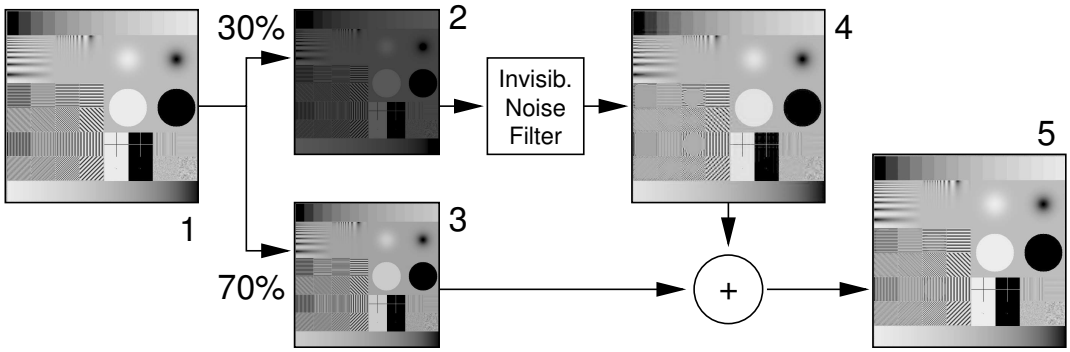


Figure 5.25: Invisible noise filtering applied to a test image. 1 – input image; 2,3 – 30% and 70% of the original image; 4 – 30% image after invisible noise filtering (luminance rescaled to better show removed details); 5 – resulting image. Test image courtesy of Scott Daly.

MTF of a monitor) for typical viewing conditions and the low-frequency flare effect would require much larger kernels or operations in the Fourier domain, which would slow down filtering significantly. For performance reasons we use wavelets, which do not model visual channels as accurately as other transformations designed especially for that purpose. Since we do not have precise information on the optical flow, we can not model temporal aspects of the CSF. The temporal CSF is partially taken into account by the MPEG encoder. Nevertheless, the proposed prefiltering method takes into account more perceptual factors than most state-of-the-art video compression techniques and can additionally handle HDR scenes.

Note that this is not the only possible filtering scheme and some applications may use different filters. For example, sub-sampling and reducing the resolution of residual frames, as done in [Ward and Simmons 2004], can improve both compression efficiency and encoding/decoding speed, but at the cost of blurry artifacts, especially in the regions where LDR pixels have been clamped to minimum or maximum values. If video is to be displayed on a particular type of display, there is no reason to encode the information that can not be displayed. Therefore the filter can take into account the limitations of the display, which are usually more restictive than the full capabilities of the HVS.

5.6.6 Implementation Details

The implementation of MPEG-4 Advanced Simple Profile ISO/IEC 14496-2, available from http://www.xvid.org/, was used as a base MPEG encoder/decoder. However, our method is not restricted to any particular implementation and any other video or image encoder could be used instead. The backward compatible HDR encoder/decoder has been implemented as a dynamic library to simplify integration with external software. We separately implemented a set of command-line tools for encoding and decoding video streams to and from HDR image files and integrated them with the *pfstools* framework (http://pfstools.sourceforge.net/). An LDR stream can be played back using any video player capable of decoding MPEG-4 video. To play back an HDR stream, we have developed a custom HDR video player, which can display video on

both LDR and HDR displays [Seetzen et al. 2004].

Since HDR video playback involves decoding two MPEG-4 streams, an LDR and a residual stream, achieving an acceptable frame rate is more challenging than in the case of an ordinary LDR video. To boost playback frame rate, we moved some parts of the decoding process to graphics hardware. We found that both color space conversion and up-sampling of color channels are computationally expensive when executed on a CPU while the same operations can be performed in almost no time on a GPU as fragment programs. The remaining parts of the decoding and encoding algorithm were implemented using the *SSE* instruction set whenever possible. Additionally, some color conversion functions were significantly accelerated with the use of fixed point arithmetic and lookup tables. All those optimizations let us achieve real-time software playback of HDR movies (25–50 frames per second for the VGA resolution, depending on a hardware configuration and quality settings of the compression).

5.6.7 Results

To test the performance of our backward compatible HDR MPEG compression, we have executed an extensive set of over 1,500 tests on images and video sequences. A good video compression should produce a video stream of the smallest size (measured in our tests as the number of bits per pixel) at the highest quality. Although simple arithmetic metrics, such as Signal to Noise Ratio (SNR), are usually used to measure the quality of compressed images, we follow a common practice in CG [Ward and Simmons 2004, Xu et al. 2005] and also use advanced metrics that account for the aspects of the HVS. We used the following metrics to evaluate the quality of the decoded images and video sequences:

HDR VDP — Visual Difference Predictor for High Dynamic Range images [Mantiuk et al. 2005a]. This is a fidelity metric that can predict the differences between two images that are likely to be noticed by a human observer. This metric has been especially designed for HDR images and takes into account such effects as light scattering in the optics of the eye, luminance masking for the visible range of luminance, spatial contrast sensitivity, local adaptation and visual masking. The result of the HDR VDP is a probability of detection map, which assigns for each pixel a probability that the difference can be noticed. For easier interpretation of the results we have summarized the prediction of the HDR VDP with a single number, which is a percentage of pixels in an image that exceed 75% probability of detection. The lower percentage denotes a better quality, as fewer pixels are noticeably affected by compression distortions. We used the original implementation of the HDR VDP provided by the authors.

UQI — Universal Image Quality Index [Wang and Bovik 2002]. This quality metric models any image distortion as a combination of three factors: loss of correlation, luminance distortion, and contrast distortion. The index, although it does not employ any model of the HVS, shows consistency with a subjective quality measurement and performs better than the mean squared error. The quality index can range from -1 (the worst quality) to 1 (the best quality). We have implemented this metric according to the original paper [Wang and Bovik 2002]. To adapt this metric to HDR images, we provide for input luma values computed with Equation 5.9.

SNR — Signal to Noise Ratio. This is the simplest but also the most commonly used metric, which does not model any aspects of the HVS and may not be consistent with

a subjective quality measurement. We used the standard formulas to compute the SNR for the luma values computed with Equation 5.9. The larger value of SNR usually results in higher quality.

In the following sections we analyze several aspects of our encoding scheme based on the collected test results.

Influence of Tone Mapping Operator

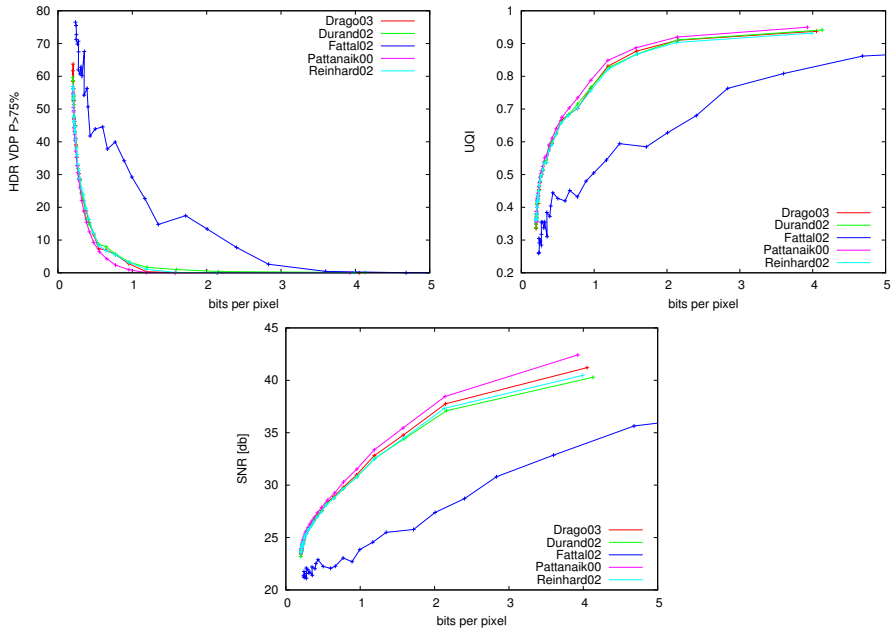


Figure 5.26: Comparison of compression performance for different tone mapping operators. See Section 5.6.7 for the description of the quality metrics. "+" denotes measurement points for a selected image.

Although there are no restrictions on tone mapping / gamut mapping or stylizing used to obtain LDR frames, the choice of such processing will obviously affect the efficiency of compression. We tested our encoder with five tone mapping operators (TMOs) from the *pfstmo* package[9] (labels in italics): *Pattanaik00* — Time-Dependent Visual Adaptation [Pattanaik et al. 2000]; *Durand02* — Fast Bilateral Filtering [Durand and Dorsey 2002b]; *Reinhard02* — Photographic Tone Reproduction [Reinhard et al. 2002b]; *Fattal02* — Gradient Domain [Fattal et al. 2002]; *Drago03* — Adaptive Logarithmic Mapping [Drago et al. 2003]. We used the default parameters for all TMOs. To prevent temporal flickering in tone-mapped video sequences, we added extensions to the original TMOs that ensured time-coherence of the TMO parameters. The extension ensured that the maximum difference of selected parameters (e.g. L_{White} for the *Reinhard02* TMO) between frames is always below the visibility threshold.

[9] More details on the *pfstmo* at: http://www.mpii.mpg.de/resources/tmo/

Figure 5.26 shows how the efficiency of compression is affected by a TMO. The results for most TMOs are in fact similar, with the exception of *Fattal02*, which results in significantly larger streams. This is mainly because the operator introduces the largest changes of local contrast in LDR frames, which results in the high variance of residual values. The result is consistent with our earlier considerations in Section 5.6.3, which suggested that global TMOs are better approximated by the reconstruction function and therefore result in smaller magnitudes of the residual. If *Fattal02* is used to generate LDR video, the size of the LDR stream is also affected since high frequencies, which are poorly compressed by the MPEG encoding, are enhanced (we expect similar problems with the tone mapping approach proposed by Li et al. [Li et al. 2005]). Nevertheless, *Fattal02* gave the most attractive LDR images. Therefore the selection of a proper TMO for compression is often a combined aesthetic and economic choice.

The Effect of Invisible Noise Filtering

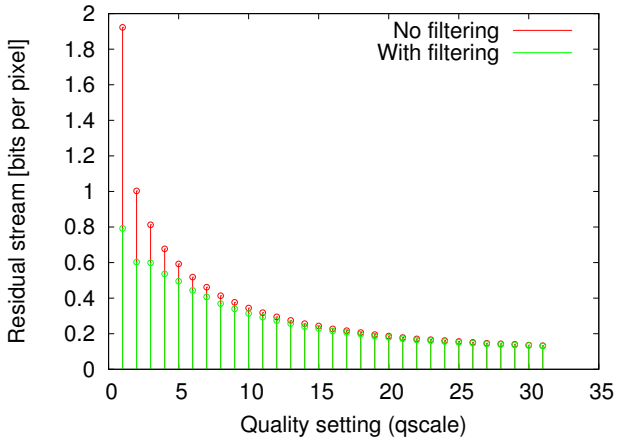


Figure 5.27: The size of a residual stream with and without invisible noise filtering with respect to the quality settings. The largest savings are achieved for the best quality settings.

We validate the algorithm for filtering invisible noise, described in Section 5.6.5, for a range of MPEG quality settings. Figure 5.27 illustrates how the size of a residual stream is reduced when the filtering is used. Note that the largest savings are possible for the best quality settings. This is because the strength of the filtering is determined by the visibility thresholds, which do not depend on quality settings. The filtering has a minimal impact on the stream size for low quality settings since the distortions introduced by the aggressive DCT quantization are far above the visibility thresholds used in the filtering. Figure 5.31 shows how both the total stream size and quality are affected when the residual frames are filtered. Although the filtering in fact introduces changes that are detected by the HDR VDP (probably due to a mismatch in the visual models used by the filtering and the HDR VDP), the loss of quality is fully compensated by the bit-rate savings (see Figure 5.31). Moreover, we observe that the subjective quality of filtered video is better than predicted by the HDR VDP. This is because the blurry artifacts due to the wavelet based filtering are less objectionable than blocky artifacts

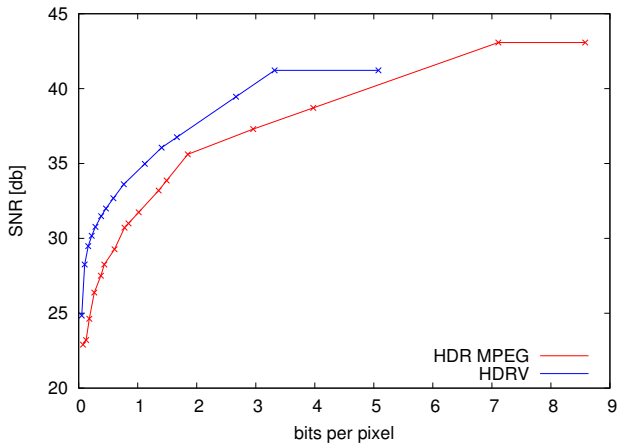


Figure 5.28: Comparison of lossy HDR compression algorithms. Averaged results for two video sequences.

of DCT coding (see Figure 5.30). Although HDR VDP can predict the existence of visible distortions, it can neither estimate their magnitude, nor their impact on perceived quality.

Comparison with Lossy HDR Compression Methods

The performance of the proposed method (labeled as **HDR MPEG**) has been compared with two others lossy HDR compression methods:

HDRV — Perception-motivated HDR Video Encoding [Mantiuk et al. 2004a], described in detail in Section 5.4. This is the first lossy HDR video compression method, which, however, does not offer backward compatibility. The method encodes HDR pixels using 11 bits for luminance and twice 8 bits for chrominance. Since the resulting video stream does not contain any information on LDR frames, it can be expected that this compression method gives better results than backward compatible methods. We used the original implementation provided by the authors.

JPEG HDR — Subband encoding of high dynamic range imagery [Ward and Simmons 2004, Ward and Simmons 2005] introduced in Section 5.5.2. This is a backward compatible HDR image encoding, which is conceptually the closest to our method. A detailed comparison of both our approach and JPEG HDR is given in Section 5.6.8. We used the original encoding/decoding library provided by the authors.

To evaluate the performance of intra-frame (image) compression, we ran the tests on eight representative HDR images. We chose the *Reinhard02* TMO to compare our algorithm with other lossy compression methods. This TMO performed similar to the others and is also used in JPEG HDR. Figure 5.32 shows the averaged results. The HDRV encoding clearly shows the best performance for all three quality metrics. This can be explained by the lack of any information on an LDR stream, which reduces the amount of information that needs to be stored but also makes this encoding incompatible with the LDR MPEG format. For the HDR VDP and the UQI, JPEG HDR performs almost the same as our method for the pre-correction and the post-correction

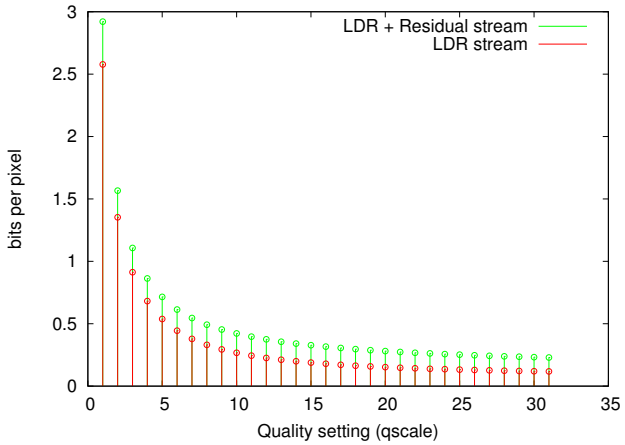


Figure 5.29: The size of a backward compatible HDR stream (LDR+Residual+Auxiliary) compared with the size of the LDR stream alone with respect to the quality settings (low *qscale* denotes high quality) for the *Reinhard02* TMO. Results averaged over a set of images.

approach, but is worse for the full-sampling. Note that our compression method does not involve sub-sampling and therefore is closer to the full-sampling than the other two approaches. JPEG HDR performs worse than our method for the SNR metric. The improved performance of our encoding over JPEG HDR for images is surprising, since the image encoding algorithms, such as JPEG, are known to perform better than intra-frame video encoding. This is due to better arithmetic encoding and a quantization matrix, which is especially optimized for images. Another difference between two methods that affects the performance is that HDR MPEG encodes information on all color channels in the residual stream while the JPEG HDR encodes only luminance in the additional subband layer (see details in Section 5.6.8)

The performance of inter-frame (video) compression was tested on two video sequences for both HDR MPEG and HDRV, while JPEG HDR was not included in these tests. Since both the VDP and the UQI are designed for images and are less suitable for video (large computational cost, lack of temporal processing), we computed the SNR over all video frames to measure quality. The averaged results for two video sequences are shown in Figure 5.28. Similarly as for images, HDRV gave better SNR than HDR MPEG for the same number of bits. HDR MPEG, however, could achieve a higher SNR than HDRV for very high bit-rates.

The Cost of Encoding Residual Stream

The proposed HDR encoding method is designed to be an extension to the existing MPEG formats. Therefore, it is interesting to know how much more data must be stored to include additional HDR information. We plot the size of the total HDR stream (LDR + Residual + Auxiliary stream) against the size of the LDR stream in Figure 5.29. The size of the auxiliary stream is negligible. The residual stream does not seem to depend on the quality settings as much as the LDR stream. Therefore its share in the total stream size is the smallest for high quality settings. This can be expected since the

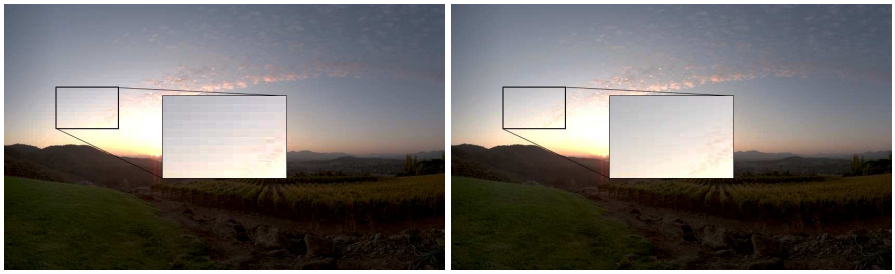


Figure 5.30: Quality comparison for an image compressed without filtering (left) and with invisible noise filtering (right). Both images were compressed to streams of approximately the same size. The strongly visible blocky artifacts in the image compressed without filtering become barely noticeable in the "filtered" image. Note that the artifacts may not be visible in print and should be observed on a gamma-2.2 monitor. No filtering: $qscale = 6$, bpp $= 1.37$, HDR VDP 75%$= 3.12\%$; With Filtering: $qscale = 2$, bpp $= 1.23$, HDR VDP 75%$= 1.11\%$.

residual stream encodes the difference between LDR and HDR frames, including those differences that result from lossy compression of the LDR stream (refer to the MPEG encoding and decoding stages in Figure 5.19). The lower quality LDR stream means that more information needs to be stored in the residual stream. Overall, the share of residual stream ranges from a 5% to 70%, depending on the image, quality settings and a TMO. A well chosen TMO and a decent quality settings result in a residual stream that is 25–30% of the LDR stream.

5.6.8 Discussion

Although the proposed backward compatible HDR encoding algorithm seems to be conceptually similar to the JPEG HDR compression [Ward and Simmons 2004], there are several important differences between the approaches, which not only enable video compression, but also result in better compression and more flexibility of HDR MPEG. As discussed in Section 5.6.3, HDR MPEG can adapt the reconstruction function to the tone/gamut mapping algorithm used to generate LDR frames and therefore reduces the magnitude of the residual values. This results in better compression ratios as compared to JPEG HDR (refer to Section 5.6.7), although the results would be even more favorable if we had used the JPEG algorithm instead of MPEG intra-frame compression to encode images. Further bit-rate savings in MPEG HDR come from perceptually optimized color spaces for HDR pixels (refer to Section 5.6.2).

HDR MPEG offers perceptually conservative and time coherent encoding of residual values, while JPEG HDR suggests an ad-hoc approach to encoding subband, which is not suitable for video. The JPEG HDR encoder transforms subband values to the logarithmic domain and then linearly scales them so that the minimum and the maximum values fit in the 0–255 range. Since the minimum and the maximum subband value can differ from image to image, the scaling factor can also change from frame to frame for video sequences, which would result in temporal flickering and lack of temporal coherence in subband frames. Such a lack of temporal coherence can significantly impact the performance of MPEG inter-frame compression. HDR MPEG, on the

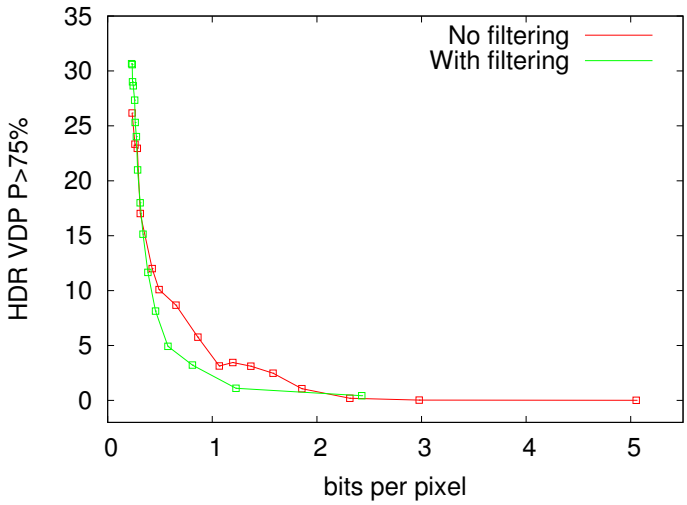


Figure 5.31: Performance of HDR video compression with and without invisible noise filtering (refer to Section 5.6.5). The bit-rate savings gained on the filtering compensate for the loss of quality.

other hand, guarantees the temporal coherence of residual frames. Moreover, the linear scaling of subband values in JPEG HDR makes the quantization of the subband layer difficult to predict and control. JPEG HDR will quantize subband values with high accuracy for those images that lead to small magnitude of subband values, perhaps wasting some bit-rate on invisible contrast details. For another set of images, which result in large magnitude of subband values, JPEG HDR can quantize too coarsely, leading to contouring artifacts. To reduce quantization, JPEG HDR skips a small percent of the brightest and the darkest pixels in an image, which however can lead to loss of some details (see Figure 5.33). HDR MPEG quantizes color values consistently for consecutive frames and the quantizer is based on the visibility thresholds of the HVS rather than frame content.

Unlike JPEG HDR, the proposed compression method does not impose any restrictions on the choice of a TMO and a gamut mapping algorithm. A TMO for HDR MPEG can saturate both luminance and color, change color values and enhance local contrast. Such changes may result in a lower compression ratio, but both LDR and HDR frames will be preserved in the resulting video stream. JPEG HDR will lose most color differences between HDR and LDR since it does not store color in the subband layer. Such unrestricted control over the appearance of both LDR and HDR streams is very important for our major application - a storage format for digital movies whose appearance cannot be compromised.

Finally, a sub-sampling of the subband layer in JPEG HDR may lead to the loss of visible details. Although the *pre-correction* may be used to avoid loss of high frequency details, this leads to distorted LDR frames, which, similar to the companding approach [Li et al. 2005], is not acceptable for applications requiring uncompromised quality of tone mapped images. The *post-correction*, although it does not modify the source image, also does not give as good results as the *pre-correction*. *Full-sampling*, on the other hand, does not give as good compression ratio as the other two approaches.

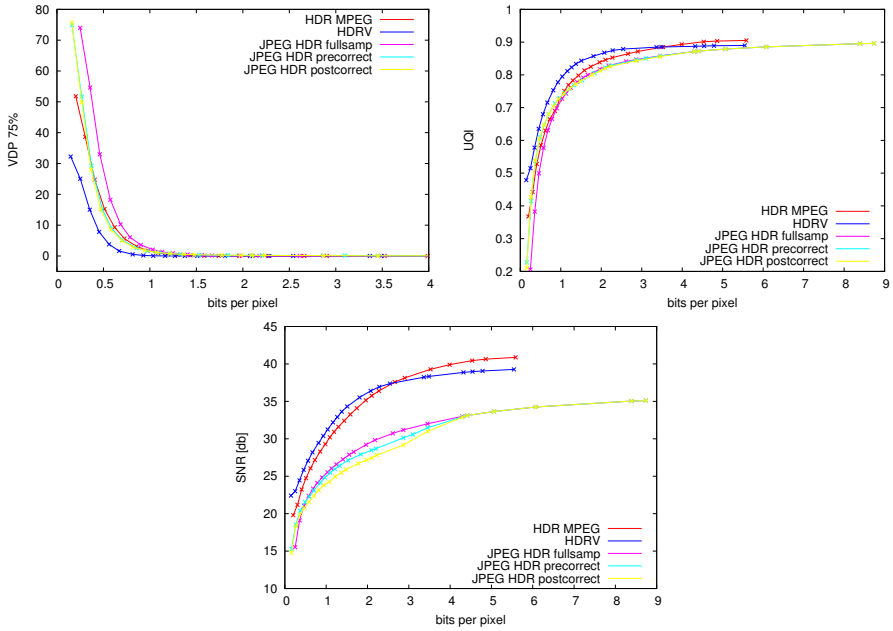


Figure 5.32: Comparison of lossy HDR compression algorithms. Averaged results for a set of images.

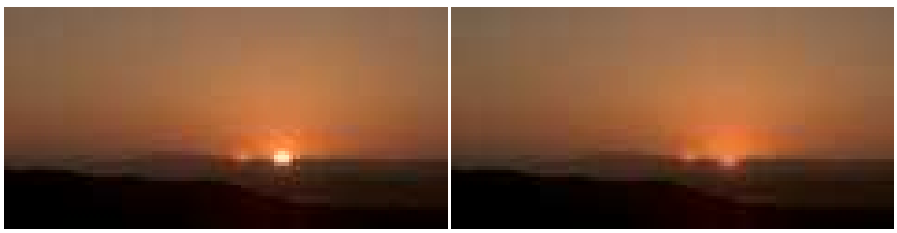


Figure 5.33: Very bright pixels in the original image (left) are lost after compression with JPEG HDR (right) at quality setting 90 and with the *precorrection* (default settings). This is because a small percentage of the brightest and darkest pixels is skipped when computing minimum and maximum value of the subband image.

The counterpart of sub-sampling in the proposed HDR MPEG is filtering of invisible noise (see Section 5.6.5). The filtering has a similar goal as the sub-sampling — to reduce high frequency noise and improve compression, but it does it in a more selective manner. The proposed filtering removes only those high frequency details which are not visible and therefore can be smoothed out without impairing the visual quality of the resulting video. Sub-sampling, obviously, cannot give such a guarantee (refer to Figure 5.22).

5.6.9 Conclusions and Future Work

This section presents the backward compatible HDR MPEG video compression method that can facilitate a smooth transition from LDR to HDR content. The storage cost of a backward compatible HDR stream is modest (about 30% overhead), compared to the huge storage requirement of High Definition video. The proposed format is especially suitable for DVD movie distribution, which must ensure the compatibility with existing DVD players that are not capable of HDR playback. The format design conforms to standard 8-bit MPEG decoding chips. The method allows for separate tone/gamut mapping of LDR and HDR video, which is essential for top quality movie production. We introduced a pair of compatible color spaces that facilitate comparisons between LDR and HDR pixels. The nonlinear function used to encode HDR luminance can be regarded as an extended "gamma correction" that can be used for the full range of visible luminance values. To achieve even better compression performance, we employed an advanced model of the HVS, which is tuned for the full range of visible luminance and is suitable for HDR image processing. We introduced an HDR filtering solution based on this model which selectively and conservatively removes imperceptible high-frequency details from the video stream prior to its compression. We believe that our computationally efficient HVS model and HDR filtering solution are general enough to find other applications in computer graphics and digital imaging.

We implemented and tested a dual video stream encoding for the purpose of a backward compatible HDR encoding, however, we believe that other applications that require encoding multiple streams can partly or fully benefit from the proposed method. For example, a movie could contain a separate video stream for color blind people. Such a stream could be efficiently encoded because of its high correlation with the original color stream. Movie producers commonly target different audiences with different color appearance (for example *Kill Bill 2* was screened with a different color stylization in Japan). The proposed algorithm could be easily extended so that several color stylized movies could be stored on a single DVD. This work is also a step towards an efficient encoding of multiple viewpoint video, required for 3D video [Matusik and Pfister 2004].

Chapter 6

Image Processing in the Contrast Domain

An image stored as a matrix of pixel values is the most common representation for image processing, but unfortunately it does not reflect the way we perceive images. This is why most image compression algorithms apply a transformation, such as the *Discrete Cosine Transform* or the *Discrete Wavelet Transform* before storing images, so that the visually important information is separated from visually less important one. Besides image and video compression, there are many fields that benefit from a representation of images that is correlated with visual perception, such as tone mapping, visual difference prediction, color appearance modeling, or seamless image editing. The goal of such "perceptual" representations is to linearize values that encode images so that the magnitude of those values correspond to the visibility of features in an image. For example, large magnitudes of low and medium frequency coefficients of the Fourier Transform correspond to the fact that the visual system is the most sensitive for those frequencies. In this chapter we derive a framework for image processing in a perceptual domain of image contrast. We base our work on the gradient domain methods, which we generalize and extend to account for perceptual issues, such as the sensitivity for superthreshold contrast in HDR images. This chapter extends work published in [Mantiuk et al. 2005b] and [Mantiuk et al. 2006d]. More information on this project and a gallery of examples can be found at:
http://www.mpi-inf.mpg.de/~mantiuk/contrast_domain/.

6.1 Previous Work

The research on perceptual representation of images has involved many areas of science. We briefly list some of these areas, pointing to the relevant works and describing major issues of these approaches.

Image Transformations. A need for better image representation, which would partly reflect the processing of the Human Visual System (HVS), has been noticed in image processing for a long time. However, practical issues such as whether a transformation is invertible and computational costs, were often of more concern than an accurate

modeling of the HVS. This resulted in numerous image transformations based in mathematics and signal processing, such as the Fourier transform, the discrete cosine transform (DCT), pyramids (Gaussian, Laplacian) or wavelets, which are now considered as standard tools of image processing.

Color Appearance Models. Color appearance models, such CIECAM [CIE 2002] or iCAM [Fairchild and Johnson 2004], convert physical color values to a space of perceptual correlates, such as lightness, chroma and colorfulness. Such correlates are useful for the prediction of color appearance under different visual conditions, for finding visible differences in images and for tone mapping. The drawback of those models is that they usually do not account for aspects of spatial vision such as contrast sensitivity or contrast masking. The reason for this is that the majority of fundamental studies on color appearance have been done with a uniform square patterns on a uniform field without considering spatial or temporal issues, therefore there is not enough data on which spatial models could be build.

Multi-scale Models of Human Vision. Spatial issues are better modelled with multi-scale models, such as those described in [Watson 1987, Simoncelli and Adelson 1989, Watson and Solomon 1997, Pattanaik et al. 1998, Winkler 2005], which separate an image into several band-pass channels. Such channels correspond to the visual pathways that are believed to exist in the HVS. Such models have been successfully applied for the prediction of visible differences in images [Daly 1993] and the simulation of color vision under different luminance adaptation conditions [Pattanaik et al. 1998]. However, they also pose many problems when images are modified in such multi-scale representations. If an image is modified in one of such band-pass limited channels while the other channels remain unchanged, the image resulting from the inverse transformation often contains severe halo artifacts.

Retinex. A different set of problems has been addressed by the Retinex theory of color vision, introduced by Land [Land 1964]. The original goal of Retinex was to model the ability of the HVS to extract reliable information from the world we perceive despite changes in illumination, which is referred as a color constancy. The latter work on the Retinex algorithm formalized the theory mathematically and showed that the problem is equivalent to solving a Poisson equation [Horn 1974, Hurlbert 1986]. Interestingly, most of the gradient methods also involve a solution of a Poisson equation although their goal is different.

Gradient Methods. Operations on image gradients have recently attracted much attention in the fields of tone mapping [Fattal et al. 2002], image editing [Perez et al. 2003, Agarwala et al. 2004], image matting [Sun et al. 2004], image stitching [Levin et al. 2004], and color-to-gray mapping [Gooch et al. 2005]. The gradient methods can produce excellent results in areas where other methods usually result in severe artifacts. For instance, tone mapping and contrast enhancement performed in the gradient domain gives almost no halo artifacts while such artifacts are usually inevitable in the case of the multi-scale methods [Fattal et al. 2002]. The gradients methods can also seamlessly blend stitched images while other methods often result in visible discontinuities [Levin et al. 2004]. Even some advanced painting tools of Adobe Photoshop are based on the gradient methods [Georgiev 2005]. However, all these works focus mainly on image processing aspects without considering perceptual issues. In this work we generalize the gradient domain methods and incorporate perceptual issues by deriving a framework for processing images in perceptually linearized visual response space. Unlike the gradient or multi-scale methods, we impose constraints on the entire

set of contrasts in an image for a full range of spatial frequencies. This way, even a severe image modification does not lead to reversing a polarity of contrast.

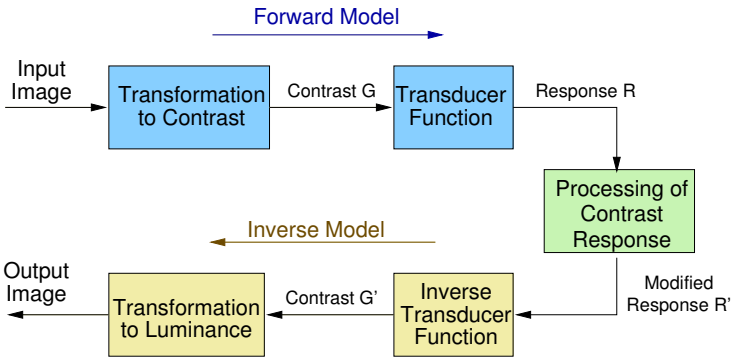


Figure 6.1: Data flow in the proposed framework of the perceptual contrast processing.

The overview of our framework is shown in Figure 6.1. Pixel luminance values of an image are first transformed to physical contrast values, which are then transduced to response values of the HVS. The resulting image is then modified by altering the response values, which are closely related to a subjective impression of contrast. The modified response values can later be converted back to luminance values using an inverse transformation. As an application of our framework we demonstrate two tone mapping methods which can effectively compress dynamic range without losing low-contrast information. We show that a complex contrast compression operation, which preserves textures of small contrast, is reduced to a linear scaling in our visual response space.

In Section 6.2 we review less well known psychophysical data that was measured for high-contrast stimuli. Based on this data we derive a model of suprathreshold contrast discrimination for high contrast images. In Section 6.3 we introduce the components of our framework, in particular a multi-scale representation of low-pass contrast and a transducer function designed for HDR data. As an application of our framework, we propose two tone mapping methods in Sections 6.4 and 6.5, and a saliency preserving color to gray mapping in Section 6.6. Details on how the framework can be implemented efficiently are given in Section 6.7. We discuss strengths and weaknesses of the proposed framework in Section 6.9. Finally, we conclude and suggest future directions in Section 6.10.

6.2 Background

In the following two sections we review some fundamentals of the perception of contrast and summarize the results of a study on the HVS performance in contrast discrimination for HDR images. We use this contrast discrimination characteristic to derive our contrast processing framework.

W – contrast expressed as a Weber fraction (see Table 6.2)
G – contrast expressed as a logarithmic ratio (see Table 6.2)
$\Delta W(W)$, $\Delta G(G)$ – function of threshold contrast discrimination for contrast W and G respectively
$\Delta G_{simpl}(G)$ – simplified function of threshold contrast discrimination for contrast G
$G^k_{i,j}$ – contrast between pixels i and j at the k'th level of a Gaussian pyramid (see Equation 6.6)
$\hat{G}^k_{i,j}$ – modified contrast values, corresponding to $G^k_{i,j}$. Such contrast values usually do not form a valid image and only control an optimization procedure
L^k_i – luminance of the pixel i at the k'th level of a Gaussian pyramid
x^k_i – $\log_{10}$ of luminance L^k_i
Φ_i – set of neighbors of the pixel i
$T(G)$, $T^{-1}(G)$ – transducer and inverse transducer functions
R – response of the HVS scaled in JND units
$\hat{R}$ – modified response R

Table 6.1: Used symbols and notation.

6.2.1 Contrast

The human eye shows outstanding performance when comparing two light patches, yet it almost fails when assessing the absolute level of light. This observation can be confirmed in a ganzfeld, an experimental setup where the entire visual field is uniform. In fact, it is possible to show that the visual system cannot discern mean level variations unless they fluctuate in time or with spatial signals via eye movements, thus having a higher temporal frequency component. The Retinex theory postulated that low sensitivity to absolute luminance can be easily explained by the adaptation of the HVS to the real world conditions. Because the HVS is mostly sensitive to relative luminance ratios (contrast) rather than absolute luminance, the effect of huge light changes over the day is reduced and therefore we perceive the world in a similar way regardless of the light conditions. This and other sources of evidence strongly suggest that the perception of contrast (difference between two light stimuli) is the fundamental ability of the HVS.

Many years of research on contrast have resulted in several definitions of contrast, some of them listed in Table 6.2. The variety of contrast definitions comes from the different stimuli they measure. For example, the Michelson contrast [Michelson 1927] is commonly used to describe a sinusoidal stimulus, while the Weber fraction is often used to measure a step increment or decrement stimulus. In the next section we show that certain contrast definitions are more suitable for describing the performance of the HVS than others.

6.2.2 Contrast Discrimination

Contrast detection and *contrast discrimination* are two of the most thoroughly studied perceptual characteristics of the eye [Barten 1999]. The contrast detection threshold

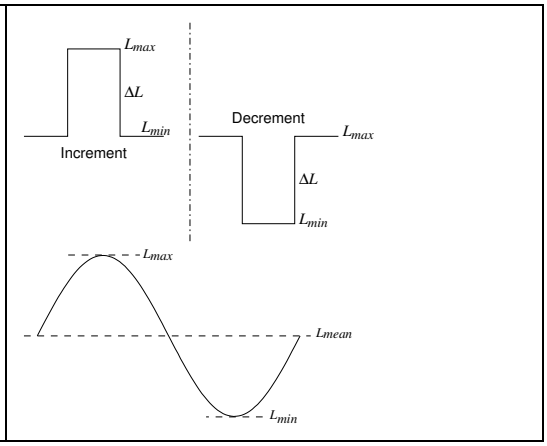
Simple Contrast $C_s = \frac{L_{max}}{L_{min}}$ Weber Fraction $W = \frac{\Delta L}{L_{min}}$ Logarithmic Ratio $G = \log_{10}(\frac{L_{max}}{L_{min}})$ Michelson Contrast $M = \frac{\lvert L_{max} - L_{min} \rvert}{L_{max} + L_{min}}$ Signal to Noise Ratio $SNR = 20 \cdot \log_{10}(\frac{L_{max}}{L_{min}})$	

Table 6.2: Definitions of contrast and the stimuli they measure.

is the smallest visible contrast of a stimulus presented on a uniform field, for example a Gabor patch on a uniform adaptation field. The contrast discrimination threshold is the smallest visible difference between two nearly identical signals, for example two sinusoidal patterns that differ only in their amplitudes. Detection can be considered as a special case of discrimination when the masking signal is uniform (has zero amplitude) and only elevates luminance. For this reason the effect of luminance on the detection threshold is sometimes called *luminance masking*.

A stimulus can be considered *suprathreshold* when its contrast is significantly above the detection threshold. When the contrast is lower or very close to the detection threshold, a stimulus is considered *subthreshold* or *threshold*. Contrast discrimination is associated with the suprathreshold characteristics of the HVS and in particular with *contrast masking*. Contrast detection, on the other hand, describes the performance of the HVS for subthreshold and threshold stimulus, which can be modelled by the *Contrast Sensitivity Function* (CSF), the *threshold versus intensity* function (t.v.i.), or Weber's law for luminance thresholds. A more detailed discussion on threshold and suprathreshold effects can be found in Section 3.9 and in [Wandell 1995, Chapter 7].

Since suprathreshold contrast plays a dominant role in the perception of HDR images, we will consider contrast discrimination data (suprathreshold) in detail and simplify the character of contrast detection (threshold). Although discrimination thresholds of the HVS have been thoroughly studied in psychophysics for years, most of the measurements consider only small contrast levels up to $M = 50\%$. Such limited contrast makes the usefulness of the data especially questionable in the case of HDR images, for which the contrast can easily exceed 50%. The problem of insufficient scale of contrast in psychophysical experiments was addressed by Whittle [Whittle 1986]. By measuring detection thresholds for the full range of visible contrast, Whittle showed that the discrimination data plotted with the Michelson contrast does not follow increasing slope, as reported in other studies (refer to Figure 6.2). He also argued that the Michelson contrast does not describe the data well. Figure 6.2 shows that the data is very scattered and the character of the threshold contrast is not clear, especially for large contrast values. However, when the same data is plotted as Weber's fraction $W = \Delta L / L_{min}$, the discrimination thresholds for all but the smallest contrast

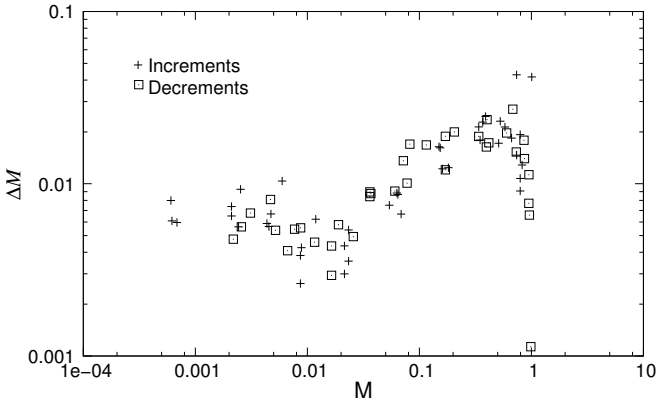


Figure 6.2: Contrast discrimination thresholds plotted using the Michelson contrast, M. The Michelson contrast does not give a good prediction of the discrimination performance, especially for high contrast.

values follow the same line on a log-log plot, which resembles Weber's law, but for suprathreshold contrast: $\Delta W/W = c$ (see Figure 6.3). The sensitivity[1] to contrast improves for low contrast just above the detection threshold and then deteriorates as the contrast reaches the threshold ($W \approx 0.025$). Whittle calls this effect "crispening" while contrast discrimination studies usually describe it as a facilitation or "dipper" effect.

Interestingly, typical models of contrast discrimination, such as Barten's model [Barten 1999, Chapter 7], closely follow Whittle's data for low contrast[2], but wrongly predict discrimination thresholds for high contrast (see the green solid line in Figure 6.3). The wrong prediction is a result of missing measurements for high contrast. Obviously, such models are not adequate for high contrast data, such as HDR images.

To construct a model for contrast discrimination, which would be suitable for High Dynamic Range images, we fit a continuous function to Whittle's original data [Whittle 1986, Figure 2]:

$$\Delta W(W) = 0.0928 \cdot W^{1.08} + 0.0046 \cdot W^{-0.183} \tag{6.1}$$

The *chi-square* test proves that the function approximates the data ($Q = 0.56$) assuming a relative error $\Delta W/W \pm 8\%$. The shape of the fitted function is shown as a red solid line in Figure 6.3. In Section 6.3.2 we use the above function rather than Whittle's original model $\Delta W/W = c$ to properly predict discrimination thresholds for low contrast values.

It is sometimes desirable to operate on contrast measure G rather than Weber fraction W (for contrast definitions refer to Table 6.2). In Section 6.3.1 we show that the proposed framework operates on contrast G since such contrast can be represented as a difference in logarithmic domain, which let us formulate a linear problem. Knowing that the relation between W and G is:

$$G = log_{10}(W + 1) \tag{6.2}$$

[1] Sensitivity is defined as an inverse of the detection or discrimination threshold.

[2] The parameters for Barten's model have been chosen to fit the measurements by Foley and Legge [Foley and Legge 1981]. The detection threshold m_t has been chosen so that it compensates for differences between the stimuli used for Whittle's and Legge & Foley's measurements.

and the relation between ΔW and ΔG is:

$$\Delta G \approx log_{10}(W + \Delta W + 1) - log_{10}(W + 1) = log_{10}\left(\frac{\Delta W}{W+1} + 1\right), \tag{6.3}$$

we plot Whittle's measurement points for contrast G in Figure 6.4. We can now fit the model from Equation 6.1 to the new data, to get a function of contrast discrimination for contrast G:

$$\Delta G(G) = 0.0405 \cdot G^{0.6628} + 0.00042435 \cdot G^{-0.38072} \tag{6.4}$$

The *chi-square* test for the fitted function gave $Q = 0.86$ assuming a relative error on $\Delta G/G \pm 7\%$. If we do not need to model the facilitation effect or the loss of sensitivity for low contrast, we can approximate the data with a simpler function, which is both reversible and integrable, but does not consider data for $G < 0.03$:

$$\Delta G_{simpl}(G) = 0.038737 \cdot G^{0.537756} \tag{6.5}$$

The *chi-square* test for the fitted function gave $Q = 0.88$ assuming a relative error $\Delta G/G \pm 3\%$. Both fitted functions are shown in Figure 6.4.

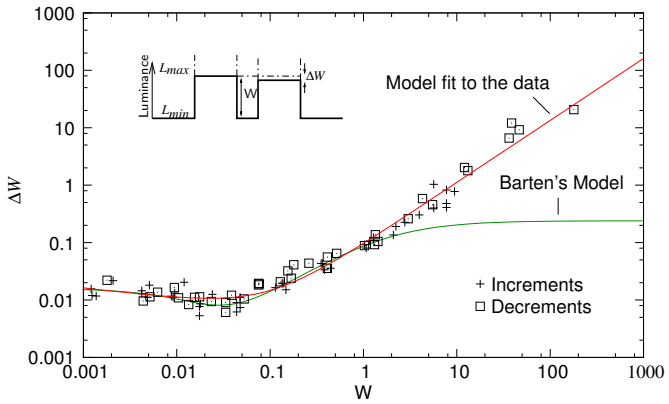


Figure 6.3: Contrast discrimination thresholds plotted as a function of contrast W. Data points – Whittle's measurements; red solid line – a function fit to Whittle's data; green solid line – Barten's model fit to the measurements by Foley and Legge [Foley and Legge 1981] ($k = 3$, $m_t = 0.02$); inset – the stimulus used to measure increments for Whittle's data.

Before we utilize the above discrimination functions, we have to consider whether it can be generalized for different stimuli and spatial frequencies. In a later study Kingdom and Whittle [Kingdom and Whittle 1996] showed that the character of the suprathreshold discrimination is similar for both a square-wave and sine-wave patterns of different spatial frequencies. This is consistent with other studies that show little variations of suprathreshold contrast across spatial frequencies [Georgeson and Sullivan 1975, Barten 1999]. Those variations can be eliminated if a contrast detection function is normalized by the contrast detection threshold for a particular spatial frequency [Legge 1979].

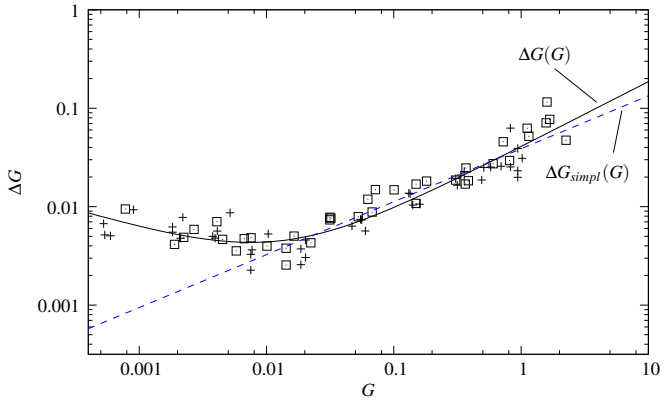


Figure 6.4: Contrast discrimination thresholds plotted as a function of the contrast G. The solid line – a full contrast discrimination model (Equation 6.4); the dashed line – a simplified contrast discrimination model (Equation 6.5).

6.3 A Framework for Perceptual Contrast Processing

In the next two sections we introduce a framework for image processing in a visual response space. Section 6.3.1 proposes a method for transforming complex images from luminance to physical contrast domain (blocks *Transform to Contrast* and *Transform to Luminance* in Figure 6.1). Section 6.3.2 explains how physical contrast can be converted into a response of the HVS, which is a perceptually linearized measure of contrast (blocks *Transducer Function* and *Inverse Transducer Function* in Figure 6.1).

6.3.1 Contrast in Complex Images

Before we introduce contrast in complex images, let us consider the performance of the eye during discrimination of spatially distant patches. We can easily observe that contrast can be assessed only locally for a particular spatial frequency. We can, for example, easily see the difference between fine details if they are close to each other, but we have difficulty distinguishing the brighter detail from the darker if they are distant in our field of view. On the other hand, we can easily compare distant light patches if they are large enough. This observation can be explained by the structure of the retina, in which the foveal region responsible for the vision of fine details spans only about 1.7 visual degrees, while the parafoveal vision can span over 160 visual degrees, but has almost no ability to process high frequency information [Wandell 1995]. When seeing fine details in an image, we fixate on a particular part of that image and employ the foveal vision. But at the same time the areas further apart from the fixation point can only be seen by the parafoveal vision, which can not discern high frequency patterns. The contrast discrimination for spatial patterns with increasing separation follows Weber's law when the eye is fixed to one of the patterns and this is the result of the increasing eccentricity of the other pattern [Wilson 1980]. Therefore, due to the structure of the retina, the distance at which we can correctly assess contrast is small for high frequency signals, but grows for low frequency signals.

While several contrast definitions have been proposed in the literature (refer to Ta-

ble 6.2), they are usually applicable only to a simple stimulus and do not specify how to measure contrast in complex scenes. This issue was addressed by Peli [Peli 1990] who noticed that the processing of images is neither periodic nor local and therefore the representation of contrast in images should be quasi-local as well. Drawing analogy from the center-surround structures in the retina, he proposed to measure contrast in complex images as a difference between selected levels of a Gaussian pyramid. However, the resulting difference of Gaussians leads to a band-pass limited measure of contrast, which tends to introduce halo artifacts at sharp edges when it is modified. To avoid this problem, we introduce a low-pass measure of contrast. We use a logarithmic ratio G as the measure of contrast between two pixels, which is convenient in computations since it can be replaced with the difference of logarithms. Therefore, our low-pass contrast is defined as a difference between a pixel and one of its neighbors at a particular level, k, of a Gaussian pyramid, which can be written as:

$$G^k_{i,j} = log_{10}(L^k_i / L^k_j) = x^k_i - x^k_j \tag{6.6}$$

where L^k_i and L^k_j are luminance values for neighboring pixels i and j. For a single pixel i there are two or more contrast measures $G^k_{i,j}$, depending on how many neighbouring pixels j are considered (see Figure 6.5). Note that both L and x cover a larger and larger area of an image when moving to the coarser levels of the pyramid. This way our contrast definition takes into account the quasi-local perception of contrast, in which fine details are seen only locally, while variations in low frequencies can be assessed for the entire image. The choice of how many neighboring pixels, x_j, should be taken into account for each pixel, x_i, usually depends on the application and type of images. For tone mapping operations on complex images, we found that two nearest neighbors are sufficient. For other applications, such as a color-to-gray mapping, and for images that contain flat areas (for example vector maps), we consider 20–30 neighboring pixels.

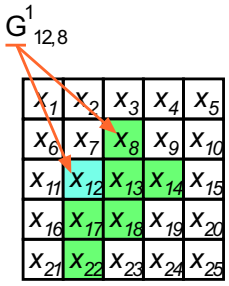


Figure 6.5: Contrast values for the pixel x_{12} (blue) at a single level of the Gaussian pyramid. The neighboring pixels x_j are marked with green color. Note that the contrast value $G_{12,7}$ (upper index $k = 1$ omitted for clarity) will not be computed, since $G_{7,12}$ already contains the same difference. Contrast values $G_{12,8}$ and $G_{12,18}$ encode contrast for diagonal orientations. Unlike wavelets, contrast values, $G^k_{i,j}$ can represent both $-45°$ and $45°$ orientation.

Equation 6.6 can be used to transform luminance to contrast. Now we would like to perform the inverse operation that restores an image from the modified contrast values $\hat{G}$. The problem is that there is probably no image that would match such contrast values. Therefore, we look instead for an image whose contrast values are close but not necessarily exactly equal to $\hat{G}$. This can be achieved by the minimization of the

distance between a set of contrast values $\hat{G}$ that specifies the desired contrast, and G, which is the contrast of the actual image. This can be formally written as the minimization of the objective function:

$$f(x_1^1, x_2^1, \ldots, x_N^1) = \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j \in \Phi_i} p_{i,j}^k (G_{i,j}^k - \hat{G}_{i,j}^k)^2 \tag{6.7}$$

with regard to the pixel values x_i^1 on the finest level of the pyramid. Φ_i is a set of the neighbors of the pixel i (e.g. set of green pixels in Figure 6.5), N is the total number of pixels and K is the number of levels in a Gaussian pyramid. We describe an efficient solution of the above minimization problem in Section 6.7.

The coefficient $p_{i,j}^k$ in Equation 6.7 is a constant weighting factor, which can be used to control a mismatch between the desired contrast and the contrast resulting from the solution of the optimization problem. If the value of this coefficient is high, there is higher penalty for a mismatch between $G_{i,j}^k$ and $\hat{G}_{i,j}^k$. Although the choice of these coefficients may depend on the application, in most cases we want to penalize contrast mismatch relative to the contrast sensitivity of the HVS. A bigger mismatch should be allowed for the contrast magnitudes to which the eye is less sensitive. This way, the visibility of errors resulting from such a mismatch would be equal for all contrast values. We can achieve this by assuming that:

$$p_{i,j}^k = \begin{cases} \Delta G^{-1}(\hat{G}_{i,j}^k) & \text{if } \hat{G}_{i,j}^k \geq 0.001 \\ \Delta G^{-1}(0.001) & \text{otherwise,} \end{cases} \tag{6.8}$$

where ΔG^{-1} is an inverse of the contrast discrimination function from Equation 6.4 and the second condition avoids division by 0 for very low contrast.

When testing the framework with different image processing operations, we noticed that the solution of the optimization problem may lead to reversing polarity of contrast values in an output image, which happens when $G_{i,j}^k$ is of a different sign than $\hat{G}_{i,j}^k$, and which leads to halo artifacts. This problem concerns all methods that involve a solution of the optimization problem similar to the one given in Equation 6.7 and is especially evident for the gradients domain method (based on Poisson solvers). The problem is illustrated in Figure 6.6. To simplify the notation, the upper index of a Gaussian pyramid level is assumed to be 1 and is omitted. A set of desired contrast values $\hat{G}$ quite often contains the values that cannot lead to any valid pixel values (6.6a). The solution of the optimization problem results in modified contrast values G that can be used to construct an image with pixel values x_1, x_2, x_3 (6.6b). The problem is that this solution results in a reversed polarity of contrast ($G_{3,1}$ in 6.6b), which leads to small magnitude, but noticeable, halo artifacts. More desirable would be solution (6.6c), which gives the same value of the objective function f and does not result in reverse contrast values. To increase probability that the optimization procedure results in solution (6.6c) rather than (6.6b), the objective function should be penalized for mismatches at low contrast. This can be combined together with penalizing mismatches according to the sensitivity of the HVS if we replace the contrast discrimination function ΔG in Equation 6.8 with the simplified model ΔG_{simpl} from Equation 6.5:

$$p_{i,j}^k = \frac{1}{\Delta G_{simpl}(\hat{G}_{i,j}^k)} \tag{6.9}$$

The simplified model overestimates sensitivity for low contrast, which is desirable as it makes the value of $p^k_{i,j}$ large near zero contrast and thus prevents the reversal of contrast polarity.

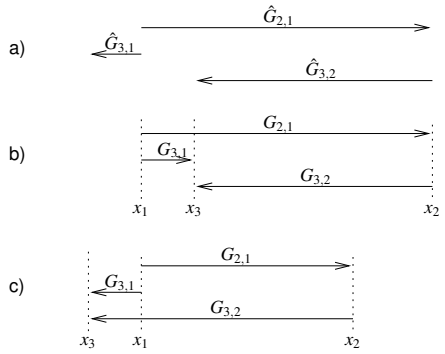


Figure 6.6: For a set of desired contrast values $\hat{G}_{i,j}$ that cannot represent a valid image (a), the optimization procedure may find a solution which contains reversed contrast values ($G_{3,1}$ in b). An alternative solution without such reversed contrast values gives images without halo artifacts (c).

6.3.2 Transducer Function

A transducer function predicts the hypothetical response of the HVS for a given physical contrast. As can be seen in Figure 6.1, our framework assumes that the image processing is done on the response rather than on the physical contrast. This is because the response closely corresponds to the subjective impression of contrast and therefore any processing operations can assume the same visual importance of the response regardless of its actual value. In this section we would like to derive a transducer function that would predict the response of the HVS for the full range of contrast, which is essential for HDR images.

Following [Wilson 1980] we derive the transducer function $T(G) := R$ based on the assumption that the value of the response R should change by one unit for each Just Noticeable Difference (JND) both for threshold and suprathreshold stimuli. However, to simplify the case of threshold stimuli, we assume that:

$$T(0) = 0 \text{ and } T(G_{threshold}) = 1 \tag{6.10}$$

or

$$T^{-1}(0) = 0 \text{ and } T^{-1}(1) = G_{threshold} \tag{6.11}$$

for the inverse transducer function $T^{-1}(R) := G$. The detection threshold, $G_{threshold}$, is approximated with 1% contrast ($G_{threshold} = log_{10}(0.01+1) \approx 0.0043214$), commonly used for digital images [Wyszecki and Stiles 2000, Section 7.10.1]. This simplification assumes that the detection threshold is the same for all spatial frequencies and all luminance adaptation conditions. For a suprathreshold stimulus we approximate the response function T by its first derivative:

$$\Delta T \approx \frac{dT(G)}{dG} \Delta G(G) = 1 \tag{6.12}$$

where $\Delta G(G)$ is the discrimination threshold given by Equation 6.4. The above equation states that a unit increase of response R (right hand side of the equation) should correspond to the increase of G equal to the discrimination threshold ΔG for the contrast G (left side of the equation). The construction of the function $R = T(G)$ is illustrated in the inset of Figure 6.7. Although the above equation can be solved by integrating its differential part, it is more convenient to solve numerically the equivalent differential equation:

$$\frac{dT^{-1}(R)}{dR} = \Delta G(T^{-1}(R)) \tag{6.13}$$

for the inverse response function $T^{-1}(R) = G$ and for the boundary condition from Equation 6.11. G is a non-negative logarithmic ratio (refer to Table 6.2) and R is the response of the HVS. Since the function T^{-1} is strictly monotonic, finding the function T is straightforward. We numerically solve Equation 6.13 to find the transducer function $T(G) = R$ shown in Figure 6.7.

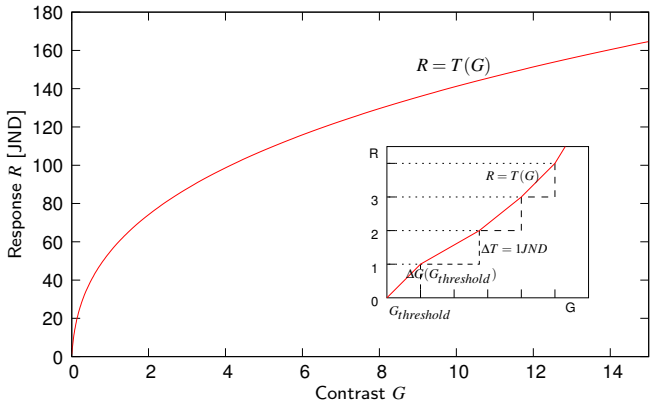


Figure 6.7: Transducer function derived from the contrast discrimination data [Whittle 1986]. The transducer function can predict the response of the HVS for the full range of contrast. The inset depicts how the transducer function is constructed from the contrast discrimination thresholds $\Delta G(G)$.

For many applications an analytical model of a transducer function is more useful than a lookup table given by the numerical solution of Equation 6.13. Although the curve shown in Figure 6.7 closely resembles a logarithmic or exponential function, neither of these two families of functions give an exact fit to the data. However, if an accurate model is not necessary, the transducer can be approximated with the function:

$$T(G) = 54.09288 \cdot G^{0.41850} \tag{6.14}$$

The average and maximum error of this approximation is respectively $R \pm 1.9$ and $R \pm 6$. Equation 6.14 leads directly to an inverse transducer function:

$$T^{-1}(R) = 7.2232 \cdot 10^{-5} \cdot R^{2.3895}. \tag{6.15}$$

The transducer function derived in this section has a similar derivation and purpose as the Standard Grayscale Function from the DICOM standard [DICOM PS 3-2004 2004] or the capacity function in [Ashikhmin 2002]. The major difference is that the

transducer function operates in the contrast domain rather than in the luminance domain. It is also different from other transducer functions proposed in the literature (e.g. [Wilson 1980, Watson and Solomon 1997]) since it is based on the discrimination data for high contrast and operates on contrast measure G. This makes the proposed formulation of the transducer function especially suitable to HDR data. The derived function also simplifies the case of the threshold stimuli and assumes a single detection threshold $G_{threshold}$. Such a simplification is acceptable, since our framework focuses on suprathreshold rather than threshold stimuli.

6.4 Application: Contrast Mapping

In previous sections we introduce our framework for converting images to perceptually linearized contrast response and then restoring images from the modified response. In this section we show that one potential application of this framework is to compress the dynamic range of HDR images to fit into the contrast reproduction capabilities of display devices. We call this method contrast mapping instead of tone mapping because it operates on contrast response rather than luminance.

Tone mapping algorithms try to overcome either the problem of the insufficient dynamic range of a display device (e.g. [Tumblin and Turk 1999, Reinhard et al. 2002a, Durand and Dorsey 2002b, Fattal et al. 2002]) or the proper reproduction of real-world luminance on a display (e.g. [Pattanaik et al. 1998, Ashikhmin 2002]). Our method does not address the second issue of trying to make images look realistic and natural. Instead we try to fit to the dynamic range of the display so that no information is lost due to saturation of luminance values and at the same time, small contrast details, such as textures, are preserved. Within our framework such non-trivial contrast compression operation is reduced to a linear scaling in the visual response space. Since the response $R^k_{i,j}$ is perceptually linearized, contrast reduction can be achieved by multiplying the response values by a constant l:

$$\hat{R}^k_{i,j} = R^k_{i,j} \cdot l \tag{6.16}$$

where l is between 0 and 1. This corresponds to lowering the maximum contrast that can be achieved by the destination display. Since the contrast response R is perceptually linearized, scaling effectively enhances low physical contrast W, for which we are the most sensitive, and compresses large contrast magnitudes, for which the sensitivity is much lower. The result of such contrast compression for the Memorial Church image is shown in Figure 6.8.

In many aspects the contrast compression scheme resembles the gradient domain method proposed by Fattal et al. [2002]. However, unlike the gradient method, which proposes somewhat ad-hoc choice of the compression function, our method is entirely based on the perceptual characteristic of the eye. Additionally, our method can avoid low frequency artifacts as discussed in Section 6.9.

We tested our contrast mapping method on an extensive set of HDR images. The only visible problem was the magnification of the camera noise on several HDR photographs. Those pictures were most likely taken in low light conditions and therefore their noise level was higher than in the case of most HDR photographs. Our tone mapping method is likely to magnify camera noise if its amplitude exceeds the threshold

Figure 6.8: The results of the contrast mapping algorithm. The images from left to right and top to bottom were processed with the compression factor $l = 0.1, 0.4, 0.7, 1.0$. After the processing images were rescaled in the log_{10} domain to use the entire available dynamic range. *Memorial Church image courtesy of Paul Debevec*.

contrast $W_{threshold}$ of the HVS. Therefore, to obtain good results, the noise should be removed from images prior to the contrast mapping.

In Figure 6.10 we compare the results of our method with other tone mapping algorithms. Our contrast mapping method produces very sharp images without introducing halo artifacts. Sharpening is especially pronounced when the generated images are compared to the result of linear scaling in the logarithmic domain (see Figure 6.11).

6.5 Application: Contrast Equalization

Figure 6.9: Top left – the linear rescaling of luminance in the logarithmic domain; top right – contrast mapping; bottom left – contrast equalization; bottom right – the result of [Reinhard et al. 2002a]. *Image courtesy of Grzegorz Krawczyk.*

Histogram equalization is another common method to cope with extended dynamic range. Even if high contrast occupies only a small portion of an image, it is usually responsible for large dynamic range. The motivation for equalizing the histogram of contrast is to allocate dynamic range for each contrast level relative to the space it occupies in an image. To equalize a histogram of contrast responses, we first find the Cumulative Probability Distribution Function (CPDF) for all contrast response values in the image $R^k_{i,j}$ [Gonzalez and Woods 2001, Section 3]. Then, we calculate the modified response values:

$$\hat{R}^k_{i,j} = sign(R^k_{i,j}) \cdot CPDF(\|R^k_i\|) \tag{6.17}$$

where $sign()$ equals -1 or 1 depending on the sign of the argument and $\|R^k_i\|$ is a

Figure 6.10: Comparison of the result produced by our contrast mapping (top left) and contrast equalization (top right) to those of Durand and Dorsey [2002b] (bottom left) and Fattal et al. [2002] (bottom right). *Tahoma image courtesy of Greg Ward*.

root-mean-square of the contrast response between a pixel and all its neighbors:

$$\|R_i^k\| = \sqrt{\sum_{j \in \Phi_i} {R_{i,j}^k}^2} \tag{6.18}$$

The histogram equalization scheme produces very sharp and visually appealing images, which may however be less natural in appearance than the results of our previous method (see some examples in Figures 6.9, 6.10, and 6.11). Such a tone mapping method can be especially useful in those applications, where the visibility of small details is paramount. For example, it could be used to reveal barely visible details in forensic photographs or to improve the visibility of small objects in satellite images.

The results of the contrast equalization algorithm may appear like the effect of a sharpening filter. Figure 6.12 shows that the result of the contrast equalization (b) results in an image of much better contrast than the original image (a) while preserving low frequency global contrast. Sharpening filters tend to enhance local contrast at the cost of global contrast, which results in images that have flat appearance (c,d). Sharpening filters also introduce ringing and halo artifacts, especially in the areas of high local contrast, such as the border of the window in Figure 6.12 (c,d). The results of the contrast equalization algorithm are free of these artifacts.

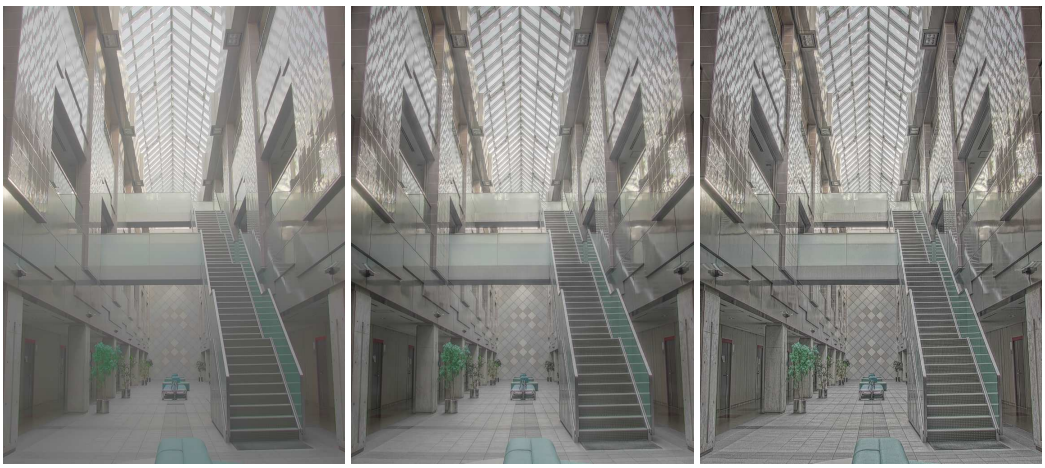


Figure 6.11: The linear rescaling of luminance in the logarithmic domain (left) compared with two proposed contrast compression methods: contrast mapping (middle) and contrast equalization (right).

6.6 Application: Color to Gray

Color images can often lose important information when printed in grayscale. Take for example Figure 6.14, where the sun disappears from the sky when only luminance is computed from the color image. The problem of proper mapping from color to grayscale has been addressed in numerous works, recently in [Gooch et al. 2005, Rasche et al. 2005]. We implemented the approach of Gooch et al. [2005] since their solution can be easily formulated within our framework. Their algorithm separately computes luminance and chrominance differences in a perceptually uniform CIE $L^*a^*b^*$ color space for low-dynamic range. Such differences correspond to contrast values, $G^1_{i,j}$, in our framework (the finest level of a Gaussian pyramid). To avoid artifacts in flat areas (more on this in Section 6.9), their algorithm computes differences between all pixels in the image, which is equivalent to considering for each pixel, x_i, all remaining pixels in the image as neighbors, x_j. Next, each luminance difference that is smaller than the corresponding chrominance difference is replaced with that chrominance difference. The algorithm additionally introduces parameters that control polarity of the chrominance difference, and the amount of chromatic variation applied to the luminance values. Finally, they formulate an optimization problem that is equivalent to Equation 6.7 restricted to the finest level of a pyramid ($k = 1$). The result of the optimization gives a gray-scale image that preserves color saliency. The authors show that their method produces results without artifacts for a broad range of images.

The algorithm, while giving excellent results, is prohibitively computationally expensive and feasible only for very small images. This is because it computes differences (contrast values) between all pixels in an image, what gives a minimum complexity of $O(N^2)$ regardless of the optimization method used. The number of considered differences can be limited, however at the cost of possible artifacts in isoluminant regions. Our framework involves a more efficient approach, in which the close neighborhood of a pixel is considered on fine levels of a Gaussian pyramid while far neighborhood is covered on coarser levels. This let us work with much bigger images and perform computations much faster.

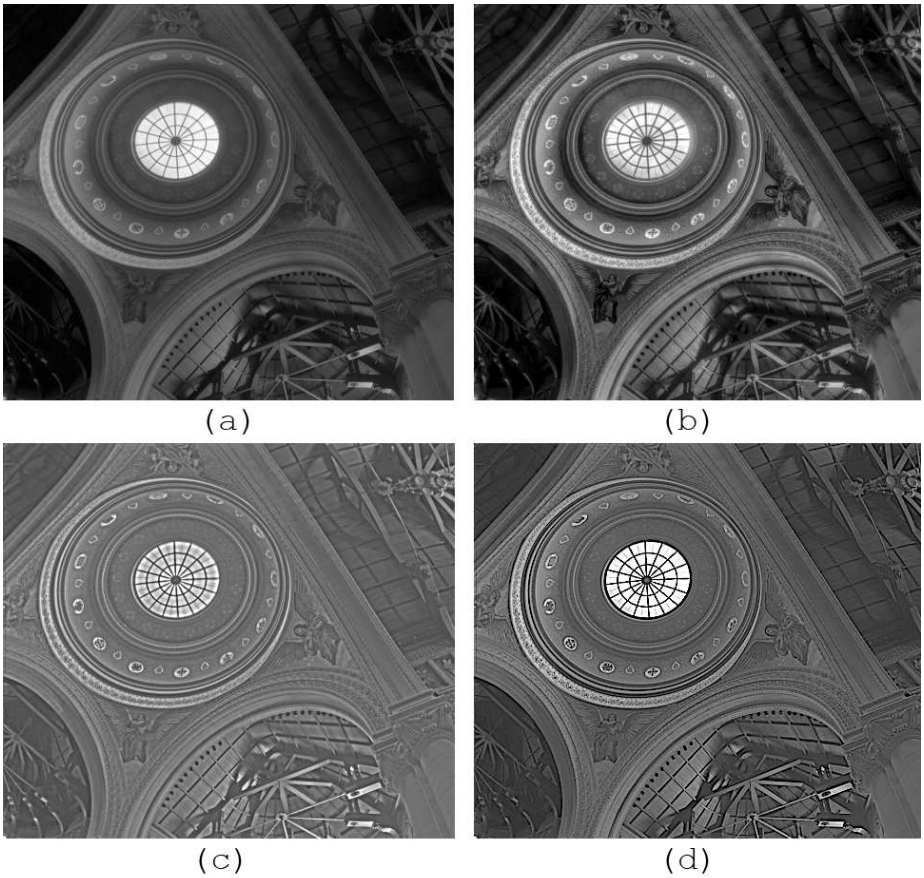


Figure 6.12: The contrast equalization algorithm compared with sharpening filters. (a) the original image; (b) the result of contrast equalization; (c) the result of a 'local adaptation' sharpening; (d) the result of a sharpening filter.

Following [Gooch et al. 2005] we transform input images into a CIE $L^*a^*b^*$ color space. Then, we transform each color channel into a pyramid of contrast values using Equation 6.6 (but x_i^k denotes now the values in color channels). Next, we compute the color difference:

$$\|\Delta C_{i,j}^k\| = \sqrt{(G(a^*)_{i,j}^k)^2 + (G(b^*)_{i,j}^k)^2} \tag{6.19}$$

and selectively replace $G(L^*)_{i,j}^k$ with a signed $\|\Delta C_{i,j}^k\|$, like in [Gooch et al. 2005]. We consider difference values for each level of a Gaussian pyramid and for 20–30 neighboring pixels. There is no need to apply the transducer function to the data. The reconstructed images can be seen in Figures 6.13 and 6.14. We achieve images of similar quality as [Gooch et al. 2005], but at a significantly lower computational cost.

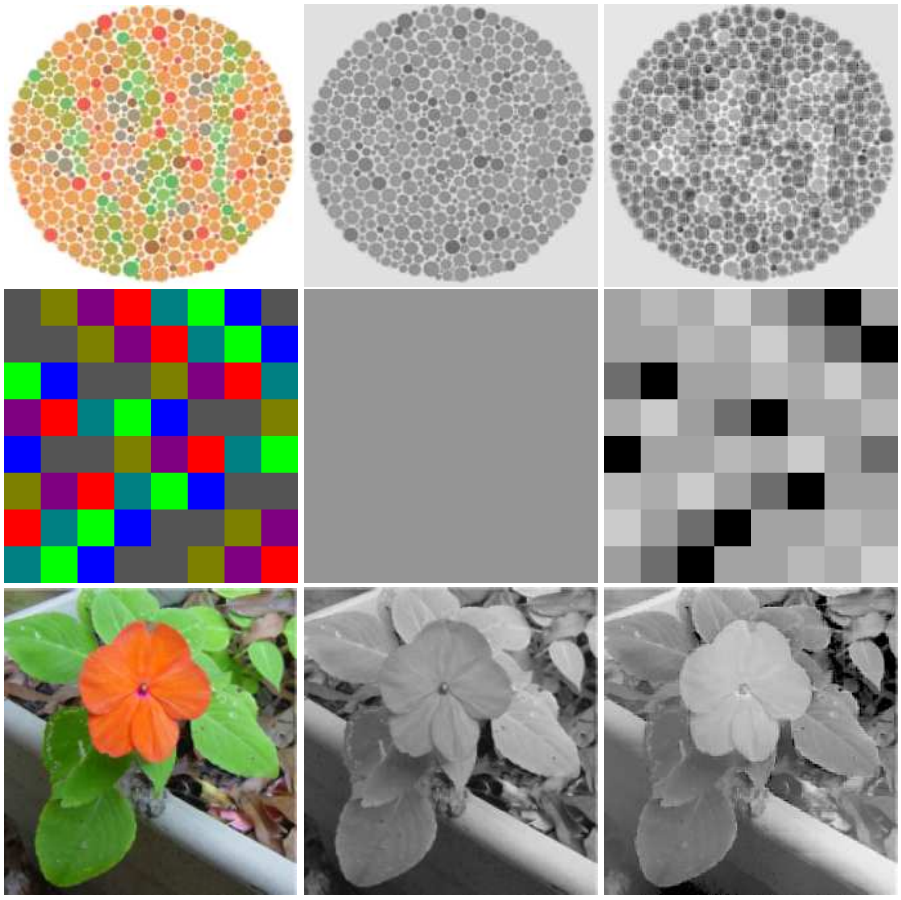


Figure 6.13: Examples of a saliency preserving color to gray mapping. Left – original image; center – luminance image; right – the result of the color to gray algorithm. *Images courtesy of Jay Neitz(top) and Karl Rasche(bottom)*

6.7 Image Reconstruction from Contrast

In this section we give an efficient solution to the optimization problem stated in Section 6.3.1. By solving the optimization problem, we can reconstruct an output image from modified contrast values.

The major computational burden of our method lies in minimizing the objective function given in Equation 6.7. The objective function reaches its minimum when all its derivatives $\frac{\partial f}{\partial x_i}$ equal 0:

$$\frac{\partial f}{\partial x_i} = \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j \in \Phi_i} 2p_{i,j}^k (x_i^k - x_j^k - \hat{G}_{i,j}^k) = 0 \tag{6.20}$$

for $i = 1, \ldots, N$. The above set of equations can be rewritten using a matrix notation:

$$A \cdot X = B \tag{6.21}$$

Figure 6.14: An example of a saliency preserving color to gray mapping. Left – original image; center – luminance image; right – the result of the color to gray algorithm. *Image: Impressionist Sunrise by Claude Monet*

where X is a column vector of $x_1, \ldots, x_N$, which holds pixel values of the resulting image, A is an $N \times N$ square matrix and B is an N-row vector. For a few mega-pixel images N can equal several million and therefore Equation 6.21 involves the solution of a huge set of linear equations. For a sparse matrix A a fast solution of such a problem can be found using multi-grid methods. However, since we consider contrast at all levels of a Gaussian pyramid, the matrix A in our case is not sparse. From the visualization of the matrix A (see Figure 6.15), we can conclude that the matrix has a regular structure, but certainly cannot be considered sparse. Such multi-resolution problem seems to be well suited for the Fourier methods [Press et al. 2002, Chapter 19.4]. However, the problem cannot be solved using those methods either, since they require matrix coefficients to be of the same value while the constant factors $p^k_{i,j}$ introduce variations between matrix coefficients. We found that the *biconjugate gradient method* [Press et al. 2002, Chapter 2.7] is appropriate for our problem and gives results in acceptable time. The *biconjugate gradient method* is normally considered to be slower than more advanced multi-grid methods, however we found that it converges equally fast for our problem. This is because the structure of the A matrix enforces that iterative improvements are performed for all spatial frequencies of an image, which is also the goal of multi-grid methods. The biconjugate gradient method is also often used as a part of a multi-grid algorithm.

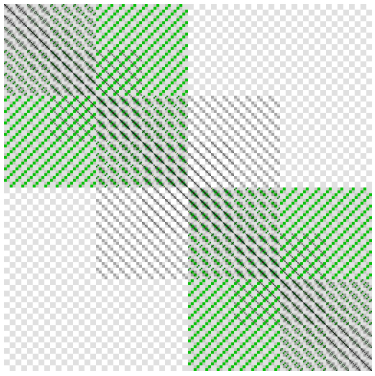


Figure 6.15: Visualization of the matrix A, which is involved in the solution of the optimization problem for a 1-mega-pixel image. White color denotes zero coefficients, which increase in magnitude with darker colors. Gray corresponds to positive and green to negative coefficients.

The biconjugate gradient method involves an iterative procedure, in which an image stored in the vector X is refined in each iteration. The attractiveness of this method is that it requires only an efficient computation of the product $\Psi = A \cdot X$. For clarity consider only the nearest neighborhood of each pixel, although the algorithm can be easily generalized to a larger pixel neighborhood at moderate computational cost. The contrast is computed between a pixel and its four neighbors within the same level of a Gaussian pyramid. Let X^k be a matrix holding pixel values of an image at the k-level of a Gaussian pyramid. Then, we can compute the product Ψ using the following recursive formula:

$$\Psi^k(X^k) = X^k \times \mathscr{L} + upsample[\Psi^{k+1}(downsample[X^k])], \tag{6.22}$$

where X^k is a solution at the k-th level of the pyramid, the operator $\times$ denotes convolution, $\mathscr{L}$ is the kernel

$$\mathscr{L} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{6.23}$$

and $upsample[]$ and $downsample[]$ are image upsampling and downsampling operators. The recursion stops when one of the image dimensions is less than 3 pixels after several successive downsamplings. The right-hand term B can be computed using another recursive formula:

$$\begin{aligned} B^k(\hat{G}^k) &= \hat{G}^k_{:,x} \times Dx + \hat{G}^k_{:,y} \times Dy + \\ &+ upsample[B^{k+1}(downsample[\hat{G}^k])] \end{aligned} \tag{6.24}$$

where $\hat{G}^k$ is the modified contrast at the k-th level of the pyramid, $\hat{G}^k_{:,x}$ and $\hat{G}^k_{:,y}$ are the subsets of contrast values $\hat{G}^k$ for horizontal and vertical neighbors, and Dx, Dy are the convolution kernels:

$$Dx = \begin{bmatrix} 1 & -1 \end{bmatrix} \quad Dy = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{6.25}$$

For simplicity, we did not include the coefficients $p^k_{i,j}$ in the above equations. Note that if only the first level of the pyramid is considered, the problem is reduced to the solution of Poisson's equation as in [Fattal et al. 2002]. To account for the boundary conditions, we can pad each edge of an image with a line or column that is a replica of the image edge.

6.8 Reconstruction of Color

Many applications, including the majority of tone mapping algorithms, focus on the processing of luminance while chrominance is transferred from an original image. The goal is to preserve the same perceived hue and color saturation while altering luminance. Hue can be easily preserved if a color space that decorrelates chrominance from luminance is used (such as *LHS* or *Yxy*). Preserving the perceived color saturation is much more difficult since it is strongly and non-linearly correlated with luminance. Additionally, the perceived color saturation may change if luminance contrast is modified. A transfer of color saturation seems to be a difficult and still unsolved problem. Therefore, for the proposed tone mapping algorithms, we follow the method employed

in most tone mapping algorithms, which involves rescaling red, green and blue color channels proportionally to the luminance and desaturating colors to compensate for higher local contrast. For each pixel, we compute:

$$C_{out} = \frac{X - l_{min} + s(C_{in} - L_{in})}{l_{max} - l_{min}} \tag{6.26}$$

where C_{in} and C_{out} are the input and output pixel values for the red, green or blue color channel, L_{in} is the input luminance, and X is the result of the optimization (all values are in the logarithmic domain). The resulting values C_{out} are within the range from 0 to 1. The parameter s is responsible for the saturation of colors and is usually set between 0.4 and 0.6. If P_k is k-th percentile of X and $d = \max(P_{50} - P_{0.1}, P_{99.9} - P_{50})$, then $l_{min} = P_{50} - d$ and $l_{max} = P_{50} + d$. This way, the average gray level is mapped to the gray level of the display ($r = g = b = 0.5$) and overall contrast is not lost due to a few very dark or bright pixels. Note that fine tuning of l_{max} and l_{min} values is equivalent to so called *gamma-correction* used as a last step of many tone mapping algorithms. This is because a power function in the linear domain corresponds to a multiplication in the logarithmic domain: $\log(x^\gamma) = \gamma \cdot \log(x)$. Equation 6.26 is similar to formulas proposed by Tumblin and Turk [1999] but it is given in the logarithmic domain and includes a linear scaling. The resulting color values, C_{out}, can be linearly mapped directly to the pixel values of a gamma corrected (perceptually linearized) display.

6.9 Discussion

The proposed framework is most suitable for those problems where the best solution is a compromise between conflicting goals. For example, in the case of contrast mapping (Section 6.4), we try to compress an overall contrast by suppressing low frequencies (low frequency contrast has large values and thus is heavily compressed), while preserving details. However, when enhancing details we also lessen compression of overall contrast since details can span a broad range of spatial frequencies (the lower levels of low-pass Gaussian pyramid) including low-frequencies, which are primarily responsible for an overall contrast. The strength of our method comes from the fact that the objective function given in Equation 6.7 leads to a compromise between the conflicting goals of compressing low-frequency large contrast and preserving small contrast of the high frequency details.

The minimization problem introduced in Equation 6.7 seems similar to solving Poisson's equation in order to reconstruct an image from gradients, as proposed by Fattal et al. [Fattal et al. 2002]. The difference is that our objective function takes into account a broader neighborhood of a pixel (summation over j) and puts additional optimization constraints on the contrast at coarser levels of the pyramid (summation over l), which improves a restoration of low frequency information. When an objective function is limited only to the finest level of the Gaussian pyramid (as it is done in Poisson's equation), the low frequency content may be heavily distorted in the resulting image[3]. This is illustrated on the examples of a 1-D signal in Figure 6.16 and a tone-mapped image in Figure 6.17. In general, Poisson solvers may lead to the reduction (or even reversal)

[3]Loss of low-frequency contrast is also visible in Figure 3 in the paper by Fattal et al. [2002], where low intensity levels of the left and middle peaks in the original image (a) are strongly magnified in the output image (f), so that they eventually become higher than the originally brightest image part on the right side.

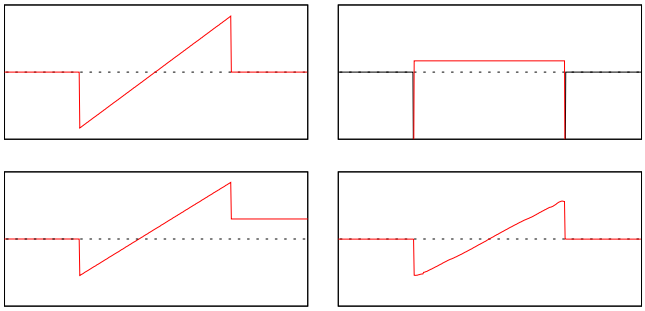


Figure 6.16: When an original signal (upper left) is restored from attenuated gradients (upper right) by solving Poisson's equation (or integration in 1-D), the flat parts of the restored signal are shifted relative to each other (lower left). However, if the minimization constraints are set for multiple levels of the pyramid as in our proposed method, the flat parts can be accurately restored although the sharp peaks are slightly blurred (lower right).

of global low-frequency contrast measured between disconnected image fragments. Other researchers have also noticed this problem. Gooch at al. [2005] experimented with Poisson solvers and found that they do not work well for "large disconnected isoluminant regions because they compute gradients over nearest neighbors, ignoring difference comparison over distances greater than one pixel". They overcome this problem by including a larger number of neighbors for each pixel in the objective function. The importance of global contrast and the fact that considering only local contrast gives wrong results was also discussed in [Rasche et al. 2005, Figure 2]. Our framework can be considered as a generalization of the gradient domain methods based on Poisson solvers. We consider larger neighborhoods for local contrast and also several levels of a Gaussian pyramid for global contrast. Such an approach is both perceptually plausible and computationally much more efficient than solving the optimization problem for contrast values between all pixels in the image [Gooch et al. 2005].

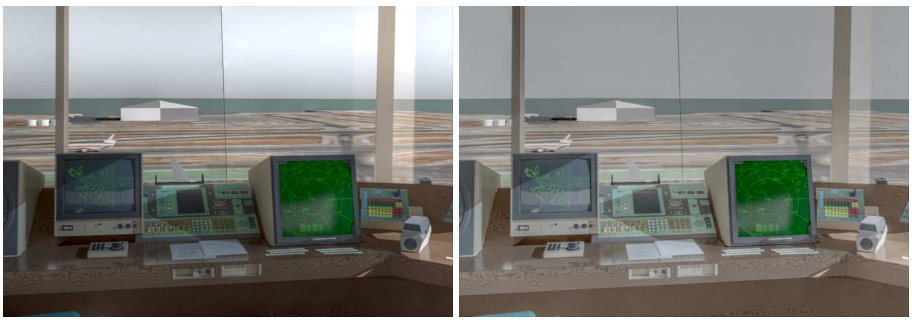


Figure 6.17: The algorithm by Fattal et al. [2002] (left) renders windows panes of different brightness due to the local nature of the optimization procedure. The contrast compression on the multi-scale contrast pyramid used in our method can maintain proper global contrast proportions (right). *Image courtesy of Greg Ward.*

The most computationally expensive part of the proposed framework is the contrast-to-luminance transformation. The solution of the minimization problem for 1–5 Mpixel

images can take from several seconds to half a minute to compute on a modern PC. This limits the application of the algorithm to off-line processing. However, our solution is not much less efficient than multi-grid methods (for example [Fattal et al. 2002]) as discussed in Section 6.7.

6.10 Conclusions and Future Work

In this chapter we presented a framework for image processing operations that work in the visual response space. Our framework is in many aspects similar to the gradient methods based on solving Poisson's equation, which prove to be very useful for image and video processing. Our solution can be regarded as a generalization of these methods which consider contrast on multiple spatial frequencies. We express a gradient-like representation of images using physical and perceptual terms, such as contrast and visual response. This gives perceptual basis for the gradient methods and offers several extensions from which these methods can benefit. For instance, unlike the solution of Poisson's equation, our pyramidal contrast representation ensures proper reconstruction of low frequencies and does not reverse global brightness levels. We also introduce a transducer function that can give the response of the HVS for the full range of contrast amplitudes, which is especially desired in case of HDR images. Other applications can also make use of the contrast discrimination thresholds, which describe suprathreshold performance of the eye from low to high contrast. As a proof of concept, we implemented two tone mapping algorithms and a saliency preserving color to gray mapping inside our framework. The tone mapping was shown to produce sharper images than the other contrast reduction methods. We believe that our framework can also find many applications in image and video processing.

In the future, we would like to improve the performance of reconstructing the image from the contrast representation, which would make the framework suitable for real-time applications. We would also like to include color information using a representation similar to luminance contrast. The framework could be extended to handle animation and temporal contrast. Furthermore, the accuracy of our model can be improved for the threshold contrast if the Contrast Sensitivity Function were taken into account in the transducer function. A simple extension is required to adapt our framework to the task of predicting visible differences in HDR images: since the response in our framework is in fact scaled in JND units, the difference between response values of two images gives the map of visible differences. One possible application of such HDR visible difference predictor could be the control of global illumination computation by estimating visual masking [Ramasubramanian et al. 1999, Dumont et al. 2003]. Finally, we would like to experiment with performing common image processing operations in the visual response space.

Chapter 7

Conclusions and Future Work

In the following sections we briefly summarize contributions of this dissertation, draw conclusions and give an outlook on future work.

7.1 Conlusions

The main motivation of the work presented in this dissertation was to create a complete HDR pipeline, from acquisition, through storage, to display. Together with the existing solutions, the proposed algorithms form an end-to-end pipeline: we can acquire HDR video sequences with HDR cameras, we can compress them using one of the proposed video compression algorithms, and finally display them directly on an HDR display, or on LDR display after applying tone mapping.

Our solutions are strongly influenced by aspects of human visual perception. Most of the digital imaging algorithms are the result of a trade-off between quality (fildelity) and performance (computational cost). The best compromise between these two conflicting goals can be found only if the human visual perception is taken into account. Considered as an optimization problem, the human visual perception defines both the boundary conditions (threshold characteristics) and the weighting factors (suprathreshold characteristics) for this compromise. Many digital imaging problems lack a well defined and objective goal. Due to the lack of such a goal, their only objective is producing subjectively pleasing images. Computational models of the human visual perception can help in defining an objective goal instead, making the objectives of the results of the proposed solution measurable and therefore possible to evaluate.

The performance of the human visual system at contrast detection defined our objectives for the derivation of the color space for HDR pixels (Section 5.3). The derived color space gave the best trade-off between the quality (visibility of contouring artifacts) and efficiency (number of required bits), which made this color space well suited for image and video compression. The proposed extension of MPEG-4, as the first of this kind, demonstrated potentials of HDR video, which includes not only better reproduction on HDR displays, but also tone mapping that is selectable at playback and the simulation of various perceptual and optical effects, such as glare, night vision and motion blur (Section 5.4). To facilitate a smooth transition from the traditional content to

HDR video material, we proposed the backward-compatible HDR MPEG compression algorithm, which can efficiently encode HDR content together with its low-dynamic range counterpart (Section 5.6). The key component of this compression algorithm was a perceptually motivated filter for removing invisible noise.

Although a common practice is to evaluate video compression algorithms using the simple mean-square-error or peak-to-signal-ratio metrics, which do not account for the human perception factors, we took an effort to evaluate the proposed compression algorithms using a visual difference predictor. The predictor models the early stages of the human visual system to predict noticeable differences, for example between the original and compressed image. To make such a predictor suitable for HDR content, we extended the original implementation with models of eye optics, photoreceptors' response and local adaptation, which play an important role in the perception of HDR images (Chapter 4).

An understanding of the processing that is performed in the retina not only helps to design better compression algorithms, but can also facilitate image editing. The proposed framework for contrast-domain image processing mimics contrast processing in the retina, including the non-linear response to contrast (Chapter 6). Images modified within the framework do not suffer from contrast reversal artifacts, which makes this framework especially suitable for contrast-enhancing tone mapping operators.

The solutions presented in this dissertation provide a set of tools for storage, editing and quality assessment of HDR content. Many of the presented solutions are provided together with a software implementation, which is available for download from the web pages (more information on the software can be found in Appendix A). We believe that the proposed algorithms together with the accompanying software can not only facilitate further research on HDRI, but also find many practical applications.

7.2 Future Work

Since HDRI is a new field with much research taking place only recently, it is an abundance of unsolved problems. Certainly, rendering of HDR images on a variety of display devices (tone mapping) is still not fully solved and not necessarily a well defined problem, despite the large number of recently published papers on this topic. Although effective luminance range reduction algorithms has been proposed, the treatment of color is not adequate. There is also a question how tone mapping should automatically adjust its results to a display device and the viewing conditions. Another important topic in the HDRI field is the standarization of common image and video formats, so that the content could be easily exchanged between the applications. The lack of standards can delay adoption of HDRI by the digital imaging industry.

There are also several improvements that we would like to introduce to the solutions presented in this dissertation. The color space for HDR pixels can efficiently encode luminance, but we are not satisfied with its performance for chrominance data. The simplistic approach taken for the CIE uniform chromacity scales u', v' does not seem to offer the best solution for our application. Chromatic information should probably undergo a similar non-linear compression as luminance, which would be determined by the color difference detection thresholds.

Also, the performance of the proposed video compression algorithms can be improved. For example, while testing the backward-compatible HDR MPEG compression, we noticed that the proper choice of the reconstruction function has a major effect on the compression performance. Different approaches for estimating the reconstruction function should be validated on a range of test sequences to choose the best solution. Also, significant bit-rate savings can be achieved without much loss of quality if an encoder is allowed to distort video as long as the distortion are smartly concealed so that they do not appear as artifacts. We observed this effect during sub-sampling of the residual channel in the backward-compatible HDR MPEG compression.

The visual difference predictor for HDR images (HDR-VDP) requires further validation in subjective experiments. Especially the detection of differences on high contrast stimuli, which cause glare and stimulate local adaptation effects, needs to be verified.

We presented only a limited number of applications for the contrast processing framework. We believe that the framework can also be applied to other purposes, such as image compositing, inpainting and non-photorealistic stylization.

Index

Bibliography

[Adams 1981] ADAMS, A. 1981. *The Print, The Ansel Adams Photography Series 3*. New York Graphic Society.

[Agarwala et al. 2004] AGARWALA, A., DONTCHEVA, M., AGRAWALA, M., DRUCKER, S., COLBURN, A., CURLESS, B., SALESIN, D., AND COHEN, M. 2004. Interactive digital photomontage. *ACM Transactions on Graphics 23*, 3, 294–302.

[Ahumada and Peterson 1993] AHUMADA, A., AND PETERSON, H. 1993. Luminance-model-based DCT quantization for color image compression. In *Human Vision, Visual Processing and Digital Display*, SPIE, volume 3299, 191–201.

[Ashikhmin 2002] ASHIKHMIN, M. 2002. A tone mapping algorithm for high contrast images. In *Rendering Techniques 2002: 13th Eurographics Workshop on Rendering*, 145–156.

[Barten 1990] BARTEN, P. 1990. Subjective image quality of high-definition television pictures. In *Proc. of the Soc. for Inf. Disp.*, vol. 31, 239–243.

[Barten 1999] BARTEN, P. G. 1999. *Contrast sensitivity of the human eye and its effects on image quality*. SPIE – The International Society for Optical Engineering.

[Bennett and McMillan 2005] BENNETT, E. P., AND MCMILLAN, L. 2005. Video enhancement using per-pixel virtual exposures. *ACM Transactions on Graphics 24*, 3, 845–852.

[Blackwell and Blackwell 1971] BLACKWELL, O., AND BLACKWELL, H. 1971. Visual performance data for 156 normal observers of various ages. *Journal of the Illuminating Engineering Society 1*, 1, 3–13.

[Bodmann 1973] BODMANN, H. 1973. Visibility assessment in lighting engineering. *Journal of the Illuminating Engineering Society 2*, 4, 437–443.

[Bogart et al. 2003] BOGART, R., KAINZ, F., AND HESS, D. 2003. OpenEXR image file format. In *ACM SIGGRAPH 2003, Sketches & Applications*.

[Bolin and Meyer 1998] BOLIN, M. R., AND MEYER, G. W. 1998. A perceptually based adaptive sampling algorithm. In *Proc. of ACM SIGGRAPH 1998*, 299–309.

[Border and Guillotel 2000] BORDER, P., AND GUILLOTEL, P. 2000. Perceptually adapted MPEG video encoding. In *Proc. of Human Vision and Electronic Imaging V*, SPIE, volume 3959, 168–175.

[Bradley 1999] BRADLEY, A. P. 1999. A wavelet visible difference predictor. *IEEE Transactions on Image Processing 8*, 5, 717–730.

[Brun and Tremeau 2003] BRUN, L., AND TREMEAU, A. 2003. Color quantization. In *Digital Color Imaging Handbook*, G. Sharma, Ed. CRC Press, ch. 9, 589–638.

[Caelli et al. 1985] CAELLI, T., HUBNER, M., AND RENTSCHLER, I. 1985. Detection of phase-shifts in two-dimensional images. *Perception and Psychophysics 37*, 536–542.

[Campbell and Robson 1968] CAMPBELL, F., AND ROBSON, J. 1968. Application of Fourier analysis to the visibility of gratings. *Journal of Psychology 197*, 551–566.

[CIE 1981] CIE. 1981. *An Analytical Model for Describing the Influence of Lighting Parameters Upon Visual Performance*, vol. 1. Technical Foundations, CIE 19/2.1. International Commision on Illumination.

[CIE 1986] CIE. 1986. *Colorimetry*, vol. CIE 15.2. International Commision on Illumination.

[CIE 2002] CIE. 2002. *A Colour Appearance Model for Colour Management Systems: CIECAM02*, vol. CIE 159:2004. International Commision on Illumination.

[Daly and Feng 2003] DALY, S. J., AND FENG, X. 2003. Bit-depth extension using spatiotemporal microdither based on models of the equivalent input noise of the visual system. In *Color Imaging VIII: Processing, Hardcopy, and Applications*, SPIE, volume 5008, 455–466.

[Daly and Feng 2004] DALY, S. J., AND FENG, X. 2004. Decontouring: Prevention and removal of false contour artifacts. In *Proc. of Human Vision and Electronic Imaging IX*, SPIE, vol. 5292, 130–149.

[Daly 1993] DALY, S. 1993. The Visible Differences Predictor: An algorithm for the assessment of image fidelity. In

Digital Image and Human Vision, Cambridge, MA: MIT Press, A. Watson, Ed., 179–206.

[Deeley et al. 1991] DEELEY, R., DRASDO, N., AND CHARMAN, W. N. 1991. A simple parametric model of the human ocular modulation transfer function. *Ophthalmology and Physiological Optics 11*, 91–93.

[Deering 2005] DEERING, M. F. 2005. A photon accurate model of the human eye. *ACM Transactions on Graphics 24*, 3, 649–658.

[Demos 2004] DEMOS, G. 2004. High quality, wide dynamic range, compression system. In *SMPTE Technical Conference Proceedings*.

[DICOM PS 3-2004 2004] DICOM PS 3-2004. 2004. Part 14: Grayscale standard display function. In *Digital Imaging and Communications in Medicine (DICOM)*. National Electrical Manufacturers Association.

[Drago et al. 2003] DRAGO, F., MYSZKOWSKI, K., ANNEN, T., AND CHIBA, N. 2003. Adaptive logarithmic mapping for displaying high contrast scenes. *Computer Graphics Forum, Proceedings of Eurographics 2003 22*, 3, 419–426.

[Dumont et al. 2003] DUMONT, R., PELLACINI, F., AND FERWERDA, J. A. 2003. Perceptually-driven decision theory for interactive realistic rendering. *ACM Transactions on Graphics 22*, 2, 152–181.

[Durand and Dorsey 2002a] DURAND, F., AND DORSEY, J. 2002. Fast bilateral filtering for the display of high-dynamic-range images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 257–266.

[Durand and Dorsey 2002b] DURAND, F., AND DORSEY, J. 2002. Fast bilateral filtering for the display of high-dynamic-range images. *ACM Transactions on Graphics 21*, 3, 257–266.

[Fairchild and Johnson 2004] FAIRCHILD, M., AND JOHNSON, G. 2004. icam framework for image appearance differences and quality. *Journal of Electronic Imaging 13*, 1, 126–138.

[Fairchild 1997] FAIRCHILD, M. D. 1997. *Color Appearance Models*. Addison Wesley.

[Fattal et al. 2002] FATTAL, R., LISCHINSKI, D., AND WERMAN, M. 2002. Gradient domain high dynamic range compression. *ACM Transactions on Graphics 21*, 3, 249–256.

[Ferwerda et al. 1996] FERWERDA, J., PATTANAIK, S., SHIRLEY, P., AND GREENBERG, D. 1996. A model of visual adaptation for realistic image synthesis. In *Proceedings of ACM SIGGRAPH 96*, 249–258.

[Ferwerda et al. 1997] FERWERDA, J. A., SHIRLEY, P., PATTANAIK, S. N., AND GREENBERG, D. P. 1997. A model of visual masking for computer graphics. In *Proc. of ACM SIGGRAPH 1997*, 143–152.

[Foley and Legge 1981] FOLEY, J., AND LEGGE, G. 1981. Contrast detection and near-threshold discrimination in human vision. *Vision Research 21*, 1041–1053.

[Georgeson and Sullivan 1975] GEORGESON, M., AND SULLIVAN, G. 1975. Contrast constancy: Deblurring in human vision by spatial frequency channels. *Journal of Physiology 252*, 627–656.

[Georgiev 2005] GEORGIEV, T. 2005. Vision, healing brush, and fiber bundles. In *IS&T/SPIE Conf. on Hum. Vis. and Electronic Imaging X*, vol. 5666, 293–305.

[Gonzalez and Woods 2001] GONZALEZ, R. C., AND WOODS, R. E. 2001. *Digital Image Processing, 2nd Edition*. Addison-Wesley.

[Gooch et al. 2005] GOOCH, A. A., OLSEN, S. C., TUMBLIN, J., AND GOOCH, B. 2005. Color2gray: Salience-preserving color removal. *ACM Transactions on Graphics 24*, 3.

[Heeger and Teo 1995] HEEGER, D., AND TEO, P. 1995. A model of perceptual image fidelity. In *Proc. of IEEE Int'l Conference Image Processing*, 343–345.

[Hood and Finkelstein 1986] HOOD, D., AND FINKELSTEIN, M. 1986. Sensitivity to light. In *Handbook of Perception and Human Performance: 1. Sensory Processes and Perception*, Wiley, New York, K. Boff, L. Kaufman, and J. Thomas, Eds., vol. 1.

[Horn 1974] HORN, B. 1974. Determining lightness from an image. *Computer Graphics and Image Processing 3*, 1, 277–299.

[Hunt 1995] HUNT, R. 1995. *The Reproduction of Colour in Photography, Printing and Television: 5th Edition*. Fountain Press.

[Hurlbert 1986] HURLBERT, A. 1986. Formal connections between lightness algorithms. *Journal of the Optical Society of America A 3*, 10, 1684–1693.

[IEC 61966-2-1:1999 1999] IEC 61966-2-1:1999. 1999. *Multimedia systems and equipment - Colour measurement and management - Part 2-1: Colour management - Default RGB*

colour space - sRGB. International Electrotechnical Commission.

[Irawan et al. 2005] IRAWAN, P., FERWERDA, J. A., AND MARSCHNER, S. R. 2005. Perceptually based tone mapping of high dynamic range image streams. In *Proceedings of the Eurographics Symposium on Rendering*, 231–242.

[ISO-IEC 14496-2 1999] ISO-IEC 14496-2. 1999. *Information technology: Coding of audio-visual objects, Part 2: Visual*. International Organization for Standardization, Geneva, Switzerland.

[ISO/IEC 14496-10 2005] ISO/IEC 14496-10. 2005. *Information technology: Coding of audio-visual objects, Part 10: Advanced Video Coding*. International Organization for Standardization, Geneva, Switzerland.

[Jin et al. 1998] JIN, E. W., FENG, X.-F., AND NEWELL, J. 1998. The development of a color visual difference model (CVDM). In *IS&T's 1998 Image Processing, Image Quality, Image Capture, Systems Conf.*, 154–158.

[Kakimoto et al. 2005] KAKIMOTO, M., MATSUOKA, K., NISHITA, T., NAEMURA, T., AND HARASHIMA, H. 2005. Glare generation based on wave optics. *Computer Graphics Forum 24*, 2 (June), 185–194.

[Kingdom and Whittle 1996] KINGDOM, F. A. A., AND WHITTLE, P. 1996. Contrast discrimination at high contrasts reveals the influence of local light adaptation on contrast processing. *Vision Research 36*, 6, 817–829.

[Krawczyk et al. 2005a] KRAWCZYK, G., GOESELE, M., AND SEIDEL, H.-P. 2005. Photometric calibration of high dynamic range cameras. Research Report MPI-I-2005-4-005, Max-Planck-Institut für Informatik, April.

[Krawczyk et al. 2005b] KRAWCZYK, G., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2005. Perceptual effects in real-time tone mapping. In *SCCG '05: Proc. of the 21st Spring Conference on Computer Graphics*, 195–202.

[Land 1964] LAND, E. H. 1964. The retinex. *American Scientist 52*, 2, 247–264.

[Legge 1979] LEGGE, G. 1979. Spatial frequency masking in human vision: binocular interactions. *Journal of the Optical Society of America 69*, 838–847.

[Levin et al. 2004] LEVIN, A., ZOMET, A., PELEG, S., AND WEISS, Y. 2004. Seamless image stitching in the gradient domain. In *Eighth European Conference on Computer Vision*, vol. 4, 377–389.

[Li et al. 1998] LI, B., MEYER, G., AND KLASSEN, R. 1998. A comparison of two image quality models. In *Human Vision and Electronic Imaging III*, SPIE, volume 3299, 98–109.

[Li et al. 2005] 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.

[Lubin and Pica 1991] LUBIN, J., AND PICA, A. 1991. A non-uniform quantizer matched to the human visual performance. *Society of Information Display Int. Symposium Technical Digest of Papers 22*, 619–622.

[Lubin 1995] LUBIN, J. 1995. A visual discrimination model for imaging system design and evaluation. In *Vis. Models for Target Detection*, 245–283.

[Lucian et al. 2005] LUCIAN, I., FELICIA, S., CHARLES, S., AND SIEFKIEN, H. 2005. Digital encode and method of encoding high dynamic range video images. In *US Patent 6,867,717*.

[Mantiuk et al. 2004a] MANTIUK, R., KRAWCZYK, G., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2004. Perception-motivated high dynamic range video encoding. *ACM Transactions on Graphics 23*, 3, 730–738.

[Mantiuk et al. 2004b] MANTIUK, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2004. Visible difference predicator for high dynamic range images. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, 2763–2769.

[Mantiuk et al. 2005a] MANTIUK, R., DALY, S., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2005. Predicting visible differences in high dynamic range images - model and its calibration. In *Proc. of Human Vision and Electronic Imaging X*, SPIE, volume 5666, 204–214.

[Mantiuk et al. 2005b] MANTIUK, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2005. A perceptual framework for contrast processing of high dynamic range images. In *APGV '05: 2nd Symposium on Appied Perception in Graphics and Visualization*, 87–94.

[Mantiuk et al. 2006a] MANTIUK, R., EFREMOV, A., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2006. Backward compatible high dynamic range mpeg video compression. *ACM Transactions on Graphics 25*, 3.

[Mantiuk et al. 2006b] MANTIUK, R., EFREMOV, A., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2006. Design and evaluation of

backward compatible high dynamic range video compression. MPI Technical Report MPI-I-2006-4-001, Max Planck Institute für Informatik.

[Mantiuk et al. 2006c] MANTIUK, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2006. Lossy compression of high dynamic range images and video. In *Proc. of Human Vision and Electronic Imaging XI*, SPIE, San Jose, USA, vol. 6057 of *Proceedings of SPIE*, 60570V.

[Mantiuk et al. 2006d] MANTIUK, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2006. A perceptual framework for contrast processing of high dynamic range images. *ACM Transactions on Applied Perception 3*.

[Marimont and Wandell 1994] MARIMONT, D., AND WANDELL, B. 1994. Matching colour images: the effects of axial chromatic aberration. *J. Opt. Soc. Am. A 11*, 12, 2113–3122.

[Matusik and Pfister 2004] MATUSIK, W., AND PFISTER, H. 2004. 3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. *ACM Transactions on Graphics 23*, 3, 814–824.

[Michelson 1927] MICHELSON, A. 1927. *Studies in Optics*. U. Chicago Press.

[Moon and Spencer 1944] MOON, P., AND SPENCER, D. 1944. Visual data applied to lighting design. *J. Opt. Soc. Am. 34*, 605.

[Morovic and Luo 2001] MOROVIC, J., AND LUO, M. R. 2001. The fundamentals of gamut mapping: A survey. *Journal of Imaging Science and Technology 45*, 3, 283–290.

[Mullen 1985] MULLEN, K. T. 1985. The contrast sensitivity of human color vision to red-green and blue-yellow chromatic gratings. *Journal of Psychology 359*, 381–400.

[Nadenau 2000] NADENAU, M. 2000. *Integration of Human color vision Models into High Quality Image Compression*. PhD thesis, École Polytechnique Fédéral Lausane.

[Nakamae et al. 1990] NAKAMAE, E., KANEDA, K., OKAMOTO, T., AND NISHITA, T. 1990. A lighting model aiming at drive simulators. In *SIGGRAPH '90: Proceedings of the 17th annual conference on Computer graphics and interactive techniques*, 395–404.

[Pattanaik et al. 1998] PATTANAIK, S. N., FERWERDA, J. A., FAIRCHILD, M. D., AND GREENBERG, D. P. 1998. A multiscale model of adaptation and spatial vision for realistic image display. In *Siggraph 1998, Computer Graphics Proceedings*, 287–298.

[Pattanaik et al. 2000] PATTANAIK, S., TUMBLIN, J., YEE, H., AND GREENBERG, D. 2000. Time-dependent visual adap-

tation for realistic image display. In *Proceedings of ACM SIGGRAPH 2000*, Computer Graphics Proceedings, Annual Conference Series, 47–54.

[Peli 1990] PELI, E. 1990. Contrast in complex images. *Journal of the Optical Society of America A 7*, 10, 2032–2040.

[Perez et al. 2003] PEREZ, P., GANGNET, M., AND BLAKE, A. 2003. Poisson image editing. *ACM Transactions on Graphics 22*, 3, 313–318.

[Poynton 2003] POYNTON, C. 2003. *Digital Video and HDTV: Algorithms and Interfaces*. Morgan Kaufmann.

[Press et al. 2002] PRESS, W., TEUKOLSKY, S., VETTERLING, W., AND FLANNERY, B. 2002. *Numerical Recipes in C++*, second ed. Cambridge University Press.

[Ramasubramanian et al. 1999] RAMASUBRAMANIAN, M., PATTANAIK, S. N., AND GREENBERG, D. P. 1999. A perceptually based physical error metric for realistic image synthesis. In *Proceedings of ACM SIGGRAPH 1999*, 73–82.

[Rasche et al. 2005] RASCHE, K., GEIST, R., AND WESTALL, J. 2005. Re-coloring images for gamuts of lower dimension. *Computer Graphics Forum 24*, 3, 423–432.

[Reinhard et al. 2002a] REINHARD, E., STARK, M., SHIRLEY, P., AND FERWERDA, J. 2002. Photographic tone reproduction for digital images. *ACM Transactions on Graphics 21*, 3, 267–276.

[Reinhard et al. 2002b] REINHARD, E., STARK, M., SHIRLEY, P., AND FERWERDA, J. 2002. Photographic tone reproduction for digital images. *ACM Transactions on Graphics 21*, 3, 267–276.

[Reinhard et al. 2005] REINHARD, E., WARD, G., PATTANAIK, S., AND DEBEVEC, P. 2005. *High Dynamic Range Imaging. Data Acquisition, Manipulation, and Display*. Morgan Kaufmann.

[Safranek 1993] SAFRANEK, R. J. 1993. JPEG compliant encoder using perceptually based quantization. In *Human Vision, Visual Processing, and Digital Display IV*, SPIE, volume 1913, 117–126.

[Seetzen et al. 2004] SEETZEN, H., HEIDRICH, W., STUERZLINGER, W., WARD, G., WHITEHEAD, L., TRENTACOSTE, M., GHOSH, A., AND VOROZCOVS, A. 2004. High dynamic range display systems. *ACM Transactions on Graphics 23*, 3, 757–765.

[Sezan et al. 1987] SEZAN, M., YIP, K., AND DALY, S. 1987. Uniform perceptual quantization: Applications to digital radio-

graphy. *IEEE Transactions on Systems, Man, and Cybernetics 17*, 4, 622–634.

[Shen and Delp 1999] SHEN, K., AND DELP, E. 1999. Wavelet based rate scalable video compression. *IEEE Transactions on Circuits and Systems for Video Technology 9*, 1, 109–122.

[Simoncelli and Adelson 1989] SIMONCELLI, E., AND ADELSON, E. 1989. Nonseparable QMF pyramids. *Visual Communications and Image Processing 1199*, 1242–1246.

[Spaulding et al. 2003] SPAULDING, K. E., WOOLFE, G. J., AND JOSHI, R. L. 2003. Using a residual image to extend the color gamut and dynamic range of an sRGB image. In *Proc. of IS&T PICS Conference*, 307–314.

[Spencer et al. 1995] SPENCER, G., SHIRLEY, P., ZIMMERMAN, K., AND GREENBERG, D. 1995. Physically-based glare effects for digital images. In *Proceedings of ACM SIGGRAPH 95*, 325–334.

[Stevens and Stevens 1960] STEVENS, S., AND STEVENS, J. 1960. Brightness function: parametric effects of adaptation and contrast. *Journal of the Optical Society of America 50*, 11 (Nov.), 1139A.

[Stockman and Sharpe 2000] STOCKMAN, A., AND SHARPE, L. T. 2000. Spectral sensitivities of the middle- and long-wavelength sensitive cones derived from measurements in observers of known genotype. *Vision Research 40*, 1711–1737.

[Sun et al. 2004] SUN, J., JIA, J., TANG, C.-K., AND SHUM, H.-Y. 2004. Poisson matting. *ACM Transactions on Graphics 23*, 3, 315–321.

[Taylor et al. 1997] TAYLOR, C., PIZLO, Z., ALLEBACH, J. P., AND BOUMAN, C. 1997. Image quality assessment with a Gabor pyramid model of the Human Visual System. In *Hum. Vis. and Elect. Imaging*, SPIE, volume 3016, 58–69.

[Thomspon et al. 2002] THOMSPON, W. B., SHIRLEY, P., AND FERWERDA, J. A. 2002. A spatial post-processing algorithm for images of night scenes. *Journal of Graphics Tools 7*, 1, 1–12.

[Tumblin and Turk 1999] TUMBLIN, J., AND TURK, G. 1999. LCIS: A boundary hierarchy for detail-preserving contrast reduction. In *Siggraph 1999, Computer Graphics Proceedings*, 83–90.

[Van Meeteren and Vos 1972] VAN MEETEREN, A., AND VOS, J. J. 1972. Resolution and contrast sensitivity at low luminances. *Vision Research 12*, 825–833.

[Van Nes and Bouman 1967] VAN NES, F., AND BOUMAN, M. 1967. Spatial modulation transfer in the human eye. *Journal of the Optical Society of America 57*, 401–406.

[Walraven et al. 1990] WALRAVEN, J., ENROTH-CUGELL, C., HOOD, D., MACLEOD, D., AND SCHNAPF, J. 1990. The control of visual sensitivity. In *Visual perception: the neurophysiological foundations*, L. Spillmann and S. Werner, Eds. Academic Press, ch. 5, 53–101.

[Wandell 1995] WANDELL, B. 1995. *Foundations of Vision*. Sinauer Associates, Inc.

[Wang and Bovik 2002] WANG, Z., AND BOVIK, A. 2002. A universal image quality index. *IEEE Signal Processing Letters 9*, 3, 81–84.

[Ward and Simmons 2004] WARD, G., AND SIMMONS, M. 2004. Subband encoding of high dynamic range imagery. In *APGV '04: 1st Symposium on Applied Perception in Graphics and Visualization*, 83–90.

[Ward and Simmons 2005] WARD, G., AND SIMMONS, M. 2005. JPEG-HDR: A backwards-compatible, high dynamic range extension to JPEG. In *Proceedings of the 13th Color Imaging Conference*, 283–290.

[Ward Larson 1998] WARD LARSON, G. 1998. LogLuv encoding for full-gamut, high-dynamic range images. *Journal of Graphics Tools 3*, 1, 815–30.

[Ward 1991] WARD, G. 1991. Real pixels. *Graphics Gems II*, 80–83.

[Watson and Solomon 1997] WATSON, A. B., AND SOLOMON, J. A. 1997. A model of visual contrast gain control and pattern masking. *Journal of the Optical Society A 14*, 2378–2390.

[Watson et al. 1994] WATSON, A. B., SOLOMON, J. A., AHUMADA, A., AND GALE, A. 1994. DCT basis function visibility: Effects of viewing distances and contrast masking. In *Human Vision, Visual Processing, and Digital Display V*, SPIE, volume 2179, 99–108.

[Watson 1987] WATSON, A. 1987. The cortex transform: Rapid computation of simulated neural images. *Comp. Vis. Graph. and Image Proc. 39*, 311–327.

[Westheimer 1986] WESTHEIMER, G. 1986. The eye as an optical instrument. In *Handbook of Perception and Human Performance: 1. Sensory Processes and Perception*, K. Boff, L. Kaufman, and J. Thomas, Eds. Wiley, New York, 4.1–4.20.

[Whittle 1986] WHITTLE, P. 1986. Increments and decrements: Luminance discrimination. *Vision Research 26*, 10, 1677–1691.

[Wilson 1980] WILSON, H. 1980. A transducer function for threshold and suprathreshold human vision. *Biological Cybernetics 38*, 171–178.

[Wilson 1991] WILSON, H. 1991. Psychophysical models of spatial vision and hyperacuity. In *Vision and Visual Dysfunction: Spatial Vision*, D. Regan, Ed. Pan Macmillan, 64–86.

[Winkler 2005] WINKLER, S. 2005. *Digital video quality: vision models and metrics*. John Wiley & Sons.

[Wyszecki and Stiles 2000] WYSZECKI, G., AND STILES, W. 2000. *Color Science*. John Willey & Sons.

[Xu et al. 2005] XU, R., PATTANAIK, S., AND HUGHES, C. 2005. High-dynamic range still-image encoding in JPEG 2000. *IEEE Comp. Graph. and Appl. 26*, 6, 57–64.

[Yee and Pattanaik 2003] YEE, H., AND PATTANAIK, S. 2003. Segmentation and adaptive assimilation for detail-preserving display of high-dynamic range images. *The Visual Computer 19*, 457–466.

[Zeng et al. 2000] ZENG, W., DALY, S., AND LEI, S. 2000. Visual optimization tools in JPEG 2000. In *IEEE International Conference on Image Processing*, 37–40.

[Zetzsche and Hauske 1989] ZETZSCHE, C., AND HAUSKE, G. 1989. Multiple channel model for the prediction of subjective image quality. In *Human Vision, Visual Processing, and Digital Display*, SPIE, volume 1077, 209–216.

Appendix A

pfstools

Most of the traditional image processing libraries store each pixel using limited-precision integer numbers. Moreover, they offer very limited means of colorimetric calibration. To overcome these problems, we have implemented HDR imaging framework as a package of several command line programs for reading, writing, manipulating and viewing high-dynamic range (HDR) images and video frames. The package was intended to solve our current research problems, therefore simplicity and flexibility were priorities in its design. Since we found the software very useful in numerous projects, we decided to make it available for the research community as an Open Source project licensed under the GPL. The software is distributed under the name *pfstools* and its home page can be found at http://pfstools.sourceforge.net/.

The major role of the software is the integration of several imaging and image format libraries, such as *ImageMagick*, *OpenEXR* and *NetPBM*, into a single framework for processing high precision images. To provide enough flexibility for a broad range of applications, we have build *pfstools* on the following concepts:

- Images/frames should hold an arbitrary number of channels (layers), which can represent not only color, but also depth, alpha, and texture attributes;
- Each channel should be stored with high precision, using floating point numbers. If possible, the data should be colorimetrically calibrated and provide the precision that exceeds the capabilities of the human visual system;
- Luminance should be stored using physical units of cd/m^2 to distinguish between the night- and the day-light vision;
- There should be user data entries for storing additional, application specific information (e.g. colorimetric coordinates of the white point).

pfstools are built around a generic and simple image format, which requires only a few lines of code to be read or written from or to a data stream. The format offers arbitrary number of channels, each represented as a 2-D array of 32-bit floating point numbers. There is no compression as the files in this format are intended to be transferred internally between applications without writing them to a disk. A few channels have a predefined function. For example, channels with the IDs 'X', 'Y' and 'Z' are used to

store color data in the CIE XYZ (absolute) color space. This is different to most imaging frameworks that operate on RGB channels. The advantage of the CIE XYZ color space is that it is precisely defined in terms of spectral radiance and the full visible color gamut can be represented using only positive values of color components. The file format also offers a way to include in an image any number of user *tags* (name and value pairs), which can contain any application dependent data. A sequence of images is interpreted by all "pfs-compliant" applications as consequtive frames of an animation, so that video can be processed in the same way as images. The format is described in detail in a separate specification[1].

pfstools are a set of command line tools with almost no graphical user interface. This greatly facilitates scripting and lessens the amount of work needed to program and maintain a user interface. The exception is a viewer of HDR images. The main components of *pfstools* are: programs for reading and writing images in all major HDR and LDR formats (e.g. OpenEXR, Radiance's RGBE, logLuv TIFF, 16-bit TIFF, PFM, JPEG, PNG, etc.), programs for basic image manipulation (rotation, scaling, cropping, etc.), an HDR image viewer, and a library that simplifies file format reading and writing in C++. The package includes also an interface for *GNU Octave*, which is a high level mathematical language similar to matlab. The *GNU Octave* interface offers an environment for researching and prototyping HDR image and video processing algorithms, similar to the matlab toolkits for the traditional imaging. The *pfstools* framework does not impose any restrictions on the programming language. All programs that exchange data with *pfstools* must read or write the file format, but there is no need to use any particular library. The typical usage of *pfstools* involves executing several programs joined by UNIX pipes. The first program transmits the current frame or image to the next one in the chain. The final program should either display an image or write it to a disk. Such pipeline architecture improves flexibility of the software but also gives straightforward means for parallel execution of the pipeline components on multiprocessor computers. Some examples of command lines are given below:

`pfsin␣input.exr␣|␣pfsfilter␣|␣pfsout␣output.exr`

Read the image `input.exr`, apply the filter `pfsfilter` and write the output to `output.exr`.

`pfsin␣input.exr␣|␣pfsfilter␣|␣pfsview`

Read the image `input.exr`, apply the filter `pfsfilter` and show the result in an HDR image viewer.

```
pfsin␣in%04d.exr␣--frames␣100:2:200␣\
␣|␣pfsfilter␣|␣pfsout␣out%04d.hdr
```

Read the sequence of OpenEXR frames `in0100.exr`, `in0102.exr`, .., `in0200.exr`, apply the filter `pfsfilter` and write the result in Radiance's RGBE format to `out0000.hdr`, `out0001.hdr`, ...

---

pfstools is only a base set of tools which can be easily extended and integrated with other software. For example, *pfstools* is used to read, write and convert images and video frames for the prototype implementation of our image and video compression algorithms. HDR images can be rendered on existing displays using one of the several

[1] Specification of the *pfs* format can be found at:
http://www.mpi-sb.mpg.de/resources/pfstools/pfs_format_spec.pdf

implemented tone mapping algorithms from the *pfstmo* package[2], which is build on top of *pfstools*. Cameras can be calibrated and images rescaled in physical or colorimetrical units using the software from the *pfscalibration* package[3], which is also based on *pfstools*. A computational model of the human visual system – *HDR-VDP*[4] – uses *pfstools* to read its input from multitude of image formats.

We created *pfstools* to fill the gap in the imaging software, which can seldom handle HDR images. We have found from the e-mails we received and the discussion group contacts that *pfstools* is used for high definition HDR video encoding, medical imaging, variety of tone mapping projects, texture manipulations and quality evaluation of CG rendering.

[2] *pfstmo* home page: http://www.mpii.mpg.de/resources/tmo/
[3] *pfscalibration* home page: http://www.mpii.mpg.de/resources/hdr/calibration/pfs.html
[4] *HDR-VDP* home page: http://www.mpii.mpg.de/resources/hdr/vdp/index.html