Human Visual System Models in Computer Graphics

Tunç Ozan Aydın

Max-Planck-Institut für Informatik

Dissertation

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

Eingereicht am 11.Oktober.2010

Betreuender Hochschullehrer — Supervisor
Dr.-Ing. Habil. Karol Myszkowski, MPI Informatik, Saarbrücken, Germany
Prof. Dr. Hans-Peter Seidel, MPI Informatik, Saarbrücken, Germany

Gutachter — Reviewers
Dr.-Ing. Habil. Karol Myszkowski, MPI Informatik, Saarbrücken, Germany
Prof. Dr. Hans-Peter Seidel, MPI Informatik, Saarbrücken, Germany
Assoc. Prof. Dr. Jan Kautz, University College London, London, UK
Prof. Dr. Phillip Slusallek, Universität des Saarlandes, Saarbrücken, Germany

Dekan — Dean
Prof. Dr. Holger Hermanns, Universität des Saarlandes, Saarbrücken, Germany

Datum des Kolloquiums — Date of Defense
09.12.2010

Tunç Ozan Aydın
Max-Planck-Institut für Informatik
Stuhlsatzenhausweg 85
66123 Saarbrücken, Germany
`tunc@mpii.de`

Abstract

At the receiving end of visual data are humans; thus it is only natural to take into account various properties and limitations of the human visual system while designing new image and video processing methods. In this dissertation we build multiple models of human vision with different focuses and complexities, and demonstrate their use in computer graphics context.

The human visual system models we present perform two fundamental tasks: predicting the visual significance, and the detection of visual features. We start by showing that a perception based importance measure for edge strength prediction results in qualitatively better outcomes compared to commonly used gradient magnitude measure in multiple computer graphics applications. Another more comprehensive model including mechanisms to simulate maladaptation is used to predict the visual significance of images shown on display devices under dynamically changing lighting conditions.

The detection task is investigated in the context of image and video quality assessment. We present an extension to commonly used image quality metrics that enables HDR support while retaining backwards compatibility with LDR content. We also propose a new "dynamic range independent" image quality assessment method that can compare HDR-LDR (and vice versa) reference-test image pairs, in addition to image pairs with the same dynamic range. Furthermore, the design and validation of a dynamic range independent video quality assessment method, that models various spatiotemporal aspects of human vision, is presented along with pointers to a wide range of application areas including comparison of rendering qualities, HDR compression and temporal tone mapping operator evaluation.

Kurzfassung

Auf der Empfängerseite visueller Daten steht der Mensch. Beim Entwurf neuer Bild- und Videoverarbeitungsmethoden ist es daher selbstverständlich die verschiedenen Eigenschaften und Beschränkungen des menschlichen visuellen Systems zu berücksichtigen. In der vorliegenden Dissertation formulieren wir mehrere Modelle des menschlichen visuellen Wahrnehmung mit verschiedenen Schwerpunkten und verschiedenen Komplexitäten und demonstrieren ihre Verwendung im Zusammenhang mit Computergrafik.

Die Modelle des menschlichen visuellen Systems, die wir präsentieren, erfüllen zwei grundlegende Aufgaben: die visuelle Signifikanz vorhersagen und visuelle Merkmale detektieren. Wir beginnen, in dem wir zeigen, dass ein wahrnehmungsbasiertes Bedeutungsmaš für die Vorhersage von Kantenstärken im Vergleich allgemein gebräuchlichen Mašen basierend auf der Gradientenlänge qualitativ bessere Ergebnisse in Computergrafikanwendungen liefert. Ein weiteres, umfassenderes Modell, dass Verfahren beinhaltet, die Fehladaptionen simulieren, wird verwendet, um die visuelle Signifikanz von Bildern vorherzusagen, die auf Bildschirmen unter sich dynamisch ändernden Beleuchtungsverhältnissen gezeigt werden.

Die Aufgabe des Detektierens wird im Zusammenhang der Datenerhebung von Bild- und Videoqualität untersucht. Wir präsentieren eine Erweiterung zu allgemein verwendeten Bildqualitätsmetriken, die HDR Unterstützung erlaubt, während Rückwärtskompatibilität zu LDR-Inhalten erhalten bleibt. Wir schlagen auch eine neue "dynamischer-Umfang-unabhängige" Methode zur Datenerhebung der Bildqualität vor, die zusätzlich zu Bildern mit gleichem dynamischen Umfang, auch HDR-LDR-Bildpaare (und umgekehrt) von Referenztests vergleichen kann. Zusammen mit Vermerken zu einer grošen Auswahl von Anwendungsbereichen, wie zum Beispiel dem Vergleich von Renderqualität, HDR-Kompression und Operatorevaluation von temporal tone mapping, wird weiterhin der Entwurf und die Validierung der dynamischer-Umfang-unabhängigen Datenerhebungsmethode für die Videoqualität präsentiert, die verschiedene raum-zeitliche Aspekte der menschlichen Wahrnehmung modelliert.

Summary

In this dissertation we explore the use of human visual system models in computer graphics context. We develop vision models of various scopes and complexities. These models are both used as the basis of the novel techniques we propose, and also to build upon the state-of-the art. The theoretical work in this dissertation is coupled with multiple psychophysical experiments for calibration and validation of the human visual system models in order to match the perception of an average observer.

We develop a simplistic human vision model that accounts for luminance adaptation and visual masking, and integrate it to a second generation wavelet based edge preserving image decomposition framework. The visual significance prediction of the perceptual model replaces the gradient magnitude as the edge strength measure without introducing a significant computational cost. We show that the extended framework is more intuitive in edge preserving smoothing and contrast enhancement, and results in qualitative improvements in the outcomes of current HDR image retargeting, tone mapping and HDR panorama stitching methods.

While there is a significant body of research focused on making images look more plausible, very little attention has been paid on how those images would be perceived on actual display devices. Moreover, due to the proliferation of mobile devices, it is no longer possible to assume that the observer will view an image on a desktop display in a controlled lighting environment. One should rather account for the effect of dynamically changing lighting conditions on the perception of the observer. To that end we propose a model that predicts the visual significance of the image contrast shown on display devices, that in addition to the fundamental spatial aspects of human vision, also accounts for maladaptation over time.

Purely mathematical image quality assessment metrics that are limited to LDR content can be extended to support HDR images by means of simple human visual system models. We develop a transfer function to a “perceptually uniform space”, that transforms image luminance to perceptually linear units along the entire visible luminance range. The quality metrics are executed on the perceptually uniform images, resulting in meaningful predictions for HDR content, as well as backward compatible quality outcomes for LDR images.

While HDR imaging is gaining momentum, the transition has been not immediate; currently both HDR and LDR content are in use. In terms of image quality assessment, this raises an important issue: quality metrics are built on the assumption that the input reference-test image pair has the same dynamic range. We address this shortcoming by proposing a “dynamic range independent” image quality assessment method, that can handle all possible dynamic range combinations of the reference-test image pair. This has been achieved using an HDR human visual system model in conjunction with three novel distortion measures. Our work enables for the first time the objective evaluation of tone mapping operators, among other novel applications.

The same inhomogeneous dynamic range content problem is also present for video sequences. Similarly, we propose a dynamic range independent video quality assessment method, where we address temporal aspects of visual perception. We show that such a metric is useful in objective evaluation of rendering methods, the assessment of HDR compression artifacts, as well as comparison of temporal tone mapping approaches. We also discuss in detail how to validate such a metric, and show that its predictions are more accurate than other video quality assessment techniques.

In summary, the proposed methods demonstrate different approaches to designing application-specific human visual system models, and show that one can extend and improve the state-of-the-art through the use of such models.

Zusammenfassung

In der vorliegenden Dissertation untersuchen wir die Verwendung von Modellen des menschlichen visuellen Systems im Bereich der Computergrafik. Wir entwickeln Modelle unterschiedlicher Komplexität für verschiedene Anwendungsbereiche, die einerseits die Grundlage für die von uns vorgestellten neuen Techniken bilden, und andererseits dazu dienen, auf den momentanen Stand der Technik aufzubauen. Der theoretische Teil dieser Dissertation umfasst mehrere psychophysischen Experimente zur Kalibrierung und Validierung dieser Modelle, um eine Anpassung an die Wahrnehmung eines durchschnittlichen Beobachters zu erzielen.

Wir entwickeln ein vereinfachtes Modell der menschlichen visuellen Wahrnehmung, das Helligkeitsanpassung und visuelle Maskierung berücksichtigt, und integrieren es in ein waveletbasiertes, kantenerhaltendes Image-Decomposition-Framework der zweiten Generation. Die Vorhersage der visuellen Signifikanz durch das Wahrnehmungsmodell ersetzt dabei die Grsse des Gradienten als Mass für die Kantenstärke, ohne dabei signifikanten Rechenaufwand zu erfordern. Wir zeigen, dass das erweiterte Framework weitaus intuitiver für kantenerhaltendes Glätten und Kontrastverbesserung ist, und dass damit qualitative Verbesserungen der Resultate von HDR-Image-Retargeting, Tonemapping und HDR-Panorama-Stitching erzielt werden knnen.

Obwohl es viele Forschungsarbeiten mit dem Schwerpunkt auf der Erzeugung plausibler Bilder gibt, gibt es wenige Untersuchungen darüber, wie diese Inhalte dann, dargestellt auf den Endgeräten, tatsächlich wahrgenommen werden. Durch die zunehmende Ausbreitung mobiler Geräte ist auch längst nicht mehr sichergestellt, dass das Bildmaterial auf einem Desktop-Bildschirm unter kontrollierten Lichtverhältnissen betrachtet wird. Die Auswirkung sich dynamisch ändernder Lichtverhältnisse auf die Wahrnehmung des Beobachters sollte deswegen berücksichtigt werden. Wir schlagen daher ein Modell vor, dass die visuelle Signifikanz des Bildkontrasts auf dem mobilen Gerät vorhersagt, und zusätzlich zu den fundamentalen Aspekten der menschlichen visuellen Wahrnehmung auch mangelhafte Anpassung über die Zeit berücksichtigt.

Rein mathematische Metriken zur Bewertung der Bildqualität, deren Anwendungsgebiet auf LDR-Inhalte beschränkt ist, knnen durch einfach Modelle des menschlichen visuellen Systems auf die Verarbeitung von HDR-Bildern erweitert werden. Wir entwickeln eine Transferfunktion in einen "wahrnehmungstechnisch gleichfrmigen Raum", durch die Helligkeiten zu wahrnehmungstechnisch linearen Einheiten entlang des gesamten sichtbaren Helligkeitsbereichs umgewandelt werden. Da die Qualitätsmetriken dann auf die transformierten Bilder angewendet werden, ergeben sich einerseits für HDR-Inhalte, und andererseits durch die Rückwärtskompatibilität auch für LDR-Inhalte aussagekräftige Vorhersagen für die Bildqualität.

Obwohl die HDR-Bildgebung langsam an Popularität gewinnt, hat sich kein abrupter Übergang von LDR zu HDR ereignet, weshalb Inhalte beider Formate genutzt werden. Aus Sicht der Bestimmung der Bildqualität ergibt sich dadurch ein Problem: Beim Einsatz von Qualitätsmetriken wird üblicherweise davon ausgegangen, dass Eingabe- und Referenzbild denselben Dynamikumfang

aufweisen. Wir beheben diesen Mangel durch die Vorstellung einer vom Dynamikumfang unabhängigen Methode zur Bewertung der Bildqualität, die alle mglichen Kombinationen des Dynamikumfangs von Eingabe- und Referenzbild berücksichtigt. Erreicht wird das durch die Verwendung eines HDR-Modells des menschlichen visuellen Systems in Verbindung mit drei neuen Massen für die Bildverzerrung. Durch unsere Arbeit wird es unter anderem zum ersten Mal mglich, Tonemapping-Operatoren objektiv zu evaluieren.

Dasselbe Problem von Inhalten mit inhomogenen Dynamikumfang tritt auch bei Videosequenzen auf. Wir stellen daher eine dynamikumfangsunabhängige Methode zur Bewertung der Videoqualität vor, wobei wir die temporalen Aspekte der visuellen Wahrnehmung berücksichtigen. Eine solche Metrik ist sowohl für die objektive Evaluierung von Rendermethoden, die Beurteilung von HDR-Kompressionsartefakten, als auch zum Vergleich temporaler Tonemapping-Ansätze sinnvoll. Ferner zeigen wir auch, wie bei der Bewertung solcher Metriken zu verfahren ist, und dass die aus ihnen resultierenden Vorhersagen genauer als andere Methoden zur Qualitätsbeurteilung von Videos sind.

Zusammenfassend kann man sagen, dass die vorgestellten Methoden dazu verwendet werden knnen, anwendungsspezifische Modelle des menschlichen visuellen Systems zu erzeugen, und dass durch die Verwendung dieser Modelle wesentliche Verbesserungen und Erweiterungen gegenüber dem momentanen Stand der Technik erzielt werden knnen.

Acknowledgements

First and foremost I would like to thank my advisor Dr.-Ing Habil. Karol Myszkowski, who introduced me to the field of visual perception and its applications in computer graphics. I am grateful to him for his scientific contribution, as well as allowing me to pursue my own ideas and patiently supporting me during the process.

None of this would have been possible without the outstanding working environment provided by my other advisor Prof. Dr. Hans-Peter Seidel. Thanks to him, during my stay in Max Planck Institute I was able to focus solely on my research without any major distractions.

I would like to thank my co-authors Dr. Martin Čadík, Dr. Rafał Mantiuk, Dawid Pająk and Assist. Prof. Dr. Makoto Okabe for their valuable contributions to the research presented in this thesis. I would also like to single out Martin and Rafał for their influence on me as a researcher.

I thank all past and present researchers of the Computer Graphics Group in Max Planck Institute for their help and advice on countless occasions, as well as Sabine Budde, Ellen Fries, Conny Liegl and the rest of the administrative staff for organizing my travels and helping me with other issues.

I would like to thank the members of my thesis committee: Dr.-Ing Habil. Karol Myszkowski, Prof. Dr. Hans-Peter Seidel, Assoc. Prof. Dr. Jan Kautz, Prof. Dr. Phillip Slusallek and Dr.-Ing. Thorsten Thormählen for reviewing my thesis and/or participating in the defense.

Finally, I thank my mother Prof. Dr. Süheylâ Aydın, father Vahît Aydın, my fiancée İrem Dumlupınar and other family and friends for their love and continued support. Prof. Aydın has also been involved in proofreading the manuscript on multiple stages of this work.

To all the great musicians of the 90's.

Contents

Chapter 1

Introduction

"Never start with a clear idea of a storyline. Instead, commence blindly, with a vague notion of trying to include a reference to your favourite band, gift shop, or chocolate bar"
Alan C. Martin

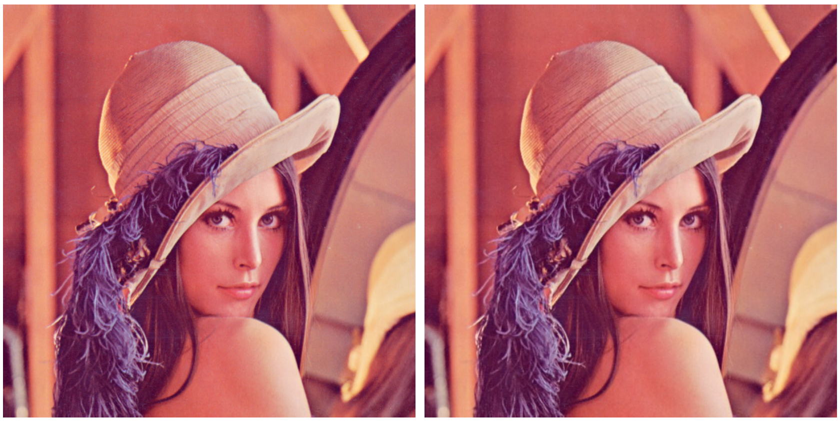


Figure 1.1: *Lossless versus lossy compression. Images can be significantly compressed without visible artifacts by exploiting limitations of the human visual system.*

Can you spot the differences between the two images in Figure 1.1? One of these images is subjected to the lossy JPEG compression, whereas the other one is stored in the lossless TIFF format. The JPEG image contains merely $1/12th$ of the information stored in the TIFF image, yet to a human observer they look very similar, if not the same. How can we remove so much information from the image without producing visible artifacts?

JPEG format achieves such a high compression rate by incorporating a simple model of the *human visual system*. The term human visual system denotes the cascade of components starting with the eye that captures reflected light from objects in a scene. The consecutive neural machinery converts light into electrical signals and relays these signals to the brain (we briefly investigate these mechanisms in Chapter 2). The initial *visual data*, that is the incoming light, is filtered, and in some cases distorted by multiple mechanisms that make up the human visual system. The JPEG compression takes advantage of this by removing image details that are not visible due to the structure of the visual system. Therefore Figure 1.1-right looks the same as Figure 1.1-left, even though it contains only a fraction of the data in the left image. In other words, the compression removes the details C_1 from the image I, such that:

$$hvs(I - C_1) \equiv hvs(I), \tag{1.1}$$

where *hvs* is a hypothetical function that simulates the human visual system. This basic principle has been a guideline for efficient acquisition and displaying of visual data.

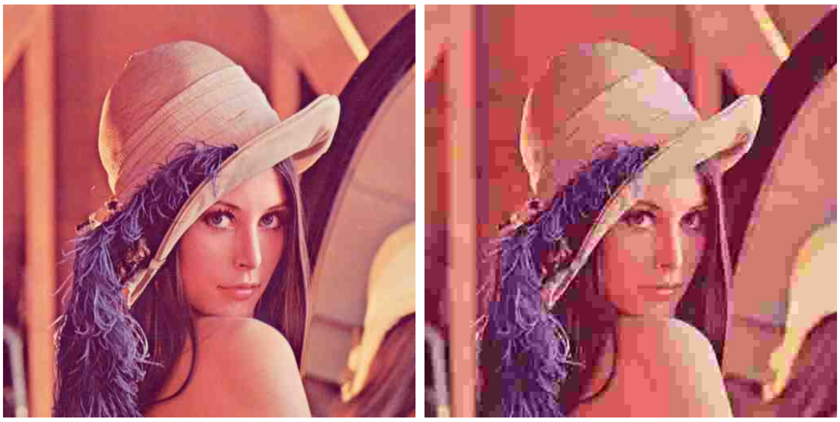


Figure 1.2: *Comparison between low (left) and high (right) compression rates.*

Increasing the compression ratio to 1 : 55 causes artifacts start appearing, as shown in Figure 1.2-left. Thus for the details C_2 removed at this compression level the relation changes to:

$$hvs(I - C_2) \not\equiv hvs(I). \tag{1.2}$$

In many computer graphics applications such visible artifacts are not acceptable, thus their *detection* is an important practical concern. Generally speaking, the detection problem consists of predicting the probability of an average user distinguishing a detail from its background, or analogously finding the point where the perception shifts from Equation 1.1 to Equation 1.2. As for the detail in focus, one can consider a distortion as in the compression case, or simply a low contrast scene detail with a magnitude near the threshold of visibility.

Conceptually, a detection framework for compression artifacts would work as follows: given the *reference* image I and the distorted *test* image $I - C_2$, we can compute the hypothetical human visual system response to the subtracted details C_2:

$$hvs(I) - hvs(I - C_2) = hvs(I) - [hvs(I) - hvs(C_2)] = hvs(C_2). \qquad (1.3)$$

In fact, the formulation above is the basis of most sophisticated *Image quality assessment metrics* that involve human visual system models. Once $hvs(C_2)$ is computed, one can determine the probability of the compression artifacts being visible through a *psychometric function* that relates perceived contrast to detection probability.

It is important to note that the formulation in Equation 1.3 assumes that the superposition principle, that is $f(x + y) = f(x) + f(y)$, holds in our system. Simplifications of this nature are common in human visual system modeling; in general assuming that the human visual system is a cascade of linear mechanisms greatly simplifies the models of the otherwise very complex and not entirely understood system [Wandell, 1995]. From a practical standpoint Equation 1.3 can still be useful if we limit our scope to a very small region near the detection threshold, where the difference between I and $I - C_2$ is hardly noticeable and thus the parameters governing the human visual system model are almost the same. Such *near-threshold models* provide a nice trade-off between complexity and scope, since in many applications accurate prediction near the detection threshold is sufficient.

Continuing with our example, Figure 1.2-right shows an aggressive 1:180 compression. At this rate the file size of the 512 × 512 image is merely 4.5K, but as a side effect the compression artifacts are highly visible. In this case, predicting the detection probability of distortions would not make sense, since the compression artifacts are strongly visible everywhere in the image. A more relevant concern in this case is the *visual significance* of these visible distortions, as a measure of their effect on quality. This problem, however, forces our human visual system model to make predictions well above the visibility threshold, where the model prediction is not accurate anymore. Thus, for the visual significance task one should use *supra-threshold* models which include an additional nonlinearity that approximates the contrast perception above the threshold. This supra-threshold nonlinearity is often modeled using a *transducer function*.

The distinction between near– and supra-threshold models are often confusing to the reader, since our visual system is able to seamlessly perform both tasks. In practice, near-threshold models focus on carefully modelling the human visual system mechanisms near the threshold, while the supra-threshold models focus more on predicting the contrast perception non-linearity above the threshold. While both types of models are based on the same physiological and psychophysical data, they differ in their focus, and consequently their simplifying assumptions. Both approaches are useful depending on the application needs.

1.1 Motivation

A wide range of computer graphics and computer vision methods can potentially benefit from human visual system models. We have already mentioned the extremely high image compression rates achievable without any visible artifacts. The same principle also applies to video compression, where one can additionally exploit the temporal aspects of visual perception. Similarly, for compression of *High Dynamic Range (HDR)* images and videos, the limitations of the human visual system are much more pronounced, and thus can greatly be taken advantage of.

Supra-threshold models have been used in HDR contrast manipulation applications such as forward– and inverse tone mapping, color to gray, color reconstruction, and others. Here, the central idea is to transform physical contrast to the "perceived" contrast which is linear in terms of perceived strength. The perceived contrast can then easily be manipulated, for example using a single multiplier, to achieve a perceptually uniform effect on the entire image or video. The perceived contrast is also a very convenient measure for visual significance of image features such as edges. One can achieve better results in applications that make use of the strength of image features, such as image retargeting and panorama stitching, by simply replacing the arbitrary feature strength measure with the perceived contrast computed by a human visual system model.

Human visual system models have found their place also in rendering. The main principle of perceptually driven rendering methods is rendering only those details that are visible to the human eye. This way one can render far less scene details without producing any visible differences. Consequently, provided that the model that predicts visibility is fast, rendering becomes more efficient.

Perhaps the application area of human visual system modelling with the largest impact is image and video quality assessment. The contributions of newly proposed computer graphics and computer vision techniques are usually demonstrated through images and videos, in which the merit of the technique is apparent. The performance of, for example a new rendering method, can be assessed by comparing sequences rendered on one hand using the proposed method, and on the other hand a more precise, but slower reference method. The point of this comparison could be to show that the proposed method produces results comparable to the reference method, but much more efficiently. Such a comparison should ideally be performed using subjective studies. However subjective studies are often costly in terms of time and resources, and require expertise in vision science. On the other hand, objective image and video quality assessment metrics are useful practical tools that could predict the outcome of such a subjective experiment automatically without the need for additional expertise or resources.

1.2 Problem Statement

In this dissertation we discuss how to design and implement human visual system models with various complexities and how to integrate them into computer graphics methods to improve their performance or expand their area of application.

The major focus of this dissertation is the improvement on multiple state-of-the-art computer graphics methods through the modeling of visual perception.

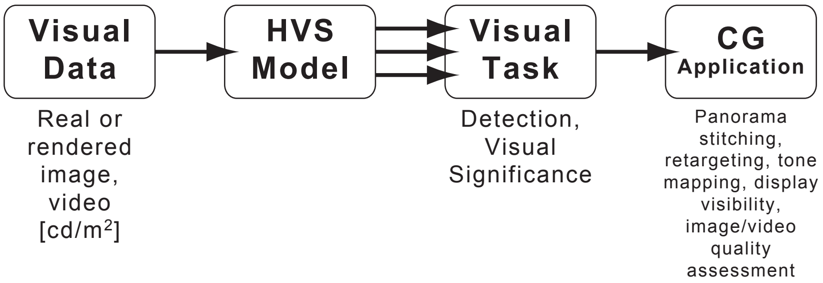

Visual Data
HVS Model
Visual Task
CG Application
Real or rendered image, video [cd/m^2]
Detection, Visual Significance
Panorama stitching, retargeting, tone mapping, display visibility, image/video quality assessment

Figure 1.3: *Generic data flow of the methods proposed in this thesis.*

To that end we propose a multitude of human visual system models geared towards performance or precision, with near– or supra threshold focus, taking into account either the spatial or spatiotemporal aspect of visual perception, and integrate these models to visual significance and detection tasks. Figure 1.3 shows the generic data flow of the proposed methods in this thesis. The visual data, either image or video, is processed by a human visual system model to compute perceived contrast, which is used to predict detection probability or visual significance depending on the application needs.

A quantitative measure for the significance of prominent image features such as edges is a requirement of great practical importance, since many image editing methods make use of such a measure. We devote our effort to design an efficient human visual system model that predicts the perceived visual significance of image edges. The prediction of visual significance of image contrast shown on a display brings in further considerations such as the sudden changes adaptation state of the observer and reflections due to dynamically changing illumination conditions. These visual significance problems are addressed in the first part of this thesis. In the second part of the thesis we investigate quality assessment in the image domain. An interesting problem relevant to computer graphics applications is automatically predicting the magnitude of visible differences between images, especially when the visual data have different dynamic ranges. We also look ways to modify existing simple image quality assessment metrics to be compatible with HDR content. The third part of the thesis focuses on quality assessment in video domain where the main challenge is the modeling of spatiotemporal characteristics of the human visual system. In parallel to the theoretical work and resulting computational models of human vision, it is also important to calibrate and validate the models using psychophysical experimentation. Accordingly, the third part also includes an in depth discussion of such a study on video stimuli.

1.3 Main Contributions

Parts of this dissertation have been published in varius venues [Aydın *et al.*, 2008b; 2008a; 2009; 2010b; 2010a]. This thesis unites these publications under the context of human visual system modeling while presenting improvements and updated results.

The investigation of the aforementioned problems resulted in the following main contributions:

- A method for estimating the visual significance of image edges, that conveniently replaces the widely used gradient magnitude measure and results in qualitative improvements in HDR image retargeting, tone mapping and panorama stitching applications.
- A metric for predicting the effect of the observer's maladaptation and reflections due to dynamically changing lighting conditions on the visibility of display devices.
- A transfer function that transforms image luminance to a "perceptually uniform space", enabling simple image quality measures, such as PSNR and SSIM, to work on HDR images. This method is also backwards compatible in the sense that the metric response for LDR images transformed to the proposed space remains approximately the same.
- An image quality assessment method that can compare LDR images with respect to an HDR reference, and vice versa, in addition to reference-test image pairs of the same dynamic range. This method enables, for the first time, the objective evaluation of forward and inverse tone mapping operators.
- A dynamic range video quality assessment metric comprising spatiotemporal aspects of visual perception. This metric enables objective evaluation of a large number of computer graphics methods such as rendering, compression and temporal tone mapping.

1.4 Chapter Organization

In the next chapter of this dissertation we present an introduction to human visual sytem modeling, which is meant to serve as the background for the following three parts. In the first part of this thesis we investigate two visual significance problems. In Chapter 3, we show the use of visual significance in image edge weighting, computed through a simplistic human visual system model geared towards computational efficiency, improves results of various techniques relying on edge strength computation. In Chapter 4 we propose a more sophisticated model including temporal aspect of luminance adaptation for predicting display visibility under dynamically changing lighting. The second part comprises two detection problems in the context of image quality assessment. Chapter 5 introduces a practical extension of a pair of simple quality measures, PSNR and SSIM, to HDR imaging through the use of a simple transfer function that accounts for the nonlinear photoreceptor response to luminance. A more comprehensive human visual system model is coupled with three novel distortion measures in Chapter 6 enabling image quality assessment where the reference and test images have different dynamic ranges. In the final part of this thesis we investigate temporal aspects of the human visual system in the context of video quality assessment. Chapter 7 introduces a video fidelity metric and demonstrates its applications to rendering, HDR compression and temporal tone mapping. Finally, in Chapter 8 we elaborate on the psychophysical validation study of the metric from Chapter 7.

Chapter 2

Background on Human Visual System Models

In this chapter we give a brief overview of the human visual system's relevant mechanisms and discuss the approaches on modeling them. The models outlined here result from decades of psychophysical and physiological studies. Even though the exact relations between the anatomical structures of the human visual system and the various aspects of human vision is currently not known, whenever possible, we make an effort to justify the proposed models with the corresponding physiological findings.

The aspects of human vision related to color perception are omitted in this section, since all models in this thesis are luminance based. For such omitted mechanisms, as well as for a more in-depth treatment of the mechanisms discussed in this section we refer the reader to the excellent book by Wandell [1995].

In the reminder of this chapter we discuss certain characteristics of the human visual system that have significant influence on visual perception, such as glare due to the eye's optics, luminance adaptation, contrast sensitivity, frequency and orientation selective visual channels and visual masking. These characteristics and corresponding models will become relevant in the following three parts of this thesis where we explore them further in computer graphics context. We also discuss contrast and its multiple interpretations found in the literature. Finally we define two fundamental problems, visual significance and detection, that will serve a basis for more complex methods we present in the following parts of the thesis.

2.1 Optics of the Eye

Eyes are the entry points of the light that carries the visual information about the observer's surroundings into the human visual system. The incoming light is absorbed by the photoreceptors in the retina and converted to electrochemical signals, and these signals are relayed to the consequent mechanisms of the visual pipeline. The image that falls onto the retina is not an exact copy of the real world image; as in every optical system, the eye distorts the light while it passes through. The combined effect of the scattering and diffraction within the optical component of the human visual system is referred to as *glare*. The glare effect is most obvious near bright light sources in otherwise dark scenes, such as the candle shown in Figure 2.1.

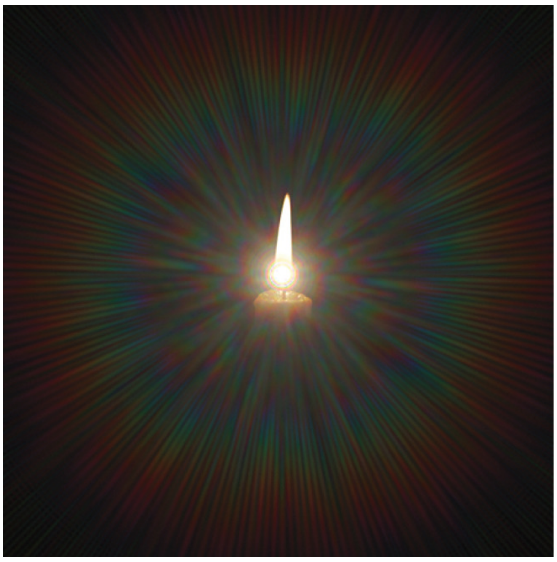


Figure 2.1: *A faithfull simulation of the glare effect cite, courtesy of Tobias Ritschel.*

A closer inspection of Figure 2.1 reveals that the glare effect, rather than being homogeneous, is composed of three main components. The radial streaks emanating from the bottom of the candle fire are called the *ciliary corona*. The ciliary corona is caused by the semi-random density fluctuation due to the motion of the particles in the lens and vitreous humor (Illustrated as black dots in Figure 2.2). The colorful ring around the light source is known as *lenticular halo*, caused by the circular grating formed by the radial fibers at the periphery of the lens (Figure 2.2, see the lens inset). Light only passes through these gratings under darker illumination conditions when the pupil diameter is greater than $3mm$. Thus, during daylight (pupil diameter is $\sim 2mm$) no lenticular halo is observed. The decrease of perceived contrast near the candle light is referred to as *blooming* (or disability glare, veiling luminance). This effect is attributed to light scattering in eyelashes, cornea ($25-30\%$), lens (40%), iris ($< 1\%$), vitreous humor (10%) and retina (20%), where the relative contribution of each eye component is denoted in paranthesis [Ritschel *et al.*, 2009b]. Additionally, if the size of the light source is large, the ciliary corona can blur and contribute to blooming as well [Spencer *et al.*, 1995].

An approach to modelling the glare effect is convolving the scene luminance with a 2D spatial filter that approximates the light scattering in the eye [Nakamae *et al.*, 1990; Spencer *et al.*, 1995]. These filters can be thought as the *point spread*

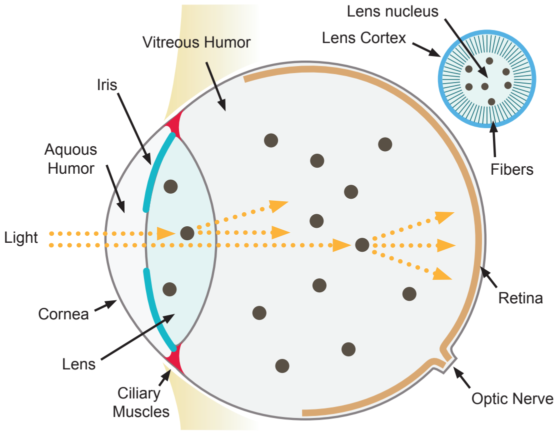


Figure 2.2: *Components of the eye that are involved in the forming of glare, courtesy of Tobias Ritschel.*

function (PSF) of the eye, that describes the blurring of the retinal image of a point source in focus, or analogously the probability that a photon will appear at a given location on retina. Spencer et al. [1995] model the radial streaks of the ciliary corona by introducing random antialiased lines to the PSF filter. Ritschel et al. [2009b] on the other hand simulate particles inside the lens and vitreous humor, along with other dynamically changing properties such as the blink state, field luminance and observer motion. The Fourier Transform of the resulting PSF is then multiplied with the Fourier Transform of the scene luminance, and the result is transformed to the spatial domain. One can also design the 2D filter directly in the frequency domain in the form of an *optical transfer function (OTF)* [Deeley *et al.*, 1991; Marimont and Wandell, 1994].

The glare effects mentioned so far are all functionally undesired in the sense that they limit visual acuity. However, an interesting side effect of blooming is the local increase in perceived contrast, that is: while details near a bright light source are harder to detect, the light source itself appears brighter than it would without the blurring near its periphery. A recent psychophysical study shows that by introducing even a very primitive blooming pattern, one can increase the perceived luminance by $20 - 35\%$ [Yoshida *et al.*, 2008].

2.2 Luminance Adaptation

The scene luminance that falls onto the retina may differ by 14 orders of magnitude from a moonless night to a cloudless sunny day. The magnitude of the electrical signals produced by the retinal photoreceptors on the other hand only vary from a millivolt to tens of millivolts. This suggests that either the photoreceptors are sensitive to even the smallest changes in electrical current to the point that the 14 orders of magnitude range can be encoded within approximately 2 orders of magnitude, or that the visual information is subject to some kind of lossy compression in the retina. As often is the case with the the human visual system, in this instance precision is traded off for more efficiency. While we are able to see the full 14 orders of magnitude, at any given time we are mostly sensitive to ∼3 orders of magnitude near the current *adaptation* level. The sensitivity to luminance levels outside this range will be very low. Thus, we don't see the stars in daylight, and our eyes need a second or two to adjust when walking out of a movie theater.

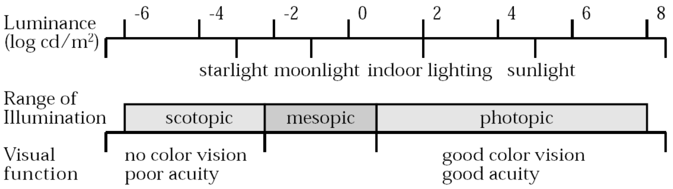


Figure 2.3: *Three different modes of vision and corresponding adaptation levels. Adopted from Ferwerda et al. [1996].*

The retina is between 0.3 mm and 0.5 mm in thickness and is composed of about 100 million *rods* and 5 million *cones.* Rods are very sensitive to light, but are achromatic and provide limited pattern sensitivity. There are three types of cones that are sensitive to short, middle and long wavelengths, and collectively they cover the range of the spectrum from 400 nm to 700 nm. The two types of photoreceptors operate in parallel and the luminance ranges to which they are sensitive complement each other. Figure 2.3 illustrates the luminance ranges where rods and cones are dominant. Rods dominate the low luminance *scotopic* vision, whereas cones are much more sensitive in the *photopic* range. As a result, we enjoy high visual acuity and color perception under indoor lighting or sunlight, whereas during the night we are sensitive to even the slightest luminance differences. Within the *mesopic* range that falls between scotopic and photopic ranges, our vision is a combination of these two mechanisms.

For a given *adaptation level* the photoreceptor response to luminance is nonlinear roughly in the form of an S-shaped curve. The curve is centered at the current adaptation level, and moving away from the center it exhibits a compressive behaviour. This means that the sensitivity is highest for scene luminance same as the current adaptation level, in other words the visual system is *adapted* to that scene luminance. On the other hand, the photoreceptor response for all other luminance levels except the adaptation luminance is compressed, and thus the observer is *maladapted* to those luminances. The compression level is still relatively low within the 2-3 orders of magnitude range around the adaptation luminance, thus we see this range well. The photoreceptor nonlinearity model proposed by Naka-Rushton [1966] is as follows:

$$\frac{R}{R_{max}} = \frac{L^n}{L^n + \sigma^n}, \tag{2.1}$$

where R is the photoreceptor response, R_{max} is the maximum response, L is the luminance falling onto the retina, σ is the half-saturation constant depending on the current adaptation state, and n is a constant that controls sensitivity and typically varies between 0.7 and 1. Note that the adaptation state is not uniform across the retina, and thus the half-saturation constant has to be computed locally.

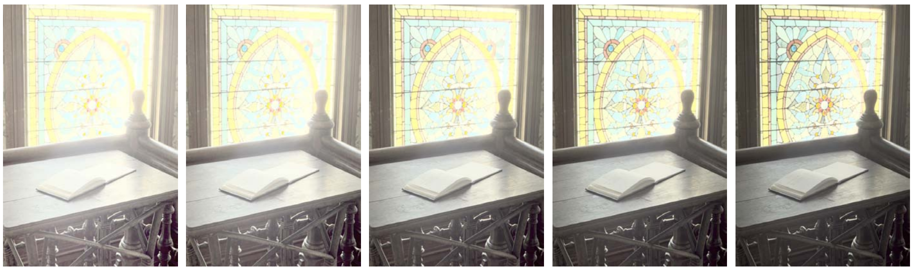


Figure 2.4: *The effect of luminance adaptation over time, a simulation of the fast adaptation from a dark environment ($10^{-4}cd/m^2$) to the stained glass ($17cd/m^2$). Columns from left to right:* $t = 0.01$s, $t = 0.02$s, $t = 0.05$s, $t = 0.1$s, $t = 60$s *(fully adapted state).*

Adaptation is a dynamic mechanism; if the illumination conditions change, so does the adaptation state as a result of mechanical, photochemical and neural processes. The most obvious adaptation mechanism is the change of the pupil size: under bright illumination the amount of incoming light is reduced by decreasing the pupil diameter (down to ∼2 *mm*), and vice versa under dark illumination (up to 7-8 *mm*). The maximum regulatory effect of this mechanism is a little more than an order of magnitude. More significant are the relatively slow photochemical processes: *bleaching* and *regeneration* and the fast *neural processes*. Bleaching occurs when exposed to a bright intensity, the photosensitive pigments in the photoreceptors are depleted faster than they are regenerated, which decreases the sensitivity at these intensities. These photochemical processes are not symmetrical, which is the primary reason for the difference of the time course between dark and bright adaptation. The neural processes are on the other hand symmetrical, and are due to the saturation of the photopigments subjected to excess light intensities. Figure 2.4 shows a simulation of dark adaptation over time.

Due to the multitude of mechanisms governing the current adaptation level, practical models of adaptation mechanisms are relatively complex [Ferwerda *et al.*, 1996; Pattanaik *et al.*, 2000; Irawan *et al.*, 2005]. An alternative practical approach is assuming that the eye is capable of adapting to a small area (such as a pixel). In terms of the model, it is like for each pixel of an image, the observer's adapted to exactly the luminance of that pixel, thus disregards maladaptation. With this assumption and taking $n = 1$, Daly [1998] proposes a simplification of Equation 2.1:

$$\frac{R}{R_{max}} = \frac{L}{L + c\sigma^b}, \tag{2.2}$$

where c and b are constants.

Using the same assumption one can also derive a *threshold versus intensity (tvi)* function, which gives the minimum luminance difference that can be noticed on a background luminance, assuming that the eye is adapted to the background luminance (Figure 2.5). This simple function behaves as a power function in low luminance levels and as logarithmic function in high luminance levels. The *tvi* function is highly useful in practice; the nonlinearity of color spaces such as sRGB and CIE $L^*u^*v^*$ mimic the *tvi* function for encoding efficiency. An-

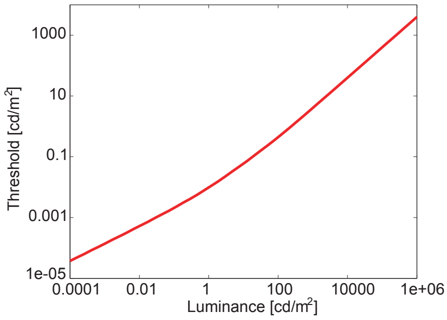


Figure 2.5: *The threshold versus intensity (tvi) function is approximately linear on a log-log plot.*

other practically useful tool is a mapping from the luminance to the number of thresholds corresponding to that luminance [Mantiuk *et al.*, 2005] (Equation 10.1). Irawan et al. [2005] proposed the generalized *threshold versus intensity and adaptation (tvia)* function, where they also take maladaptation into account. However the domain of this function is two-dimensional (retinal luminance and adaptation luminance) and thus is more complex. In Chapter 4 of this thesis we propose a display visibility metric that makes use of the *tvia* function.

2.3 Contrast Sensitivity

The signals produced by the photoreceptors leave the retina through the axons of the retinal ganglion cells. These axons comprise the optic nerve, and exit from the retina at a single location called the optic disk. Through the optic nerve, the visual data is relayed to the neurons in lateral geniculate nucleus (LGN) and primary visual cortex. Each of these neurons along the visual pathway have a *receptive field*: an area in the retina which influences the neuron's response. A crucial property of the receptive fields is that the influence of the receptive field's center is the exact opposite of the influence of the receptive field's surround. About half of the retinal ganglion cells are inhibited at the center, and excited at the surround (off-center, on-surround), and the remaining half behaves vice versa (on-center, off-surround).

As a consequence of the center-surround structure of the receptive fields, the neurons in the LGN are "tuned" for the range of spatial frequencies that matches the size of their receptive fields. Figure 2.6 illustrates the neuronal response to cosinusoidal stimuli with various spatial frequencies. In the first case, the spatial frequency is low, and the light falling on the entire receptive field is nearly constant. As a result the neuron's response will be low. In the second case, the spatial frequency is high, and as a result both positive and negative parts of the cosinusoidal stimulus fall onto both the excitatory and inhibitory regions, effectively cancelling each other out. The third case shows that the highest response is generated when the size of the grating matches a single region of

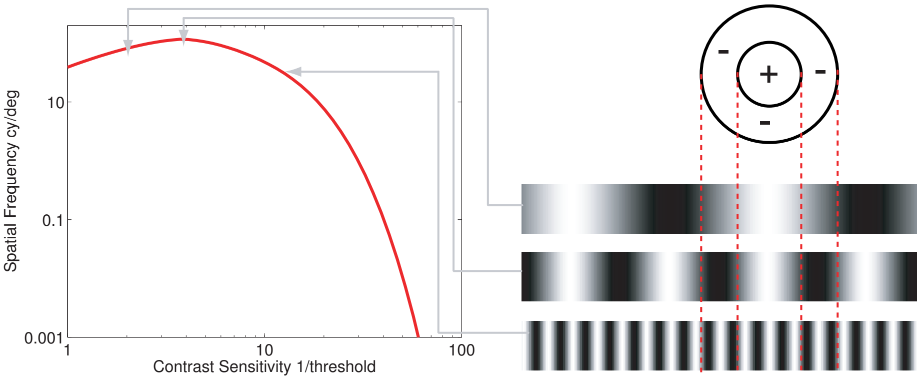


Figure 2.6: *The parts of the cosine stimuli at various frequencies that fall onto a receptive field in LGN (right) . Even though the amplitude of all three stimuli are the same, due to the center-surround structure of the receptive field the neuron's response varies. The neuronal response can be plotted as the contrast sensitivity function (left).*

the receptive field. The overall change in sensitivity with respect to spatial frequency is plotted in Figure 2.6-left, and is known as the *contrast sensitivity function (CSF)*.

From a computational point of view the CSF describes the sensitivity to harmonic stimuli as a function of spatial and temporal frequencies, where the sensitivity is defined as the inverse of the threshold Michelson contrast (Equation 2.4). The threshold contrast depends on many factors such as the background (adaptation) luminance, the grating's spatial frequency, orientation, spatial extent, and eccentricity with respect to the fovea. Consequently, popular CSF models [Daly, 1993; Barten, 1999] have multitude of input parameters. For HDR imaging, Daly's CSF [1993] as a part of the Visible Differences Predictor (VDP) is found to produce better predictions, especially in scotopic range and for adaptation levels greater than 1000 cd/m^2 (Equation 10.3). Kelly [1983] proposed a chromatic and achromatic spatiotemporal CSF, which has been improved later by accounting for the movements of the eye [Daly, 1998] (Equation 10.5). A disadvantage of spatiotemporal CSFs is the lack of a luminance adaptation model. In Chapter 7 we discuss the temporal aspects of contrast sensitivity in more detail, and show how one can incorporate luminance adaptation to a spatiotemporal CSF. In Figure 2.7 we show the effect of maladaptation to the shape of the CSF.

An important point to note is that the psychophysical studies to determine the CSF are performed on near-threshold stimuli. In supra-threshold contrast region, the CSF tends to become flatter, meaning that the human visual system becomes equally sensitive to all visible frequencies. This property is known as *contrast constancy* [Georgeson and Sullivan, 1975].

There are two approaches to implementing the CSF, as a weighting function for each visual channel [Lubin, 1993; Winkler, 2005] which offers less precision, or as a filter in frequency domain [Daly, 1993] which offers better precision but is computationally less efficient and assumes that the filter is shift invariant. Local adaptation can efficiently be approximated by interpolating between a limited number of CSF functions with logarithmically spaced adaptation luminances [Mantiuk *et al.*, 2005].

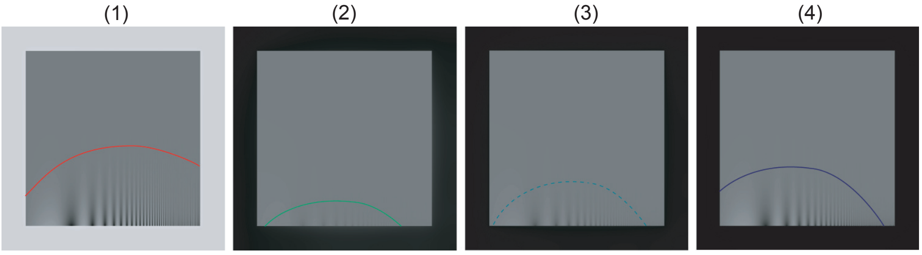


Figure 2.7: *Classical Campbell-Robson contrast sensitivity chart for dark adaptation. From left to right: (1) fully adapted state in a relatively bright environment (adaptation luminance 112 cd/m^2), (2) background luminance was decreased to 3 cd/m^2, the contrast sensitivity moves to lower frequencies, but due to maladaptation, it is basically very low, (3) sensitivity regenerates according to dark adaptation time-course, (4) final fully adapted state (adaptation luminance 3 cd/m^2). The curves show the author's thresholds observed from approximately 30 centimeters at original paper size.*

2.4 Channel Decomposition

The receptive fields of the simple neurons in the primary visual cortex differ from the receptive fields of the LGN neurons, in that they are selective to certain spatial frequencies and orientations. Figure 2.8 shows hypothetical receptive fields of the cortical neurons. As the right figure shows, the elliptical receptive fields generate a stronger response if the stimulus has exactly their preferred orientation.

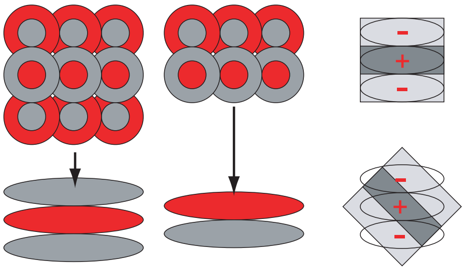


Figure 2.8: *Receptive fields in the primary visual cortex. The neighboring circular receptive fields (left, center) together form elliptical shapes with a certain orientation preference (right).*

The main difference between various approaches to modelling this mechanism is the tradeoff between, on one hand physiological plausibility, and on the other hand theoretical simplicity and computational efficiency. At the former end

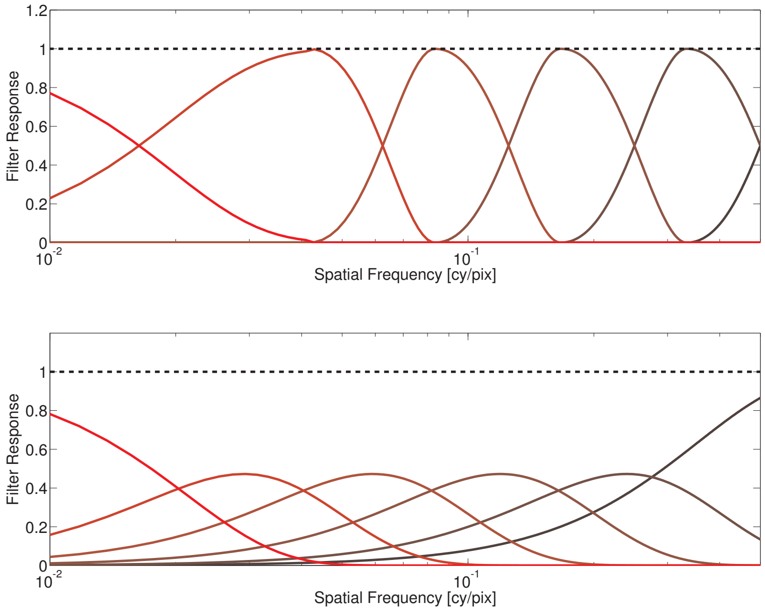


Figure 2.9: *The spatial frequency separation of the Cortex Transform (top), and Laplacian Pyramid (bottom). Note that every cortex band only contains a minor amount of contrast at the immediate neighboring bands. Laplacian Pyramid levels on the other hand have a much larger support in the frequency axis.*

of the spectrum are the Gabor filter banks that faithfully model the on/off structure of the receptive fields, however they are non-invertible and costly to compute. On the efficiency and simplicity end of the spectrum is the Laplacian Pyramid [Burt and Adelson, 1983]. It is also relatively simple to implement orientations by "steering" the pyramid [Freeman and Adelson, 1991]. However, the spatial frequency separation of the Laplacian Pyramid is low: each pyramid level receives a notable contribution from spatial frequencies other that the frequency corresponding to that pyramid level. Similarly, wavelet based decompositions are extremely fast, and recently Fattal [2009] demonstrated their use in computer graphics applications. Like the Laplacian Pyramid, wavelet based decompositions are multi-purpose tools not necessarily geared towards modelling the primary visual cortex. To that end, the Cortex Transform [Watson, 1987] offers a nice trade-off between physiological plausibility and practicality, in that it is invertible, has orientations and the frequency separation is high (refer to Section 10.4 for the derivation). Figure 2.9 shows a comparison of the frequency selectivity of the Laplacian Pyramid and Cortex Transform. Parts of the Cortex Transform of an example image is shown in Figure 2.10. A closer inspection of the Figure 2.8 left and center shows that the receptive fields of the cortical neurons can correspond to either even or odd functions. This indicates to a shortcoming of both the Cortex Transform and Laplacian Pyramid, whose responses closely resemble that of the even filters. In practice, this means

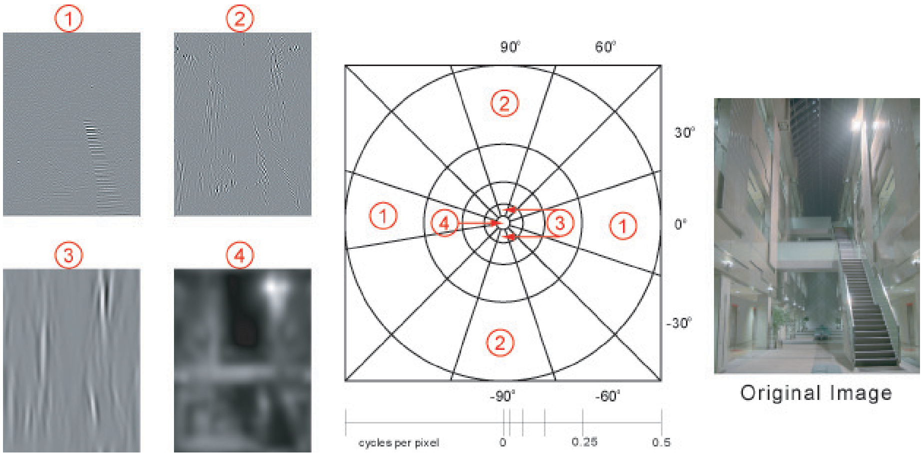


Figure 2.10: *Cortex Transform decomposes an image (right) into multiple frequency and orientation channels with boundaries shown in the frequency domain diagram (center). The four images on the left show the inverse Fourier Transform of representative channels.*

that these decompositions will produce a zero crossing at step edges. This is exactly the opposite of our visual experience, where we tend to be sensitive to edges, perhaps also due to some higher level visual mechanisms. Therefore, instead of only an even response, using a quadrature pair of filters for modelling the receptive fields of the neurons produces results that correlate better with the actual perception (Figure 2.11). The Steerable Pyramid [Freeman and Adelson, 1991] framework uses the Hilbert Transform of the second derivative Gaussian filters in addition to the second derivative Gaussian Filters. Similarly, the Cortex Transform can be extended by combining it with the corresponding quadrature filters [Lukin, 2009]. This effectively removes the phase dependency of the signal, which correlates with the insensitivity of the visual system to phase. In Chapter 7 we propose an extension of the spatial Cortex Transform to the temporal domain, as well a method to remove the phase dependency in the temporal domain.

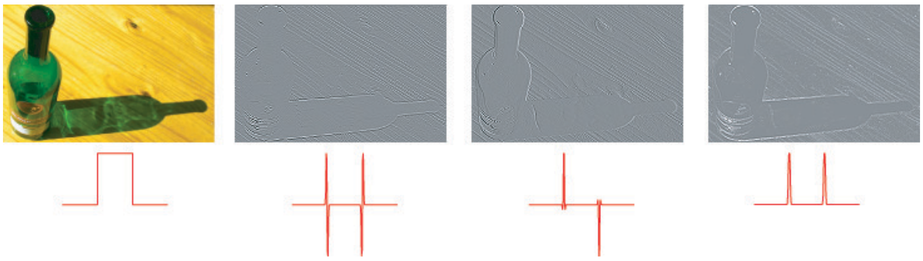


Figure 2.11: *The illustration of phase uncertainty on a complex image (top row) and a simple stimulus (bottom row). The even responses to both stimuli create zero-crossings near step edges, whereas the odd responses are centered at edge locations. Often the combination of both type of responses (quadrature pair) gives a plausible result.*

2.5 Visual Masking

The loss of sensitivity to a contrast patch due to the presence of other "similar" patches nearby is referred to as *visual masking*, as demonstrated in Figure 2.12. However this definition is simplistic, it is well known that if the masking signal's contrast is low, than it facilitates the target rather than masking it. The physiological foundations of this complex mechanism of the human visual system is not well understood, and there are a multitude of models in the literature that differ in their definition of "similarity" of the masker and target signal. Some models only take into account masking from a masker at the same spatial location and spatial frequency and orientation (self masking), while other consider also masking from neighboring spatial locations frequencies, orientations (neighborhood masking). There is however no consensus on the extent of the considered neighborhood for the latter approach. On the other hand a common point of most models is the omission of facilitation for simplicity on the basis that it is not as significant as masking especially in complex images.

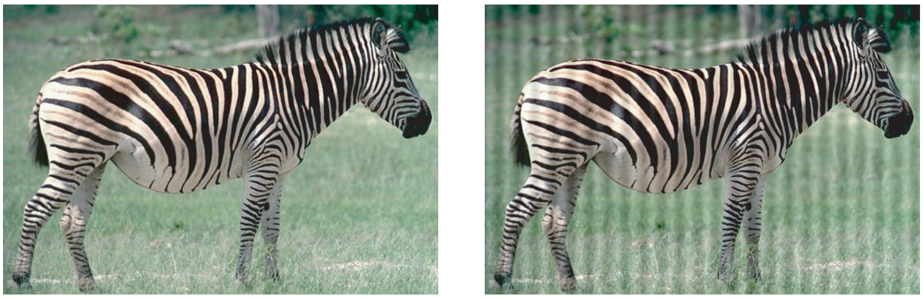


Figure 2.12: *Illustration of visual masking. Even though the distortions imposed on the reference image (left) are uniform in magnitude, they are hardly visible near the zebra's vertical stripes, whereas one can clearly see them on the grass background (right). An interesting point is that distortions are also visible near the zebra's diagonal stripes, illustrating the orientation dependency of visual masking.*

There are two main approaches to implementing visual masking. The first approach involves the use of a *threshold elevation* function, that is a nonlinearity depending on the masker signal's contrast, spatial frequency and orientation. If the frequency and orientation of signal and the masker signal are similar, the original signal is suppressed by applying the appropriate compressive nonlinearity, in effect elevating the visibility threshold of the signal. An advantage of this approach is that if the contrast is already normalized by a human visual system model, a single nonlinearity can be used for all spatial frequency and orientation bands [Daly, 1993]. The second approach involves the use of a *transducer function* that maps physical contrast to a hypothetical perceptual response that also accounts for visual masking [Legge and Foley, 1980; Wilson, 1980; Mantiuk *et al.*, 2006b; Watson and Solomon, 1997]. The transducer function is often preferred in supra-threshold models, whereas the near-threshold models make use of the threshold elevation function. In computer graphics, visual masking has been used in textured mesh simplification [Ferwerda *et al.*, 1997], rendering [Ramasubramanian *et al.*, 1999], tone mapping and color appearance [Pattanaik *et al.*, 1998], among others.

Psychophysical data on temporal stimuli [Boynton G M, 1999] reveals that visual masking also depends on the similarity of the temporal frequency of the masker and target signal. Several models that fit these measurements have been proposed. While models with many narrow band mechanisms, as well as three channels have been proposed in the past, it is now believed that there is just one low-pass, and one band-pass mechanism [Winkler, 2005]. This theory is consistent with the biological structure of the LGN where one can identify *parvocellular* and *magnocellular* pathways encoding low and high temporal frequencies, respectively. Moreover Friedericksen and Hess [1998] obtained a very good fit to large psychophysical data using only a *transient* and a *sustatined* mechanism. We investigate the temporal aspect of visual masking in more detail in Chapter 7.

2.6 Contrast

The human visual system does not have a mechanism dedicated to contrast computation in the sense of the mathematical formulations often used in the literature. The close relation between perception and contrast is due to the center-surround structures that in effect compute luminance differences at multiple frequencies. Computing physical contrast from the scene luminance is a common first step in especially supra-threshold human visual system models. These models then predict the perceived contrast from the physical contrast using a transducer function.

Contrast is the change in the image intensity relative to the local average. It can be used as a degree of distinguishability from the background. Perhaps as a consequence of these vague description there are multiple mathematical definitions of contrast that can be confusing at times. Considering a simple stimulus of a box-like luminance profile, Weber contrast is defined as:

$$W = \frac{L - L_{bg}}{L_{bg}}. \tag{2.3}$$

However, if the stimulus has a sinusoidal luminance profile, and thus is spatially variant, than the selection of L is ambiguous. Using the luminance separately at each location could be misleading, because the resulting contrast shape would be sinusoidal as well. However, we tend to perceive the grating as a whole. Michelson's contrast is a better measure for sinusoidal gratings, as it represents the contrast of the entire grating as a unit:

$$M = \frac{L_{max} - L_{min}}{L_{max} + L_{min}}. \tag{2.4}$$

The definition of contrast becomes more complicated once we consider complex images instead of simle stimuli. In this case, Michelson contrast is obviously not usable, and as for Weber the background luminance is not well defined.

To remedy this, a possible simplification is to ignore the spatial distribution of contrast alltogether and produce a single contrast number from the image, such as the root mean square (RMS) contrast:

$$RMS = \sqrt{\frac{1}{MN}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}(L_{ij} - L_{mean})^2}. \tag{2.5}$$

This measure however is often too simplistic and often a single number does not provide enough information about the image.

Local band limited contrast proposed by Peli [1990] overcomes this problem by computing the local contrast at multiple scales. This contrast measure is formulated as follows:

$$P = \frac{L - L_{lp}}{L_{lp}}, \tag{2.6}$$

where L_{lp} is the low pass filtered version of the original image. Often the contrast is computed as multiple scales, where L and L_{lp} are the neighboring scales of a lowpass pyramid. This measure can be seen as a generalization of the Weber contrast to multiple scales.

Mantiuk [2006b] proposed a lowpass contrast measure defined as

$$G = \left(\frac{L}{L_{mean}}\right), \tag{2.7}$$

which avoids the halo artifacts that appear with Peli's definition.

Among the aforementioned contrast measures, there is no obvious choice that would suppress the others in all cases. In practice the choice of the contrast measure is dictated by application needs and the design choices made in the rest of the human visual system model.

2.7 Visual Significance and Detection

The aforementioned models of the human visual system mechanisms offer merely an interpretation of the retinal image, not a description [Wandell, 1995]. Not much is known about how the human brain uses the HVS-processed visual data to perform the wide range of complex tasks such as face recognition and object tracking. For the purposes of this work, we define two simple, but fundamental tasks that we can perform using the outcome of the human visual system, namely *visual significance* prediction and *detection*. We will show that a multitude of useful methods can be built using these tasks as a basis. In fact, the applications throughout this dissertation make use of human visual systems designed to predict visual significance (Part I) and detection (Parts II and III).

The bare outcome of a human visual system model, that is a prediction of perceived contrast, is a guideline for *visual significance*. The details of the scene, that produce a stronger perceptual response are likely to be more "significant". Compared to commonly used importance measures such as the gradient magnitude, visual significance has the advantage of being scaled perceptually by taking into account mechanisms of the human visual system. In Chapter 3 we show that one can achieve qualitative improvements over non-perceptual importance measures by integrating visual significance into various computer graphics methods.

An important task for human vision is the *detection* of a barely visible signal with a certain degree of reliability. Whether a certain pattern is detectable can be determined experimentally, often using a two alternative forced choice (2AFC) experimental procedure. The magnitude of the experimental stimuli can be modulated according to the PEST procedure for more efficient use of time.

The outcome of the 2AFC experiment can be computationally predicted by using a *psychometic function* that maps the perceived contrast C' computed by a human visual system model to detection probability:

$$P(C') = 1 - \exp(-|C'|^3). \tag{2.8}$$

Often the contrast is computed at multiple scales k, and the psychometric function is applied to each of those scales separately. Finally, the detection probabilities P from multiple bands are combined using a probability summation formula:

$$\hat{P} = 1 - \prod_{k=1}^{K} \left(1 - P^k\right). \tag{2.9}$$

In the first part of the reminder of this thesis, we investigate two visual significance tasks: predicting the visual significance of image edges, and predicting the visibility of images shown on a display under dynamically varying lighting conditions. In the second and third parts we discuss image and video quality assessment methods that are based on the detection task. From this point on, we will assume that the reader is familiar with the aforementioned basics of human visual system modeling. Also, for brevity the term *human visual system* will be abbreviated as *HVS* in the rest of the thesis.

Part I

Visual Significance

Chapter 3

Visual Significance of Image Edges

In the first part of this thesis we investigate two visual significance problems. In this chapter we present an edge aware image decomposition framework based on second generation wavelets [Fattal, 2009] that uses visual significance as its edge strength metric. *The contribution of this work is the use of an HVS model to estimate visual significance as a measure of edge strength*, instead of gradient magnitude that is commonly used in computer graphics applications. The HVS model computes physical contrast at edge locations, and scales it through a cascade of simple and well known models of luminance adaptation, spatial frequency perception and visual masking. The computed visual significance is approximately scaled in perceptually linear units, which implies that similar edge strength values across multiple images correspond to similar perceived strengths.

Localizing significant variations in image luminance and chrominance, i.e. edge detection, has been a classical problem in image processing. Similarly, edge aware image decompositions have been used in numerous computer graphics applications such as image abstraction, detail enhancement and HDR tone mapping. In both contexts, the essential component is an edge model, which in the former case is used to produce a map of image edges, and in the latter case is integrated into the image decomposition algorithm that purposely avoids smoothing near strong edges.

The edge model serves two purposes: determining the location and strength of edges. The majority of the methods proposed for edge detection involve smoothing and differentiation to locate edges. A measure of edge strength is essential, since typically the result of these methods is “too many” edges, and the output is only comprehensible after the removal “less important” edges thorough thresholding. Incidentally, gradient magnitude based edge models are conveniently used in all but the most specialized edge detectors, because one can locate edges by computing local maxima of the gradient magnitude, as well as simply use the magnitude value at the edge location as a rough estimate of edge strength.

While existing methods are capable of localizing edges in a semantically meaningful way, their performance is directly influenced by the edge strength model they employ. The focus of this work is the computation of edge strength rather than edge localization and semantics. Our central idea is that the magnitude of image edges as perceived by the human eye, or the "visual significance" of an edge, should be the guideline for edge strength computation. In that respect, gradient magnitude as an edge strength measure encapsulates the well known property of the human visual system being sensitive to luminance differences, but ignores other aspects such as visual masking and luminance adaptation. Earlier research [Ferwerda *et al.*, 1997] has demonstrated how image contrast is masked by other contrast patches that are of similar spatial frequencies (refer to Chapter 2.5). Except perhaps simple stimuli designed for experimental purposes, visual masking is expected to occur in virtually any complex image and often to has a strong influence on perception. Disregarding the nonlinear perception of luminance, especially in HDR images, often leads to overestimating bright image regions. As a simple counter-measure, one can operate in log-luminance space [Fattal *et al.*, 2002] that better approximates perceived intensity in bright image regions, but fails to model the perception of lower luminance values that is not linear in log-space (see Chapter 2.2 for a discussion on luminance adaptation).

In the rest of this chapter, we first summarize related work (Section 3.1), then discuss the edge avoiding decomposition framework (Section 3.2) and the HVS model (Section 3.3), than we validate the model (Section 3.4) and show that the use of visually significant edges results in qualitatively better outcomes in image retargeting, panorama stitching and HDR tone mapping over gradient magnitude based approaches (Section 6.6).

3.1 Background

Edge Detection has been one of the fundamental problems in computer vision. In an early approach, Marr and Hildreth used the zero crossings of the Laplacian operator motivated by its rotational symmetry [Marr and Hildreth, 1980]. Later Canny focused on finding an optimal differential operator that localizes sharp intensity edges (which he approximated with the first derivative of a Gaussian), and introduced the use of non-maxima suppression and hysteresis thresholding [Canny, 1986]. Canny's method proved to be very reliable over the years and is still widely used. A notable improvement over earlier edge detectors is the use of multi-scale analysis to detect smooth edges as well as sharper edges (see [Pellegrino *et al.*, 2004] for an overview). The steerable pyramid decomposition, while designed for general purpose feature detection, is shown to perform better at small peaks of intensity by combining even and odd filter responses [Freeman and Adelson, 1991]. Lindeberg proposed an automatic scale selection method where the scale of edges is determined by finding the maximum of a strength measure over scales [Lindeberg, 1996]. This method is later employed in Georgeson's third derivative operator [Georgeson *et al.*, 2007], which provides a more compact response than the first derivative. Some effort has also been made to detect color edges [Ruzon and Tomasi, 1999]. For a detailed summary of edge detection techniques we refer the reader to [Ziou and Tabbone, 1997].

Edge detection has found various applications in computer graphics such as guidance over image editing operations [Elder and Goldberg, 2001], stylization and abstraction of photographs [DeCarlo and Santella, 2002] and texture flattening [Perez *et al.*, 2003]. The notion of edge importance understood as its "lifetime" (essentially its presence) over increasing scales in the scale-space framework similar to [Lindeberg, 1996] has been used for stylized line drawings and structure-aware image abstraction [Orzan *et al.*, 2007]. Edge-preserving techniques such as the bilateral filter have been used to decompose an image into a base and detail layers and applied to HDR tone mapping [Durand and Dorsey, 2002]. Recently, Farbman et al. [2008] proposed another decomposition with multiple detail layers and presented applications to scale selective feature enhancement and image abstraction. Fattal [2009] later showed that comparable results can be achieved much faster using a second generation wavelet decomposition with a specialized weighting function that avoids edges. Another approach to edge preserving filtering is detecting the edge strength by computing the gradient of the input image, and reconstructing the image through anisotropic diffusion [Perona and Malik, 1990]. This method decouples edge detection and smoothing, but it is inefficient due to the iterative processing. This method has later been modified by an edge strength measure based on curvature change [Tumblin *et al.*, 1999]. Gradient domain operators such as [Fattal *et al.*, 2002; Mantiuk *et al.*, 2006b], while not explicitly stated, also utilize edges since gradient magnitude operator is essentially an edge detector. Mantiuk et al.'s [2006b] method has additionally a perceptual component in the form of a simple contrast transducer.

3.2 Edge Avoiding Framework

Objects appear differently depending on the scale of observation, and thus visual significance of image features depends on the image scale. Consequently, many image processing tools including edge detection algorithms adopted multi-scale approaches. This has been physiologically justified by the finding that each simple retinal cell responds to a certain bandwidth of spatial frequencies [Wandell, 1995, Chapter 6].

Recent work [Fattal, 2009] demonstrates use of second generation wavelets computed through the lifting scheme [Sweldens, 1997] in the context of edge avoiding multi-scale image decomposition. In this section we give an overview of these concepts, for a detailed discussion refer to [Jansen and Oonincx, 2005]. Contrary to regular wavelets, second generation wavelet bases do not have to be merely translates and dilates of a single pair of scaling and wavelet functions. This generalization enables data dependent filtering through the use of a weighting function that utilizes the information obtained from the local neighborhood changes the shape of wavelet bases accordingly. In the context of edge avoiding wavelets (EAW) the weighting function assigns lower weights to locations containing strong edges, thus the wavelet bases effectively "avoid" those locations.

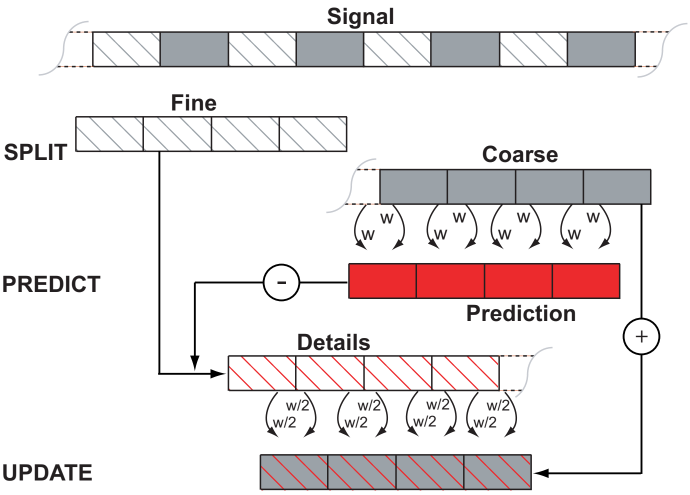


Figure 3.1: *An illustration of the lifting scheme on a 1D signal. The signal is decomposed into fine and coarse parts by arbitrarily designating odd pixels as fine, and even pixels as coarse components. The fine component is predicted from the coarse component using weights computed by the edge aware function ω, or simply by linear interpolation. The difference between the original fine component and the predicted fine component gives the details. The details are then used to update the coarse component. The same process is then iterated on the updated coarse signal.*

The data dependent filtering achieved by wavelet bases not relying on translation and dilation comes at the cost of prohibiting the use of Fourier analysis for wavelet calculation. This issue has been addressed by a discrete wavelet transform named the lifting scheme [Sweldens, 1997]. The basic idea behind the lifting scheme is to *split* a signal arbitrarily into fine and coarse samples, *predict* fine samples from coarse samples and compute the details by subtracting fine samples from their prediction, and *update* coarse samples using the details. Figure 3.1 illustrates the computation in 1D (using Uytterhoeven's coloring scheme [Uytterhoeven *et al.*, 1997]). Advantages of the lifting scheme are fast, in place computation and easily invertible decomposition.

One can achieve edge aware behavior by simply executing a weighting function at each location that assigns weights according to the edge strength at the local neighborhood. If the goal is to avoid edges, i.e. obtaining detail components free of strong edges, this can be achieved by the function ω in Equation 3.1, where m and n are intensities at the current location and some neighboring pixel, respectively:

$$\omega(m,n) = \frac{1}{(|\nu(m,n)|^{\alpha} + \epsilon)}. \tag{3.1}$$

The control parameter α is set to 0.8 as suggested in [Fattal, 2009]. Divisions by zero are prevented by setting ϵ to 10^{-5}. We will use the function ν later for the estimation of visual significance; in the original implementation it simply returns the difference of n and m. Such a decomposition is useful in contrast editing applications such as detail enhancement and image abstraction, since halo artifacts are prevented due to the absence of strong edges in detail components. The opposite goal of extracting solely strong edges can be achieved by simply using the *inverse* of ω. The detail components of the resulting decomposition closely resemble the outcome of multi-scale edge detectors, which we utilize in context aware image retargeting and panorama stitching applications (Section 6.6).

The straightforward extension to the second dimension is to repeat the 1D computation at both dimensions (Figure 3.2a). If an edge preserving weighting function is used, the results of this 2D decomposition are analogous to X and Y gradients, and thus fit naturally into the edge detection pipeline. Another splitting method by [Uytterhoeven *et al.*, 1997] with lower anisotropy produces better results coupled with an edge avoiding weighting function (Figure 3.2b).

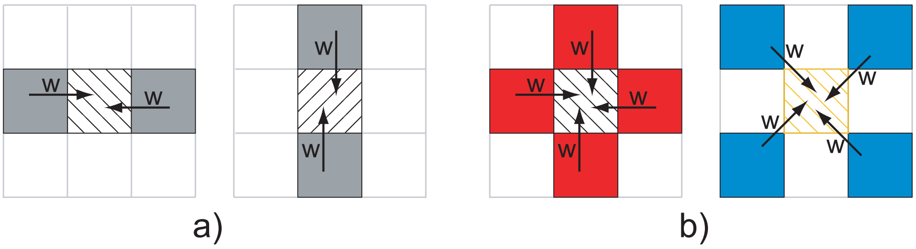


Figure 3.2: *The lifting scheme can be extended by repeating the 1D computation in X and Y directions (a), or using a lower anisotropy red-black quincunx lattice (b). Only the prediction step is illustrated for brevity.*

3.3 Human Visual System Model

We extend the EAW framework (Section 3.2) with an HVS model, where we modify the weighting function (Equation 3.1) that penalizes strong differences of image pixel values by computing visual significance of the luminance differences. The HVS model takes physical image luminance as input, therefore 8-bit images should be mapped to display luminance and HDR images should be calibrated to scene luminance before processing. The luminance contrast C (Chapter 2.6) is approximated in the EAW framework by dividing the fine samples by the local mean of the K immediate *coarse* neighbors (K equals 2 and 4 for X-Y splitting and red-black splitting, respectively):

$$C = \frac{Fine}{(\frac{1}{K})\sum_K Coarse_k} - 1. \tag{3.2}$$

Repeated at each scale, this formulation is similar to the low-pass contrast in [Mantiuk *et al.*, 2006b]. The advantage of a contrast based edge strength measure over a gradient based measure is illustrated in Figure 3.3

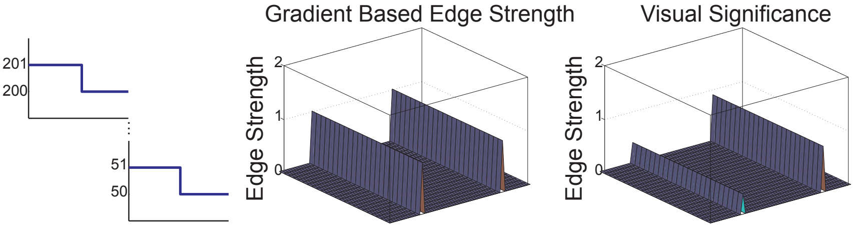


Figure 3.3: *Edge strength predictions utilizing physical contrast account for the effect of background luminance level. The perceived strength of step edges 200-201 cd/m^2 and 50-51 cd/m^2 (left) are predicted to be the same by the gradient based method, whereas a contrast based method correctly predicts the weaker perceived strength of the first profile.*

Note that the contrast C is computed solely using physical luminance. As the next step we scale C by computing the sensitivity of the visual system to obtain contrast in perceptually linear units. Two prominent factors that affect **contrast sensitivity** (Chapter 2.3) are its spatial frequency (ρ), and the adaptation luminance (L_a). These effects can easily be observed in the Campbell-Robson chart.

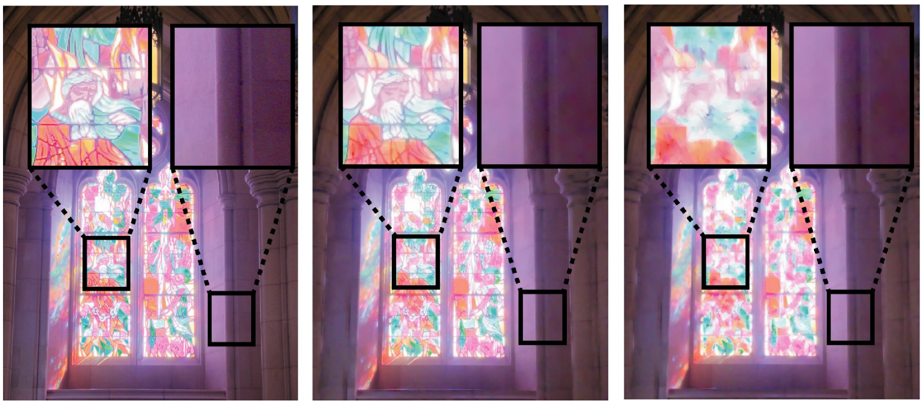


Figure 3.4: *An illustration of the effect of luminance adaptation (the practical utility of our model is shown in Section 6.6). The original HDR image (left), smoothing with EAW method (center), and smoothing with EAW method using visually significant edges (right). The strength of edges of the bright window are overestimated by EAW method in the absence of a model of luminance adaptation. All images are tone mapped [Reinhard* et al.*, 2002] for display purposes.*

We use the CSF from the Visible Differences Predictor [Daly, 1993] with corrections as indicated in [Aydın *et al.*, 2008a, Equations (10, 11)] (also in Appendix Section 10.2) to obtain the perceptually linearized contrast $C' = C \cdot CSF(\rho, L_a)$. Figure 3.4 shows an example where the difference in edge preserving smoothing is mainly due to the scaling of contrast by the CSF. This behavior is typical in HDR images, where the contrast magnitudes at very bright and very dark image regions are overestimated by the frameworks without perceptual components. As a result, the edges of the bright window are avoided unlike the edges at the window's frame (Figure 3.4 center). The CSF's scaling results in a more uniform smoothing over edges with similar magnitude of visibility (Figure 3.4 right).

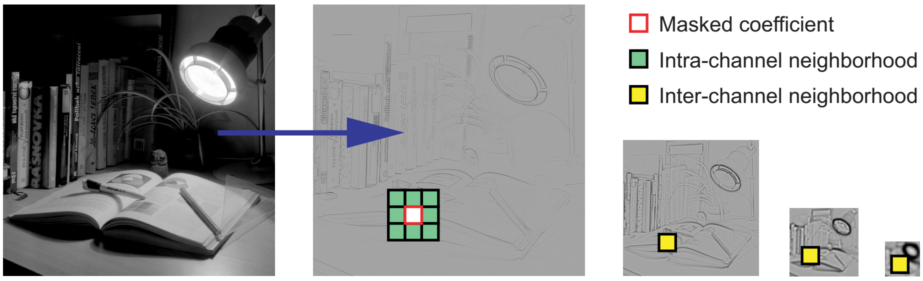


Figure 3.5: *An illustration of neighborhood masking on detail layers of a multi-scale decomposed image. At each image location, visual masking is computed as a function of the immediate 8 neighboring pixels. The same neighborhood spans a larger area in coarser scales (visualized by yellow boxes).*

Visual masking (Chapter 2.5) is the decrease in visibility of a contrast patch in the presence of other contrast patches of similar spatial frequencies. One way of modeling this effect is by computing a *threshold elevation* map for each visual channel, which when divided by the contrast at that channel accounts for the increase in detection thresholds (thus, decrease in sensitivity). This method trades off accuracy at supra-threshold contrast levels for better prediction near the threshold, and has been used in image quality assessment metrics for distortion detection. On the other hand, the *transducer* model is focused on perception of supra-threshold contrasts and thus preferred in discrimination tasks. The model relies on a transducer function that is constructed by iteratively summing up contrast detection thresholds. The use of a transducer function in computer graphics context is demonstrated in [Ferwerda *et al.*, 1997]. A more comprehensive transducer model [Watson and Solomon, 1997] also comprises masking from adjacent frequency channels (inter-channel masking). In this model, since the lower frequency channels contain information from the spatial neighborhood, a contrast patch at a certain location is effectively masked by neighboring contrast patches (See Figure 3.5 for an illustration of neighborhood masking.)

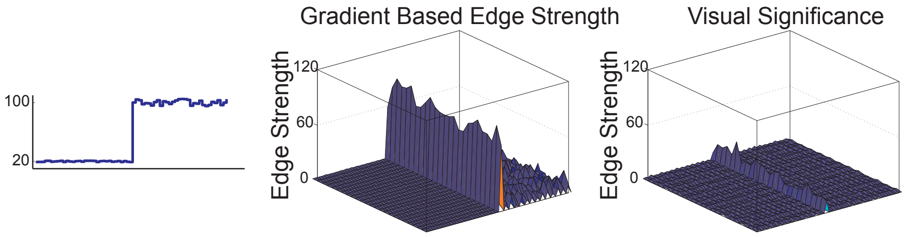


Figure 3.6: *The visual masking due to the random noise modulated by image luminance in the test stimulus (left), results in lower perceived edge strength then the gradient magnitude (center), as predicted by our method (right).*

While the visual masking due to the local neighborhood is often not significant for isolated test stimuli, natural images tend to have "busy", textured regions

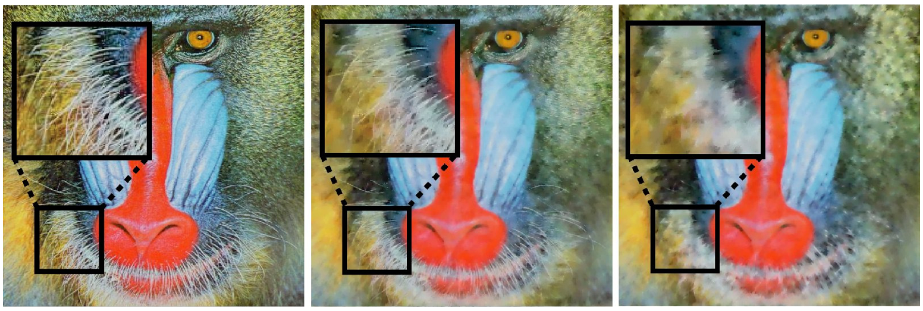


Figure 3.7: *The effect of contrast masking in a complex image. The original image (left), smoothing with EAW method (center), and smoothing with EAW method using visually significant edges (right). The masking model reduces the strength of the facial hair edges due to the presence of hair in the local neighborhood.*

where the visibility of edges are notably lesser than non-textured regions. To account for that, our ν function (Equation 3.1) comprises the point-wise extended masking model [Zeng *et al.*, 2000] which, in addition to a compressive nonlinearity, also accounts for visual masking from the local neighborhood K:

$$R = \frac{sign(C')|C'|^{0.5}}{(1 + \sum_K |C'_k|^{0.2})} \; . \tag{3.3}$$

The effect of visual masking on a simple stimulus is illustrated in Figure 3.6. Figure 3.7 shows that the involvement of the point-wise extended masking model results in a perceptually uniform smoothing near high-masking regions. Computation of the hypothetical HVS response R is the final step in function ν in EAW the framework.

3.4 Model Calibration – Perceptual Experiment

To validate and calibrate the proposed edge perception model, we conducted a simple threshold-level perceptual experiment. The motivation for this is twofold: first, we aim to calibrate the implemented supra-threshold transducer model described above (Equation 3.3) for threshold stimuli; second, as noted by [Whittle, 1986], discrimination thresholds for spatially separated patches should not be generalized for perceiving edges, thus there is a lack of usable experimental data. Furthermore, the CSF curves [Daly, 1993] reflect measurements using the Michelson's definition of contrast, which is slightly different from the implemented definition contrast (Equation 3.2).

In our experiment, two adjacent grayscale patches were presented on a calibrated display device. The luminance of the left patch is kept constant during each trial, whereas the luminance of the right patch was modulated according to the responses of the subject. Each subject was asked whether there is a visible edge between the two patches or not. The luminance of the right patch was decreased if the response was positive, and increased if the response was negative. The step sizes were determined by following the PEST procedure [Taylor and Creelman, 1967]. A random noise pattern was presented for 1s between stimuli to avoid afterimages, memory effects, etc. Each trial ended once the standard deviation of the subject's last 6 responses were below the minimum step size ($0.01cd/m^2$) or if there were more than 30 responses collected. The experiment comprised 10 trials for each subject, where the initial luminance of the left patch at each trial is selected by randomized sampling from the luminance range $1.5 - 400cd/m^2$.

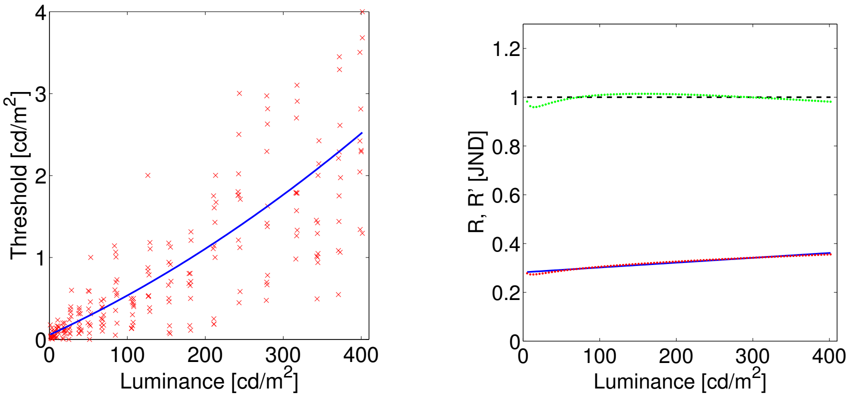


Figure 3.8: *Experimental results. Left: measured edge detection luminance thresholds as a function of adaptation luminance L_a, right: model predictions before (red dots) and after the calibration (green dots). An ideal model response is constantly 1 JND for the threshold data (dashed line).*

The stimuli were displayed on a calibrated Barco Coronis MDCC 3120 DL, a 10-bit 21-inch hi-precision LCD display, in its native resolution 2048×1536 pixels, the maximal display luminance was 440cd/m^2. The display response was measured by the Minolta LS-100 luminance meter. The experimentation room was darkened (measured light level: 1 lux), and observers sat approximately 70 cm from the display. The total of 22 observers took part in our experiment. There were both male and female observers, and all of them reported to have normal, or corrected-to-normal vision. Each subject was verbally introduced to the problem before the experiment.

The measured edge perception thresholds, see Figure 3.8 (left), were approximated by the second order polynomial function (blue curve). Using the polynomial function, we generated 100 threshold stimuli as the inputs for model calibration procedure. We assume that the model output for each stimulus at the threshold level should be R=1 JND. Therefore, we run the model for each of 100 input stimuli to obtain the error function, see Figure 3.8 (right). The threshold prediction of the uncalibrated model (red dots) was quite solid, so that we decided to perform the calibration by means of a simple linear function which should not affect the performance of the model for supra-threshold stimuli. The calibration was achieved by dividing the masking model by the calibration function (blue curve in Figure 3.8 (right)):

$$R' = \frac{R}{0.0002\ L_a + 0.2822}, \tag{3.4}$$

where L_a is the adaptation luminance in cd/m^2.

As the masking model (Equation 3.3) was verified in JPEG 2000 applications, we did not calibrate it for supra-threshold data. However, we believe that the supra-threshold performance is also improved as a consequence of the threshold calibration, and the precision of the model is more than sufficient for various applications as illustrated in the next section.

3.5 Applications

In the previous sections we showed that the use of visual significance results in smoothing that better correlates perceived strength of edges. However, applications like image abstraction through edge preserving smoothing or detail enhancement produce images whose quality is judged aesthetically. Thus, despite the obvious differences between the perceptual and non-perceptual methods, one can not objectively prove that a visually significant edge model produces better results. In this section we present three applications that rely on importance of image features, and thus the improvement through a perceptual model can be demonstrated through examples. All results are generated using the extended EAW framework. The edge maps used in image retargeting and panorama stitching are generated by using the inverse of Equation 3.1 as discussed in Section 3.2.

3.5.1 Image Retargeting

Several techniques were recently proposed to allow content-aware image and video retargeting [Avidan and Shamir, 2007; Wang *et al.*, 2008; Rubinstein *et al.*, 2009]. The central part of those approaches is usually an *importance map* (energy function) that describes the importance of areas in the image. Using the map, the retargeting operator then preserves the important areas at the expense of less-important ones. Several possibilities of the importance map construction were proposed [Avidan and Shamir, 2007], however a simple Sobel operator was utilized in many cases.

The visually significant edges are a natural candidate to construct such importance map in a perceptually more convincing way. We show the results of seam carving image resizing operator [Avidan and Shamir, 2007] using traditional importance map and the new map calculated by our technique in Figures 3.9 and 3.10. The traditional technique removes more visually significant areas than when we build importance map using our method. Our results indicate that the difference between both methods is especially significant if the visually significant details are located in dark image regions. While the perception of brighter details ($> 100\ cd/m^2$) can be approximated by a simple compressive logarithmic function, our method has the advantage of faithfully modeling perception in all luminance levels and taking masking into account, and thus overall produces more reliable results (Figure 3.10 (c) and (d)). In fact, the success of particular importance map construction varies with the input images and the absence of a universal retargeting operator led to the proposal of a hybrid approach combining several techniques [Rubinstein *et al.*, 2009]. Our results suggest that visual significance can be guideline in importance map computation and can provide a basis for more sophisticated retargeting operators. An advantage of our approach is that it allows perceptually based retargeting on not just ordinary, but also high dynamic range images. That said, we found that first producing a tone mapped "dual" image, and then performing the retargeting on the original HDR image using the edge strengths computed on the dual image to work well in some cases. However, the type of tone mapping operator and suitable parameter setting is an open question, and requires manual interaction in comparison to our fully automated method.

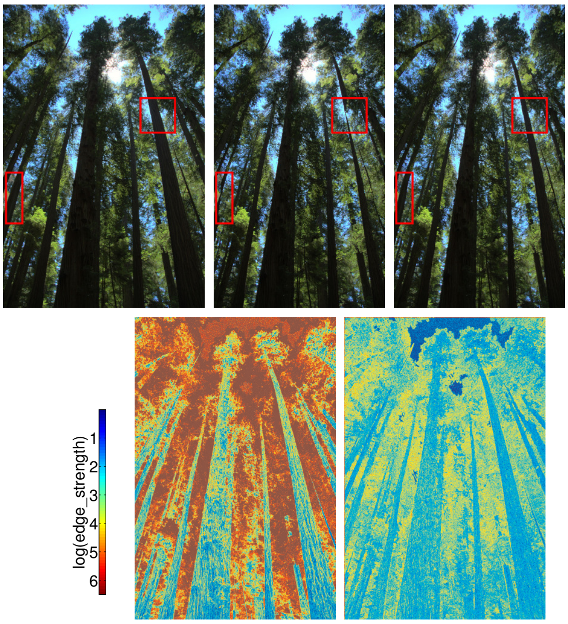


Figure 3.9: *HDR image shrinking by seam carving (150 pixels horizontally). First column left: original HDR image. Middle: result when the Sobel operator is used for importance map construction. Right: result using the proposed visually significant edges. Images are tone mapped [Drago* et al.*, 2003b] for the display purposes. Second column: edge strength maps. Left: edges detected by Sobel operator in the input HDR image. Right: visually significant edges – note the differences in absolute values and in the ratios of edge strengths (due to the JND scaling), and the structural differences in the edge map (due to the masking).*

3.5.2 HDR Tone Mapping

As mentioned in experimental evaluations [Kuang *et al.*, 2007b; Čadík *et al.*, 2008], the goal of tone mapping is manifold: some tone mapping operators are focused on compressing the image luminance while preserving the overall scene appearance. For example, the outcome of such an operator applied to a dark scene would not reproduce the details that are not visible by the human eye

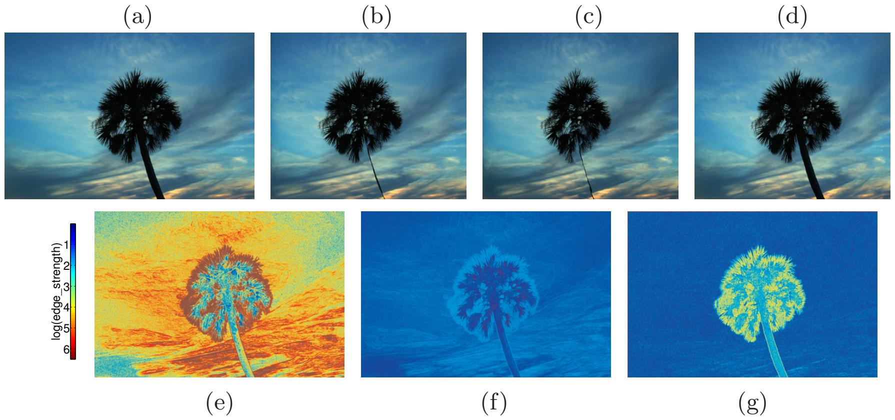
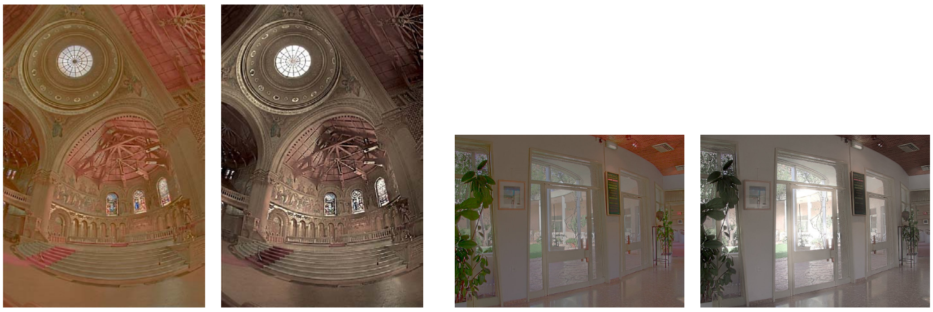


Figure 3.10: *HDR image shrinking (400 pixels horizontally) by seam carving. First row: (a) original HDR image, (b) Sobel operator overestimates the strength of edges in the sky, which results in carving of the visually important palm tree, (c) results are similar if the Sobel operator results are compressed by the logarithm function, (d) the proposed method results in less distorted image appearance, especially evident at the tree's body. Images are tone mapped [Drago* et al.*, 2003b] for the display purposes. Second row: (e,f,g) Edge strength maps for (b),(c),(d).*

due to insufficient lighting. The other group of tone mapping operators on the other hand focuses on preserving as many scene details as possible irrespective of their visibility magnitude.

The tone mapping from the original edge avoiding framework [Fattal, 2009] can be classified as strictly detail preserving. In the spirit of previous decomposition-based approaches [Tumblin *et al.*, 1999; Fattal *et al.*, 2002; Durand and Dorsey, 2002; Farbman *et al.*, 2008], the technique flattens the coarsest scale of the EAW image decomposition by factor β and the other scales are progressively compressed so that the wavelet coefficients in a coarser scale are decreased more than in a finer scale (by factor γ^k, where k is the scale). This corresponds to an observation that the coarser scales often contain very high magnitude differences and should be therefore compressed much more than the finer scales (details) that we usually aim to preserve. The technique operates on *logarithm* of the input luminance that can be thought of as a simple approximation of human luminance perception, but having not accounted for other prominent perceptual phenomena (e.g. the perception of contrast), the results look unnatural, see Figure 3.11 (left).

The results produced by the technique mentioned above may be suitable for certain scenarios (e.g. the best reproduction of details), but not for reproducing the appearance of a scene. However, we can achieve much better results (in this sense) by replacing the logarithm function with the perceptual framework proposed in this chapter. We thus obtain image decomposition coefficients that are closer to the human visual system response (accounting for phenomena described in Section 3.3) and those are then compressed in a same way as above for the display purpose. As expected, the results are then more natural renditions of the original HDR images and preserve the scene appearance, see Figure 3.11 (right).



Original with HVS Model Original with HVS Model

Figure 3.11: *HDR image tone mapping without (left columns) and with our HVS model (right columns). The original method [Fattal, 2009] preserves as many image details as possible at the cost of overall scene appearance. Our method is more balanced in terms of reproduction of scene appearance and detail preservation.*

3.5.3 Panorama Stitching

An HDR panorama generation approach proposed by Ward [2006] makes use of edge maps to stitch adjacent images of a scene. In this method images are decomposed into two layers: a low pass layer that corresponds to $1/16^{th}$ of the image's original resolution and a high frequency layer. The low frequency layers of adjacent images are blended together using a sinusoidal weighting function, whereas the high frequencies are spliced at locations containing strong edges. The method is guided by a compound edge map E obtained as a combination of edge maps of pairs of overlapping images (E_{left}, E_{right}). We adopted the following technique to construct the compound edge map:

$$E = \max(E_{left} \cdot E_{right}, 0). \tag{3.5}$$

In other words: if there is a strong edge in the left image, but not in the right image, then this is possibly due to a misalignment and should not be preferred for splicing. On the other hand, locations containing strong edges with the same sign in both images are strong candidates for splicing.

For panorama stitching application, we inverted the neighborhood masking in our model, so that it amplified the masked edges. This is motivated by observation that the masked edges also mask the seams so that they are less disturbing in the final panorama. We empirically found that multiplying R with $(2 \cdot Neighborhood\_masking)^2$ to work well in practice. We compare the results obtained using our technique and the traditional Sobel operator in Figure 3.12. The source images were inverse tone mapped prior to processing by simple contrast stretching.

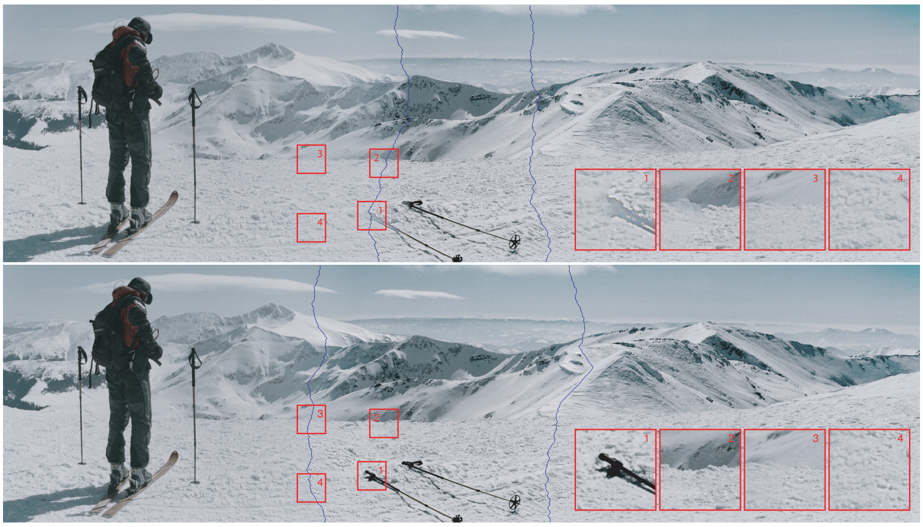


Figure 3.12: *An HDR panorama stitched from three different, not precisely aligned pictures using Ward's technique [Ward, 2006]. Top: the result obtained using Sobel operator, Bottom: the result using the proposed visually significant edges. The images are tone mapped [Reinhard* et al.*, 2002] for display purposes.*

3.6 Conclusion

We presented a method that localizes image edges and scales their strength proportionally to their visual significance. We discussed a simple and efficient HVS model that accounts for prominent features of the visual system such as luminance adaptation, spatial frequency sensitivity and visual masking. In our experience the visual significance computation in EAW framework increases the edge-map computation time by $30 - 50\%$.

The HVS model is integrated into the edge avoiding wavelet framework which provides a convenient basis for edge preserving image decomposition, and also extraction of edges by inverting the edge-stopping criterion. The choice of the framework is not crucial for specialized applications that rely either solely on image decomposition or edge extraction. For example, the HVS model can be applied to multi-scale image gradients for the former type of applications, or to an image pyramid obtained through bilateral filtering for the latter type of applications. The wavelet framework is convenient in the sense that it can serve both purposes in one framework, and is faster than others in decomposition.

The main limitation of this chapter is the absence of models for higher level mechanisms of the visual system such as gestalt properties and prior knowledge. Unfortunately modeling those mechanisms is not trivial because of their complexity and consequently the hardness of designing reproducible experimental setups to determine their effects. Moreover, it has been shown that the shape of the CSF becomes flatter at supra-threshold contrast levels [Georgeson and Sullivan, 1975]. A more precise treatment of supra-threshold contrast sensitivity could involve implementing the transducer given in Watson and Solomon [1997], but it is not clear how to perform the inhibitory pooling involved in this model within the second generation wavelet framework without notably increasing the computation time. The presented model, along with similar supra-threshold models used in computer graphics context, does not account for this behaviour for efficiency reasons.

In the light of recent work [Cole *et al.*, 2008] that shows luminance edges are in fact prominent image features, we believe that the visually significant edges are good candidates for determining the richness of detail in images. Such a measure, combined with others such as image brightness, overall contrast and colorfulness can provide a good estimate of image quality in the absence of a reference image (no-reference image quality assessment). As a future direction we would like to investigate the possibility of designing such a metric that utilizes visually significant edges.

Chapter 4

Display Visibility under Dynamically Changing Illumination

In this chapter we investigate another visual significance problem, namely the visibility of image features when viewed on a display under dynamically changing lighting conditions. The simplistic HVS we employed in the previous chapter does not take into account external factors such as the ambient illumination and reflections on the display surface, which are crucial for this application. Thus, in effect the visual significance computation from the previous chapter is adjusted to the sensitivity of an observer sitting in front of a monitor, in a room with controlled illumination that does not interfere with the observer's adaptation state. On the other hand, the display technology progresses not only towards increasing brightness, contrast and color reproduction quality, but also making display devices lighter and thinner. As a result, the use of display devices is no longer limited to indoors where the illumination shows lesser variation. How is the visibility of details on, for example, a cell phone display affected when exposed to direct sunlight? In this chapter, we investigate how dynamic changes in illumination affects the visual significance of the displayed content.

The method presented in this chapter accounts for the decrease in sensitivity due to *maladaptation* (adaptation to a different luminance level than the background luminance) that may be caused by abrupt changes in lighting as well as the observer directing her gaze to a brighter or darker object (see Chapter 2.2 for the basics of luminance adaptation). Since we do not assume static illumination, the sensitivity of the HVS changes over time due to the changing adaptation state (Section 4.2.1). Our metric, on the one hand predicts the spatially varying magnitude of visibility of the reference and associates pixels with easily interpretable visibility classes like *informative*, *warning&caution*, etc. (Section 4.2.2), and on the other hand detects the loss of details due to the reflections with respect to the reference (Section 4.2.3). The final visual significance of the displayed content is the combination of the visibility classes with the effect of reflections. We present results for various lighting conditions and visual system states in Section 8.2, and apply our method to a car interior display, where the lighting of both the car interior and the display is computed by a global illumination simulator (Section 4.4).

4.1 Background

The visibility of the displayed content is expecially crucial in automotive and aerospace applications, where the pilot or the driver often needs to react quickly to changing conditions. Dreyer [2007] proposes determining the visibility level of a displayed contrast patch as the ratio between the luminance difference between a symbol and its background, and Adrian's threshold luminance [Adrian, 1989] which is a function of the symbol size and exposure time. The temporal aspect of adaptation is modeled in the time-to-visibility metric [Krantz and Silverstein, 1992], which takes into account display contrast, ambient illumination and the adaptation luminance to determine the time when a given spatial frequency pattern becomes visible. Mantiuk et al. [Mantiuk *et al.*, 2008] proposed a quality metric which takes into account ambient illumination conditions to optimize perceived detail reproduction in a tone mapping algorithm. Other work focused on the discriminability of symbols shown on the displays in airplane cockpits [Ahumada *et al.*, 2006].

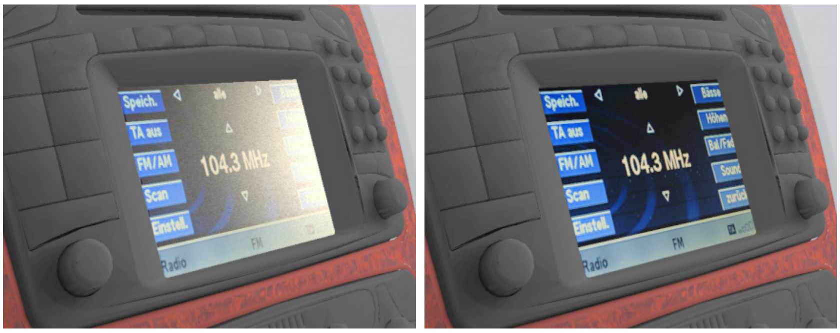


Figure 4.1: *Images of a car display with and without reflections. Note that gamma corrected images should be converted to physical luminance before being processed by our method.*

4.2 Visibility Analysis

Our method requires a reference display emission image along with the image of the display subjected to reflections, both scaled in cd/m^2 units. An example input image pair is shown in Figure 4.1. We use a combination of two measures in our display analysis: for each pixel, first we determine the visibility classes of the display emission image as a function of the perceived contrast magnitude, and second, we detect the loss of details due to the reflections. The former measure is computed at contrast levels well above the threshold (supra-threshold), while the latter happens at the vicinity of the visibility threshold (near-threshold). We employ separate methods to model both tasks, each specialized in modeling the corresponding contrast range. The predicted visual significance of the displayed content is a combination of the predictions of the outcome of both models. In the rest of this section we elaborate on modelling temporal adaptation, and discuss how we incorporate temporal adaptation to the near– and supra-threshold components of our metric.

4.2.1 Temporal Adaptation

The sensitivity variation at a certain adaptation state is commonly modeled as a sigmoid response profile centered at the corresponding luminance level [Naka and Rushton, 1966]. To cope with temporally and spatially changing real world luminance, the adaptation state is continuously readjusted. In scenarios with dynamically changing lighting conditions, the relatively slow pace of temporal adaptation plays a significant role in visual perception. The threshold luminance when the HVS is maladapted to the adaptation luminance L_a, while the actual (background) luminance is L, is typically given as a *threshold versus intensity and adaptation* function ($\Delta L = tvia(L, L_a)$).

Almost all known adaptation mechanisms operate within retina, each of which having their own time course suggesting that they should be tracked separately [Pattanaik *et al.*, 2000]. The fast but less effective neural mechanisms and slower but more effective photochemical mechanisms are responsible for shifting the response profile across the visible luminance range for both cone and rod systems. We adopt Irawan's [2005] approach, where adaptation due to pigment bleaching (σ_b), slow neural adaptation (σ_c) and fast neural adaptation (σ_n) are modeled separately, and Equation 4.1 gives the adaptation state as a function of adaptation luminance L_a:

$$\sigma(L_a) = \sigma_b(L_a)\ \sigma_c(L_a)\ \sigma_n(L_a). \tag{4.1}$$

The sigmoid shaped retinal response function R for this adaptation state as a function of background luminance L and a sensitivity control parameter n is given in Equation 4.2:

$$R(L, \sigma(L_a)) = \frac{L^n}{L^n + \sigma(L_a)^n}. \tag{4.2}$$

To obtain the threshold luminance ΔL at an adaptation level given by the tvia function, first the differential retinal response $\Delta R = R(L + tvi(L), L) - R(L, L)$, that produces a unit JND, is computed assuming perfect adaptation ($L_a = L$). The tvi (threshold versus intensity) function returns the visibility threshold of the fully adapted visual system given the background luminance. We derive the tvi function from VDP's contrast sensitivity function by iteratively computing the maximum sensitivity for each adaptation luminance along all spatial frequencies. Finally, the difference between the luminance value that generates the response $R + \Delta R$ and L gives the threshold luminance of the maladapted visual system.

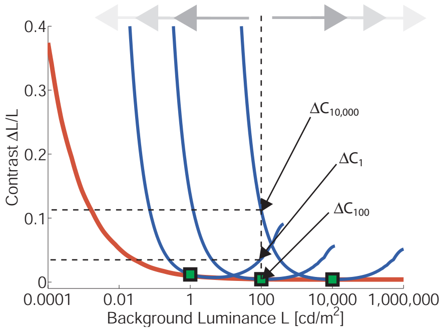


Figure 4.2: *Threshold contrasts (calculated by normalizing tvia by background luminance L) for perfect adaptation (red curve) and for adaptation luminances (L_a) of 1, 100 and 10,000 cd/m^2 (blue curves). At $L = 100$ cd/m^2, threshold contrast (ΔC) is lowest for adaptation luminance 100 cd/m^2, whereas for mal-adapted states with L_a equals 1 cd/m^2 and 10,000 cd/m^2 the threshold increases notably.*

In Figure 4.2, we plot the threshold contrasts ($1/sensitivity$) for three adaptation states at L_a equals 1, 100 and 10,000 cd/m^2 (blue curves) along with the threshold contrasts for perfect adaptation (red curve). The perfect adaptation curve is approximately the envelope of all adaptation states. We can think of the blue curves shifting horizontally as the visual system adjust to a new adaptation state. The time course of neural adaptation in the case of an abrupt change in lighting from luminance L_0 at time $t = 0$, to L_a is modeled as the exponential decay function (Equation 4.3) for neural adaptation of both rods and cones:

$$L_{a-current} = L_a + (L_0 - L_a)\, e^{\frac{-t}{t_0}}. \tag{4.3}$$

Temporal change in adaptation is modeled by updating the tvia at each time step with the current adaptation level $L_{a-current}$. We set t_0 to 0.08 seconds for cones, and 0.15 seconds for the rods as given in [Irawan *et al.*, 2005]. We consider only the steady-state behavior of relatively slow pigment bleaching, since we observed that detail visibility is almost entirely recovered within the first few seconds.

Next, we discuss the supra– and near threshold measures we employ in our analysis and introduce new building blocks based on the tvia function, that extend those measures by modeling adaptation over a time course.

4.2.2 Visibility Classes

Visibility classes relate the contrast of the reference display emission to magnitude of contrast visibility scaled in JND units. Contrast of typical displayed content is well over the visibility threshold. Thus we use a transducer based supra-threshold HVS model, that accurately predicts the magnitude of HVS response by taking into account visual masking. A numerical response value computed by the model alone is not descriptive (e.g. how much visible are 50

	Visibility Class	PJND	Description
6	Attention Getter	120	Attention getting quality must be maintained beyond the para-foveal limit and into the peripheral vision areas
5	Warning & Caution	90	Warning or cautionary information requiring predominate attention
4	Dynamic Complex	70	Complex formats with small alphanumeric characters and/or fine line analogue or graphic presentations. This data is **not** fixed in location
3	Static Complex	60	Same as Dynamic complex, but data is fixed in location
2	Status	50	Dual State (On-Off) information. Location is fixed
1	Informative	40	Fixed format single state information. Provides background information supporting controls or more complex presentations

Table 4.1: *The visibility classes and associated PJND values.*

JNDs?). Instead, a classification of HVS response intervals into *visibility classes* (VC) is easier to interpret by humans. The *perceptible just noticeable difference (PJND) model* introduces 6 classes of visibility (Table 4.1) and has been applied to airplane cockpits [Sharpe *et al.*, 2003]. The method has been calibrated by subjective experiments on airplane-pilots and civilians in separate studies, and similar values are obtained for both. The PJND value is defined as the geometric mean of luminance and chrominance JNDs. According to [Dreyer, 2007], the effect of chrominance is relatively small, therefore we consider only luminance contrast.

In the original PJND method, luminance to JND conversion is done be normalizing the logarithmic contrast by an experimentally found constant assuming that the observer is adapted to $10,000\ cd/m^2$. In environments subject to strong sunlight (such as airplane cockpits), it is reasonable to assume logarithmic HVS response and high adaptation luminance. But under dimmer lighting this model will severely underestimate observer sensitivity. Additionally, the significant effect of visual masking on supra-threshold contrast perception is neglected. In our work, we employ a multiscale luminance contrast perception model [Mantiuk *et al.*, 2008] to compute the hypothetical supra-threshold HVS response (Figure 4.3). First, we calculate the logarithmic contrast G across

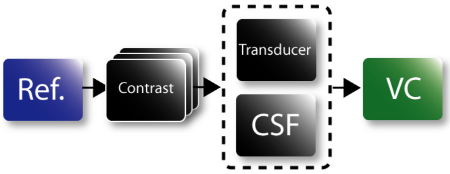


Figure 4.3: *Building blocks of the visibility class (VC) analysis. See text for details.*

scales given the image luminance, by computing the logarithm of image luminance and building a Gaussian pyramid. The logarithmic contrast at level l is then given by the difference between levels l and $l+1$ of the pyramid, where larger numbers indicate coarser scale. Considering the high frequency nature of the information conveyed through display devices (text, symbols, etc.), we focus on the loss of local details rather than distortions in the global contrast. Thus, consistent with the original method, we only consider frequencies higher than $3cy/deg$. Next, Wilson's transducer [Wilson, 1980] is used to compute the HVS response given contrast $W = \Delta L/L$ and sensitivity S as input (Equation 4.4). Note that logarithmic contrast can then easily be converted to Weber contrast ($W = 10^{|G|} - 1$) (see Chapter 2.6 for a discussion on various contrast definitions).

$$T(W,S) = \frac{3.291\;[(1+(SW)^3)^{1/3} - 1]}{0.2599\;(3.433 + SW)^{0.8}}. \tag{4.4}$$

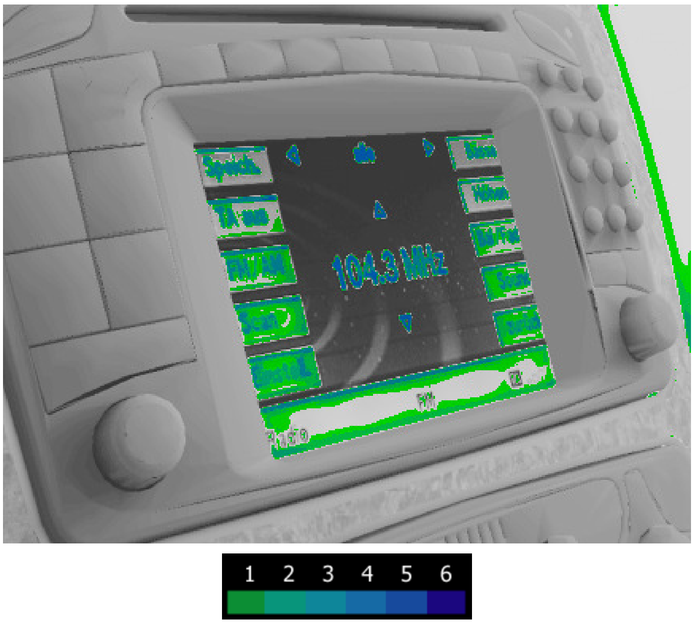


Figure 4.4: *Visibility classes of the emission of a car display. Refer to Figure 4.1-right to see the original image.*

The contrast sensitivity function [Daly, 1993] used in the original method to compute S in Equation 4.4 is designed for steady-state adaptation. Using the tvia function from Section 4.2.1, we derive Equation 4.5 that also accounts for temporal adaptation:

$$S = OTF(\rho, p)\,\frac{nCSF(\rho, L_a, d)}{tvia(L, L_a)/L}. \tag{4.5}$$

The normalized contrast sensitivity $nCSF$ (Chapter 10.2) is modeled as a function of spatial frequency ρ, adaptation level L_a and viewing distance d. The (OTF) models the disability glare due to the reflections in optics of the human eye, where p denotes observer's pupil diameter. Although not commonly observed in low dynamic range imaging, disability glare has a significant effect on our perception of HDR images. The effect of changing adaptation conditions to the HVS response is shown in Figure 4.5.

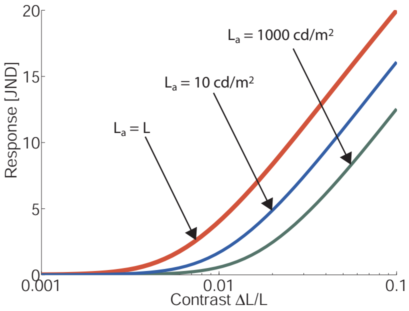


Figure 4.5: *Supra-threshold HVS response for background luminance $L = 100$ cd/m^2, at adaptation levels L_a: 100 (red), 10 (blue) and 1000 (green) cd/m^2 at 4 cy/deg.*

The HVS responses to luminance contrast across all scales are summed up using a Minkowski summation with exponent 2. In our visualization, the test image is shown in gray-scale while corresponding visibility classes are color-coded according to the scale at the bottom (Figure 4.4).

4.2.3 Loss of Details

Note that the visibility classess computed entirely from the reference display emission, and do not account for the loss of details due to reflections. Spatially varying illumination may produce specular reflections that locally reduces the display contrast. The consequent decrease in the visibility of the displayed content is modeled in a separate near-threshold method [Aydın *et al.*, 2008a], that will be discussed in detail in Chapter 6.

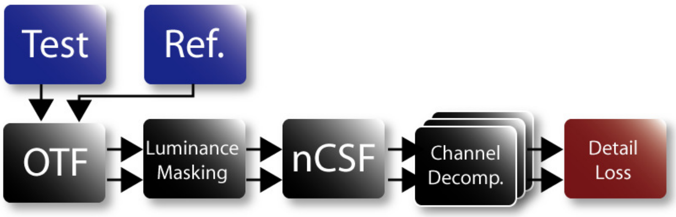


Figure 4.6: *Main processing steps of the detail loss analysis. The luminance masking step is modified as in Equation 4.6. See the Chapter 6 for further discussion of the pipeline.*

The input to the metric are the luminance values of the test image (display emission and reflections) and the reference image (display emission only), both given in cd/m^2 units (Figure 4.1). The main processing steps of the method are depicted in Figure 4.6, where both input images undergo the same processing separately, until the final distortion detection step.

To model temporal adaptation we introduce a mapping from luminance to a perceptually uniform space scaled in JND units of a maladapted visual system. Unlike the original method that assumes perfect adaptation, we derive the mapping for a given adaptation luminance L_a by iteratively adding threshold values at the maladapted state, starting from the minimum luminance L_1 (10^{-3} cd/m^2) until the maximum luminance L_N (10^{10} cd/m^2):

$$L_i = L_{i-1} + tvia(L_{i-1}, L_a) \quad i \in \{2, 3, \ldots, N\}. \tag{4.6}$$

The index i of luminance L_i gives the corresponding JND value for the maladapted visual system. The JND values at arbitrary luminance levels are interpolated from the two closest neighbors. The resulting mapping from luminance to JNDs is shown for perfect adaptation and three adaptation levels at 1, 100 and $10,000$ cd/m^2 in Figure 4.7. We calibrate both components using the calibration values from [Aydın *et al.*, 2008a] for the case when $L_a = L$, that are obtained through psychophysical experiments on the modelfest dataset [Watson, 2000].

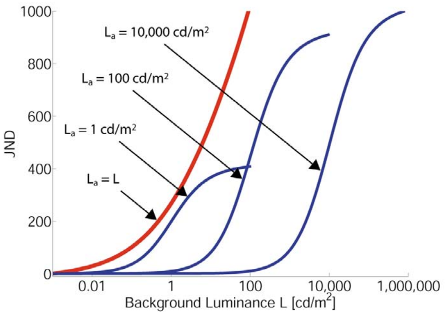


Figure 4.7: *JND values for perfect adaptation (red curve), and L_a equals 1, 100 and $10,000$ cd/m^2 (blue curves from left to right). Note the large differences between JND values for the same background luminance in different adaptation states.*

We use the same optical transfer function (OTF) and normalized contrast sensitivity function ($nCSF$) as discussed in Section 4.2.2. The orientation and spatial frequency selectivity of the neurons in the visual cortex are modeled at the *channel decomposition* step through the cortex transform [Watson, 1987] with modifications as in [Daly, 1993] (Chapter 10.4). Consistent with Section 4.2.2, we use the cortex bands down to mean frequency 3 cy/deg, while additionally performing processing for 6 orientations.

In order to predict only the detail loss due to the reflections, we calculate the detection probability of the case where visible contrast in the reference becomes invisible in the test image, separately at each frequency and orientation. The detail loss map is generated by combining distortions across frequencies and orientations (through regular probability summation [Aydın *et al.*, 2008a]). We

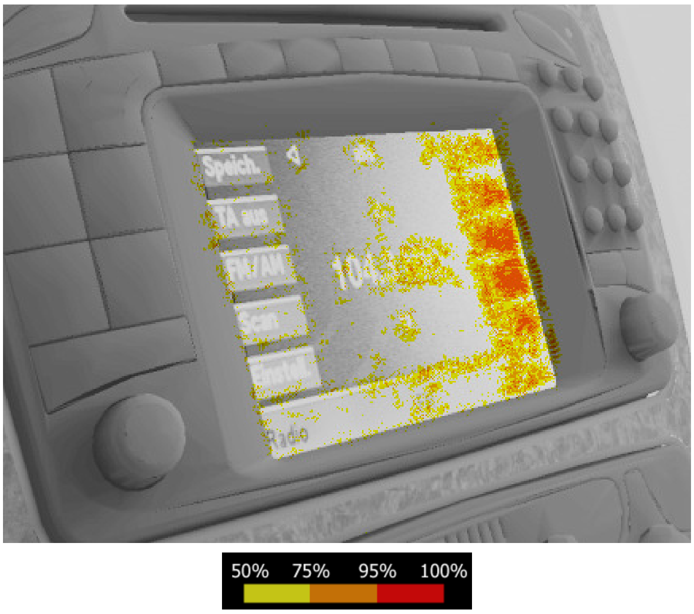


Figure 4.8: *Detail loss due to the reflections on the car display. Refer to Figure 4.1-left to see the original image. The loss of visibility of the symbols and characters on the right display side are detected by the metric. On the other hand the increase of luminance on the display background in this region due to the reflections is ignored by our metric because it does not lead to any structural changes in the image.*

take a similar in-context map approach to visualization of detail loss as in visibility classes (refer to Figure 4.8). The visual significance of the displayed content is computed by assigning the locations with $> 50\%$ detail loss probility to 0 in the visibility class map.

4.3 Results

In this section we test our method on a Barco Coronis 3MP LCD display (max. luminance 400 cd/m^2) under multiple levels of reflections. Firstly, for each image in our test set we generate a scene referred HDR image of the corresponding display emission. HDR images are generated by combining multiple shots from a Canon 5D camera with different exposures using the open source pfsCalibration package. Next, in the same way we capture a reflection component generated by the camera flash. A test set is created by amplifying the reflection components to three separate levels and combining them with captured emissions for three test images. Resulting visibility class and detail loss maps, and the combined visual significance maps for an adapted observer are shown in Figure 4.9.

Our method associates high contrast regions such as icons and text to the highest visibility classes, whereas background regions are predicted to have lower importance. The detail loss analysis detects more structural distortions with

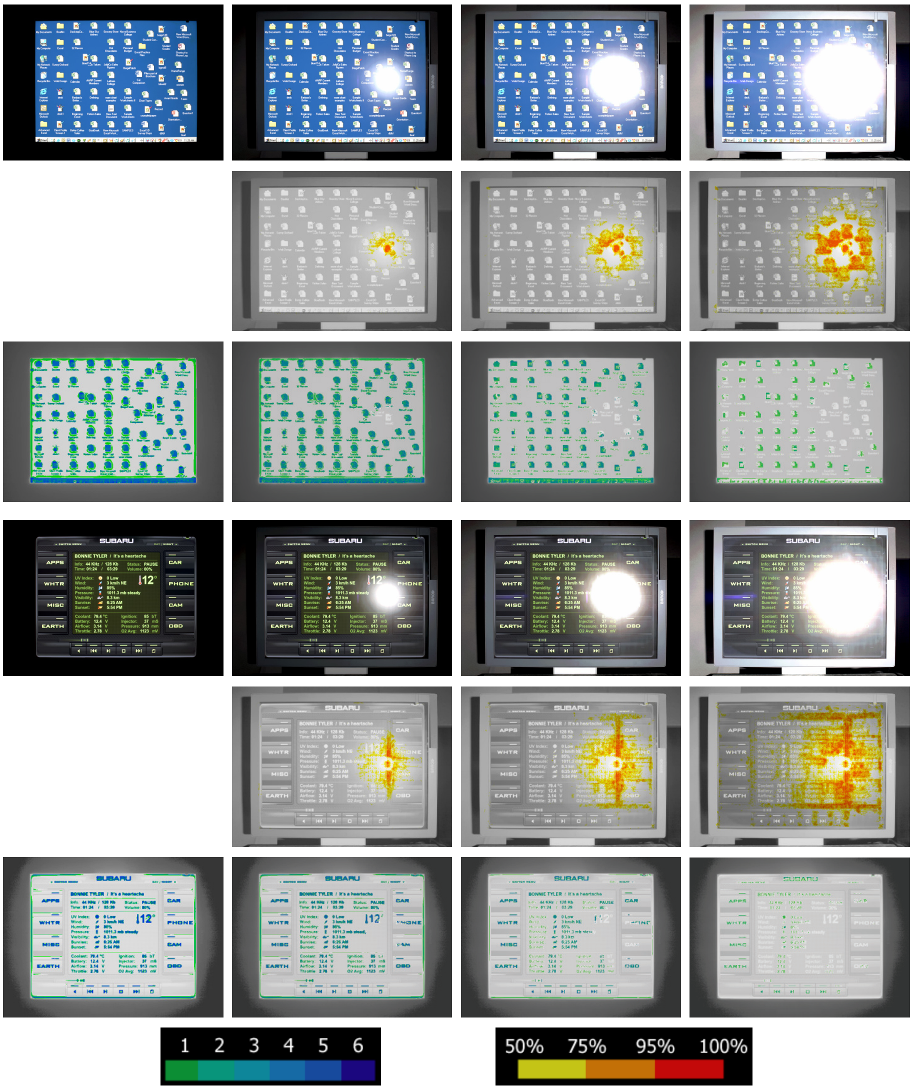


Figure 4.9: 1^{st} *and* 4^{th} *rows, from left to right: the reference display emission, and three levels of reflections in increasing order* $(100, 500, 2500$ $cd/m^2)$*. The* 2^{nd} *and* 5^{th} *rows: detail loss maps due for three levels of reflections. The* 3^{rd} *and* 6^{th} *rows: resulting visual significance maps. The visibility maps are computed for a perfectly adapted observer.*

each increase in reflections. Note that we correctly differentiate between the contrast introduced by the reflection component and the image contrast occluded by the reflection component, and detect only the latter. In Figure 4.10, we show how our measures respond to temporal recovery of sensitivity. In this scenario, the observer first adapts to an image with reflections (Figure 4.9: 4^{rd} row, 3^{rd} column). The adaptation luminance at time $t = 0$ is calculated as 5664 cd/m^2 by averaging over a 1 visual degree area near the brightest center part of the reflection. Next, the reflections are removed, leaving the display emission

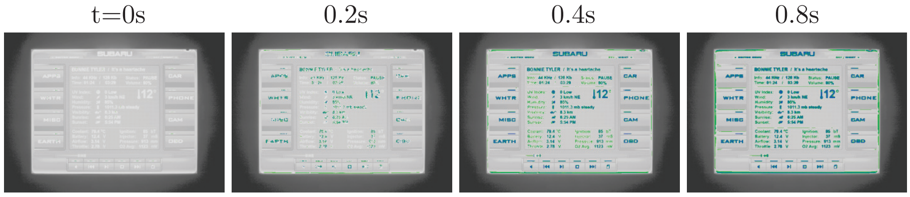

t=0s 0.2s 0.4s 0.8s

Figure 4.10: *Visibility class (first row) and detail loss (second row) maps for* $L_0 = 5664\ cd/m^2$*, after time steps indicated at the top.*

fully visible (Figure 4.9: 4^{rd} row, 1^{st} column) while the observer is still adapted to the luminance of the highlight. Our analysis shows that after 0.8 seconds nearly all details become visible, and visibility classes are improved.

4.4 Automotive Application

To demonstrate a possible application, we integrate our method to a global illumination simulator that models a car interior containing a navigation panel display [Dmitriev *et al.*, 2004]. HDR environment maps captured using an HDR camera with a fisheye lens mounted on the roof of driving car have been used to illuminate the virtual car model. Since the car geometry is static in our application, the precomputed radiance transfer (PRT) technique has been used to efficiently compute global illumination in the car interior for each environment map frame.

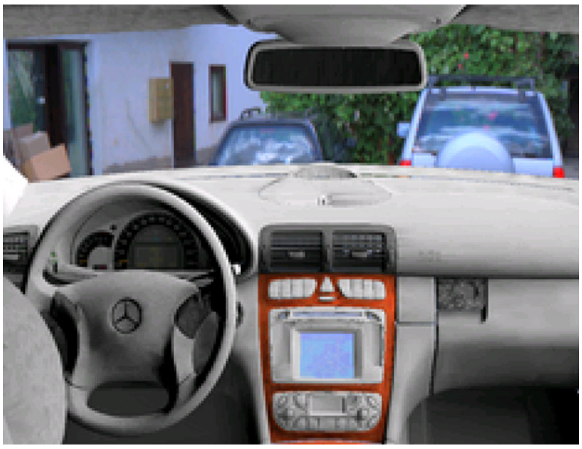


Figure 4.11: *Rendering of the car cockpit. Original HDR image is tone-mapped for displaying purposes.*

The reflectance from the display has been computed off-line using precise final gathering with importance sampling driven by the bi-directional reflectance distribution function (BRDF), which was measured for the actual car display covered with antiglare/antireflection layer. Figure shows an example view of a car interior produced by our renderer. The simulation environment allows us to conveniently test the effect of various levels of reflections reflections using our metric 4.12.

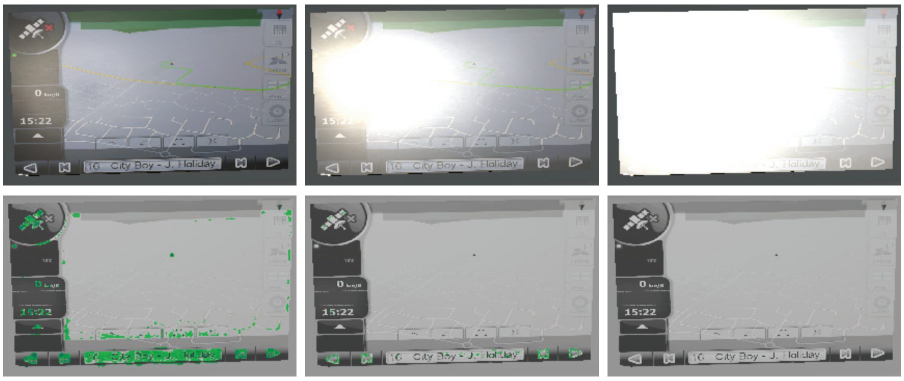


Figure 4.12: *Visibility analysis of the displayed content subject ot increasing level of reflections (*$250, 1250, 6250\ cd/m^2$*).*

4.5 Conclusions

We introduced a method for display visibility analysis that works under spatially and temporally varying illumination conditions and accounts for the temporal adaptation of the observer's visual system. Our method consists of two parts: A supra-threshold metric that associates visible contrast of the display emission with visibility classes, and a near threshold metric that detects the visible detail loss due to reflections. We extend current methods, that assume the eye is perfectly adapted at single pixel resolution, by deriving the components necessary to model the temporal change in sensitivity. The performance of our method is demonstrated on an LCD display illuminated by spatially varying ambient light of different intensity. We also integrated our method to a global illumination simulator and present visibility analysis of a car cockpit display under various lighting and adaptation conditions.

One limitation of our work is that we use a single adaptation luminance for the entire image when modeling maladaptation. A better approximation to real adaptation luminance would be found by averaging over a region at each location. However, the exact support size and type of such an averaging kernel is unknown to us. We also assume that the displayed content to be static. A higher perceived contrast can be achieved by introducing temporal variations to displayed content (e.g blinking lights). Our model can be improved by taking into account the change in contrast sensitivity due to temporal variance. It would also be interesting to compare our method to reaction time based visual performance studies such as [Rea and Ouellette, 1991] [Ueno *et al.*, 1985].

Part II

Image Quality Assessment

Chapter 5

HDR Extension for Simple Image Quality Metrics

In the second part of this dissertation we present two image quality assessment techniques. In this chapter we discuss an extension to a pair of common LDR quality metrics, whereas in Chapter 6 we present a full dynamic range independent quality assessment pipeline.

Most of the commonly used quality metrics do not take into account the brightness of display devices. Such metrics take as input 8-bit code values (luma or gamma corrected pixel values) and assume that they are perceptually uniform, regardless of how bright or dark the display is. However, the visibility of distortion can increase significantly as the display gets brighter. Taking into account the effect of display brightness is especially important for the new LCD TVs, whose peak brightness (over 500 cd/m^2) exceeds five or more times the typical peak brightness of a CRT display.

Accounting for luminance effects is also important for HDR images. They store linear radiance or luminance maps, instead of 8-bit gamma-corrected code values. The difference between luminance or radiance values has little correspondence with the actual visible difference, since the eye is sensitive to luminance ratios rather than absolute luminance values, the property sometimes referred as the luminance adaptation (Chapter 2.2). Therefore, simple measures computed on luminance or radiance maps have little correspondence with the actual image quality. In this chapter we explain how absolute luminance values can be converted to an approximately perceptually uniform encoding, which in turn can give meaningful quality predictions when used with the image quality metrics that operate on pixel values.

In this chapter we discuss how the perceived image quality is affected by the actual luminance levels. We propose an extension to a pair of well-known quality metrics in the form of a transfer function, referred as perceptually uniform (PU) encoding. The PU encoding transforms luminance values in the range from 10^{-5} to 10^{8} cd/m^2 into approximately perceptually uniform code values. The resulting code values are passed to the quality metric instead of gamma corrected RGB or luma values. The proposed PU encoding is derived from the contrast sensitivity function (CSF) that predicts detection thresholds of the HVS for a broad range of luminance adaptation conditions.

The PU encoding is designed so that it is backward-compatible with the sRGB nonlinearity within the dynamic range of a CRT display. Consequently, the quality metrics using PU encoding show similar behaviour as the original metrics for CRT displays. We test the proposed PU encoding with two widely used visual quality measures: the Peak Signal to Noise Ratio (PSNR) [Wang and Bovik, 2006] and the more sophisticated Structured Similarity Index Metric (SSIM) [Wang and Bovik, 2006].

5.1 Background

Objective visual quality metrics either model luminance adaptation (effect of luminance of the detection threshold) explicitly and include it in their processing, or implicitly, assuming that input code-values are "gamma-corrected" and thus perceptually linearized. The former group includes Sarnoff VDM [Lubin, 1995b], PDM [Winkler, 2005], DVQ [Watson *et al.*, 2001], VDP [Daly, 1993], HDR-VDP [Mantiuk *et al.*, 2005] and many other metrics that model the HVS. These metrics, however, due to their complexity, difficult calibration, on-going standardization effort or lack of freely available implementation, are not as popular as the latter group of metrics, which includes arithmetical and structural metrics. Two such popular metrics are peak signal-to-noise ratio (PSNR):

$$PSNR(x,y) = 20\,log_{10}\,\frac{D}{MSE(x,y)} \quad MSE(x,y) = \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(x_i - y_i)^2, \tag{5.1}$$

and structural similarity index metric (SSIM) [Wang and Bovik, 2006]:

$$SSIM(x,y) = l(\mu_x,\mu_y)^{\alpha}\; c(\sigma_x,\sigma_y)^{\beta}\; s(\sigma_x,\sigma_y)^{\gamma}, \tag{5.2}$$

where x and y are pixel values in reference and distorted images, D is the dynamic range, μ and σ are the mean and standard deviations of the corresponding input images. The final quality measure $SSIM$ is a weighted combination of the luminance comparison function l, contrast comparison function c and structure comparison function s. These metrics rely on the perceptual linearity of input pixel values x_i and y_i, which should account for luminance adaptation. In the following sections we show that this is reasonable assumption for CRT displays, but it is less accurate for much brighter LCD displays. This is especially the case when the same "gamma" function is used for both a bright and a regular display. Finally, such metrics cannot be applied directly to HDR images.

The proposed PU encoding in conceptually similar to the DICOM Grayscale Standard Display Function [DICOM, 2001], but is intended to handle a larger dynamic range. The proposed encoding is an adaptation of the color space used for HDR image and video encoding [Mantiuk *et al.*, 2006a] for quality metrics that ensures backward-compatibility with the sRGB color space.

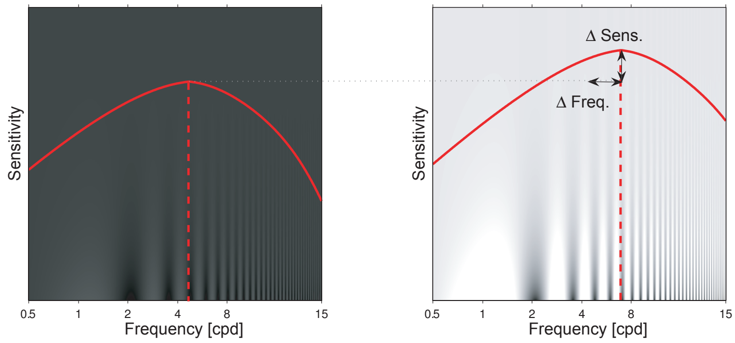


Figure 5.1: *Contrast sensitivity function (CSF) of the human eye in dark (left) and bright (right) viewing conditions. Arrows labelled as $\Delta Sens.$ and $\Delta Freq.$ denote the amount of difference in magnitude and frequency of the peak sensitivity between the dark and bright cases.*

5.2 Distortion Visibility on Regular and Bright Displays

The effect of luminance level on the sensitivity of the human visual system is often referred as luminance adaptation. Figure 5.1 shows the Campbell-Robson contrast sensitivity chart for two different background luminance levels. For the best viewing, the figure should be viewed on an LCD display of about 200 cd/m^2 and the display function close to the sRGB nonlinearity. The solid lines denote the contrast sensitivity of the HVS, which is the contrast level at which the sinusoidal contrast patterns become invisible. Even though the same scales were used for both left and right plots, the CSF is shifted upwards (higher sensitivity) and right (towards higher spatial frequencies) for the brighter pattern. This shows that we are more likely to notice contrast changes, if the stimuli is brighter, as is the case of a brighter display.

But it is not clear if this observation for simple sinusoidal pattern can be assumed valid for complex images. Consequently, we cannot assume that a difference in sensitivity due to image brightness results in a difference in quality assessment. To verify this, we performed a subjective quality evaluation of distorted images shown on the displays of different brightness.

Our 16 test subjects were within the ages 23–48, all with near perfect or corrected vision. Each subject was presented a reference and distorted image side by side for 10 seconds. After that interval, a blank screen was displayed and the subjects were asked to assess the quality of the distorted image with respect to the reference on a 5 point scale, where higher values indicate better quality. Subjects were given the opportunity to view the image pair again for additional 10 second intervals until deciding on the image quality. A set of distorted test images was generated by applying 3 types of distortions (random pixel noise, gaussian blur and JPEG compression) at 2 levels (high and low) to 3 images. Each image pair was shown on a Brightside DR-37P HDR display, which simu-

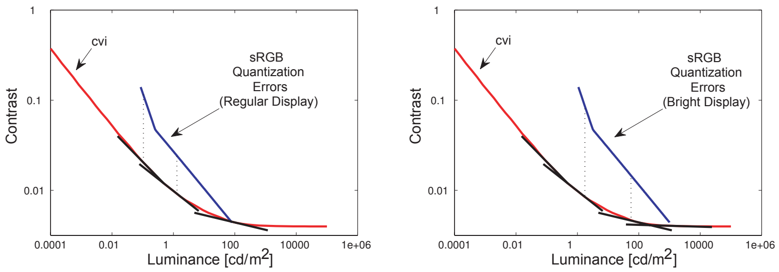


Figure 5.2: *Quantization errors of sRGB encoding for maximum luminance* $80cd/m^2$ *(left) and* $1000cd/m^2$ *(right), in comparison to contrast versus intensity (cvi) function of the HVS. The discrepancy between the slopes of both functions is large, especially for the bright case.*

lated either a regular (1–100 cd/m^2) or a bright display (10–1000 cd/m^2). The simulated displays had the same response as an actual LCD display (measured with a Minolta LS-100 luminance meter), only the absolute luminance levels were shifted for the bright display. The order of trials were entirely randomized and each image was shown 2 times to ensure subject reliability

Our experimental setup and grading scale is adopted from ITU-T Rec. P.910 standard [ITU-T, 1999]. We determined the mean quality value for the regular display as 3.15, and for the bright display as 2.85, indicating that subjects tend to perceive the quality of distorted images to be lower on the bright display. In other words, distortions of the same type and with the same magnitude are more annoying when the overall brightness of the image is higher. An evaluation of the data with the ANalysis Of VAriance (ANOVA) method resulted in an F-value of 20.57 and the corresponding p-value $\ll 0.05$ for the display brightness parameter, showing that the effect of display brightness to perceived quality is statistically significant.

5.3 Weber-Fechner Law and Luminance Adaptation

Figure 5.1 reveals that the threshold contrast $\Delta L/L$ is different for dark and bright stimuli (refer to Chapter 2.3 for a discussion on contrast sensitivity, as well as Equation 10.3 for the formula of the function used to generate the plots). This is contrary to the commonly assumed Weber-Fechner law, which would require that the ratio $\Delta L/L$ stays constant. This observation is better illustrated on the contrast versus intensity (cvi) plot shown in Figure 5.2. The cvi function indicates the threshold contrast (y-axis) at particular luminance adaptation level (x-axis). The region where such contrast is constant, and the Weber-Fechner law holds ($\Delta L/L = const.$), can be found for luminance values greater than approximately 500 cd/m^2. For lower luminance levels the detection threshold rises significantly. This indicates that the Weber-Fechner law is in fact very inaccurate model of luminance adaptation for the range of luminance shown on typical displays (from about 0.1 cd/m^2 to 100-1000 cd/m^2).

5.4 sRGB Nonlinearity and Detection Thresholds

The compressive nonlinearity (transfer function) used in the sRGB color space accounts not only for the response of a typical CRT, but also partly for the drop of the HVS sensitivity for dark luminance levels. The sRGB nonlinearity has the form:

$$l(L') = \begin{cases} \left(\frac{L'+0.055}{1.055}\right)^{2.4} & \text{if } L' > 0.04045 \\ \frac{L'}{12.92} & \text{otherwise,} \end{cases} \tag{5.3}$$

where L is the trichromatic value (for simplicity we assume luminance) normalized by the peak display luminance and L' is the "gamma-corrected" luma value. In Figure 5.2 we plot the peak quantization errors due to 8-bit coding of L', assuming the peak display luminance of 80 cd/m^2 on the left (CRT) and 1000 cd/m^2 on the right (bright LCD or Plasma). We compute the peak quantization errors as

$$e(L') = \frac{1}{2} \frac{\max |l(L'\pm 1) - l(L')|}{l(L')}, \tag{5.4}$$

but plot them in the luminance domain (L), instead of luma domain (L'), to compare different displays. The slopes of the error quantization functions give closer match to the cvi function for the darker display (80 cd/m^2), suggesting that the sRGB has better perceptual uniformity for CRT displays. The slopes start to deviate much stronger for brighter displays, making perceptual uniformity of the sRGB nonlinearity for LCD and Plasma displays questionable.

Another observation that we can make in Figure 5.2 is that the quantization errors of 8-bit code value encoding are actually larger than the detection threshold of the human eye. This means that when we display a smooth gradient on a display driven by 8-bit input, we can see contouring artifacts. This is true even for darker displays, but is more noticeable for bright displays, where the discrepancy between encoding quantization errors and the cvi gets larger. Such contouring artifacts could be easily hidden by adding random noise to the gradient (spatial or temporal dithering). For the same reason, medical displays are usually driven by signals of 10- or more bits to reduce the quantization errors to an undetectable level.

5.5 Detection Thresholds in Complex Images

Before we can derive a perceptually uniform encoding, we need to estimate contrast detection thresholds as a function of pixel luminance. Many aspects of complex images, such as spatial frequency, orientation and masking pattern, can significantly rise the detection threshold. Figure 5.3 illustrates how the sensitivity (inverse of the contrast detection thresholds) changes with spatial frequency and adapting luminance. Since the perceptually uniform encoding is a function of pixel value, we need to reduce all these factors except adapting

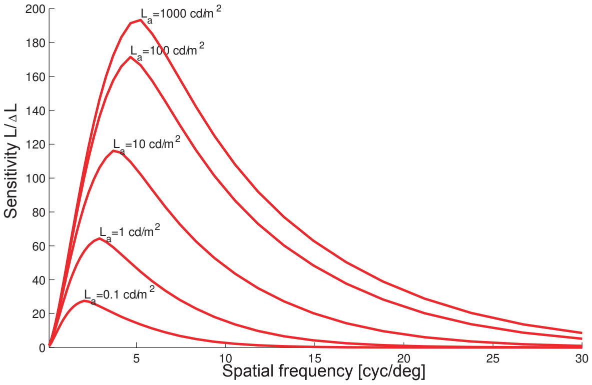


Figure 5.3: *Contrast sensitivity function variation with adaptation luminance.*

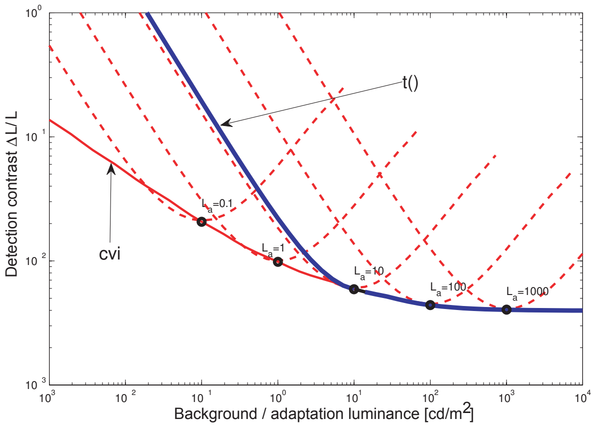


Figure 5.4: *Continous line - cvi function for different adaptation levels; Dashed lines - contrast detection thresholds for fixed adaptation and varying background luminance. Refer to the text for the description of function t.*

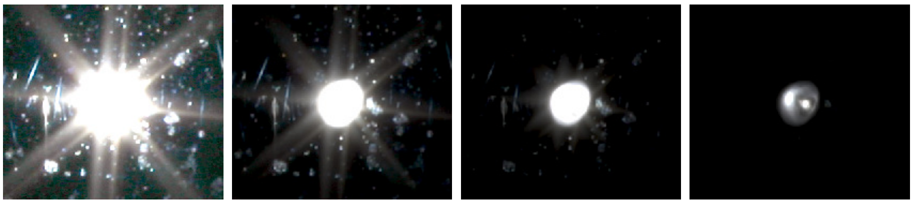


Figure 5.5: *A specular highlight on a piece of metal captured using multi-exposure technique. The exposure time decreases from left to right. The right-most image reveals the reflection of a lamp, which is not visible to the human eye in the actual setup.*

luminance L_a, and assume that they will be taken into account by the actual quality metric. To ensure that the estimated detection threshold is always conservative, we choose the value that corresponds to the maximum sensitivity for each factor we want to reduce. Therefore, we define our cvi function as:

$$cvi(L, L_a) = \left(\max_{\mathbf{x}} \left[CSF(L_a, \mathbf{x})\ MA(|L - L_a|) \right] \right)^{-1}, \tag{5.5}$$

where the CSF is the contrast sensitivity function and $\mathbf{x}$ corresponds to all the parameters (spatial frequency, orientation, stimuli size, etc.) except adapting luminance L_a and the background luminance L. The $MA()$ function estimates the loss of sensitivity due to maladaptation, as explained below. We use the CSF function from Daly's VDP [Daly, 1993], as it is valid for a large range of luminance values (both photopic and scotopic viewing).

To properly utilize the cvi function, it is important to distinguish between the adapting luminance, L_a, and the background luminance, L. When viewing a complex scene the human eye can adapt locally to small regions. For example our eyes are in one state of luminance adaptation when looking outside a window on a sunny day, and in a different state when looking at the interior of a room. However, the eye is hardly ever perfectly adapted for each tiny luminance patch in a scene. For example, when looking at bright specular reflections, we usually cannot see the reflected features of a light bulb or the sun, since we are adapted to the diffuse light reflected from an object, rather than the tiny specular spot. Figure 5.5 shows that such tiny features are in fact reflected, but we usually don't see them. The situation when the eye is maladapted has been studied in so-called *probe-on-flash* experiments [Walraven *et al.*, 1990], in which a threshold stimuli on a background was briefly flashed, thus bypassing the adaptation process. The typical characteristics measured in such experiments are shown in Figure 5.4. The plots were derived by combining the typical cvi function with an S-shaped photoreceptor response curve, as done by Irawan, et al [Irawan *et al.*, 2005].

To make our extension spatially independent and possibly compatible with the sRGB nonlinearity, we make two simplifying assumptions about the luminance adaptation process. Firstly, we assume that there is a minimum luminance level to which the eye can adapt, L_{a-min}. When viewing complex images, the darkest areas are usually affected by the glare (light scattering in the eye's optics), therefore the minimum luminance level that reaches the retina and to which the eye can adapt is elevated. Secondly, we assume that the eye is perfectly adapted for all luminance levels above L_{a-min}, that is the adapting luminance is equal the luminance of the pixel ($L_a = L$). The second assumption results in the most conservative estimates of the contrast detection thresholds (refer to Figure 5.4). Our final estimates of the detection thresholds are:

$$t(L) = cvi(L, max(L, L_{a-min})). \tag{5.6}$$

5.6 Perceptually Uniform Encoding

The goal of perceptually uniform encoding is to ensure that the distortion visibility is approximately uniform along all encoded values. This is achieved when the differentials of such encoding are proportional to the detection thresholds. The easiest way to find such mapping from the detection threshold estimates t (Equation 5.6) is to use the following recursive formula:

$$f_i = f_{i-1}\,(1 + t(f_{i-1})) \quad where \quad f : L' \longrightarrow L,\ i \in [2 \cdots N], \tag{5.7}$$

where f_1 is the minimum luminance we want to encode (10^-5 cd/m^2 in our case) and N is selected so that f_N is larger than the maximum luminance to be encoded (10^{10} cd/m^2). Note that $cvi(L)\cdot L$ gives an absolute detection threshold in cd/m^2. The values of f_i give the luminance value associated with particular luma value i, that is the inverse mapping from luma to luminance. To find a forward mapping function, which we denote with $PU : L \longrightarrow L'$, we use the values of f as a lookup table and find the nearest (or interpolated) index i for a given luminance value L. For a more information on the formulation of this problem, refer to [Mantiuk *et al.*, 2006a] and Section 10.1.

Ideally, we would like our PU encoding to be backward-compatible with the sRGB nonlinearity (Equation 5.3), meaning that it should result in similar luma values within the dynamic range of a CRT display, while still retaining perceptual uniformity. We achieve this by minimizing the squared difference between both encodings within the range $0.1-80$ cd/m^2 with respect to three parameters m, s and L_{a-min}:

$$\sum_{L=0.1}^{80} \left((s\, PU(L, L_{a-min}) + m) - l^{-1}(L)\right)^2, \tag{5.8}$$

where the summation is performed for 256 logarithmically distributed luminance values L, $l^{-1}(L)$ is the inverse of Equation 5.3, and $PU(L)$ is the inverse of Equation 5.7. The result of such fitting together with the sRGB nonlinearity is shown in Figure 5.7. The fit is not perfect, as the sRGB nonlinearity does not fully agree with the cvi function. Note that neither of the parameters m and s affect our initial assumption since the differentials of the PU encoding are still proportional to the detection thresholds. The parameter s can be understood as the absolute sensitivity factor, which in fact varies among observers. By performing the optimization we implicitly assume the same sensitivity as the sRGB encoding. The other parameter m adjusts the absolute encoding response to fit to sRGB.

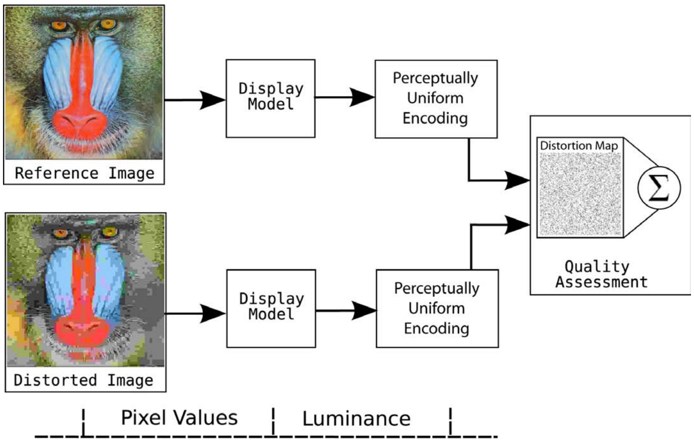


Figure 5.6: *Data flow diagram of the extended metrics for typical 8-bit images. Pixel values are converted to luminance and re-encoded with PU encoding before quality assessment.*

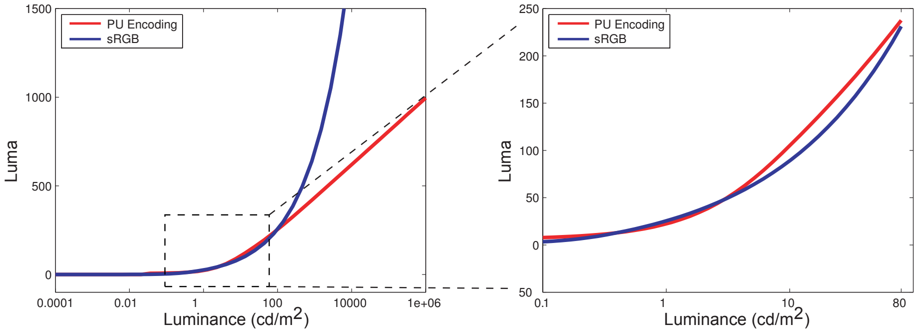


Figure 5.7: *The best fit of PU encoding to sRGB within the range* $0.1 - 80$ *cd/m*^{2} *in a least squares sense. Resulting curve is shown along the entire dynamic range (left), and only within the range that is considered for optimization (right).*

We store the resulting PU encoding as a look-up table, rather than trying to fit an analytic function. A look-up table offers better accuracy and is usually faster to compute than power or logarithmic functions approximating such encodings.

The data flow diagram of the extended metrics is given in Figure 5.6. Similar to non-extended metrics, the input is a pair of reference and distorted images. Both images are converted to display luminance values using the response function of the display on which the images are viewed. Next, the PU encoding transforms the luminance values into perceptually uniform pixel values. At the final quality assessment step, no modification on the metric part is necessary since the PU encoding merely provides perceptually uniform pixel values, which was the metric's assumption in the first place (Section 5.1).

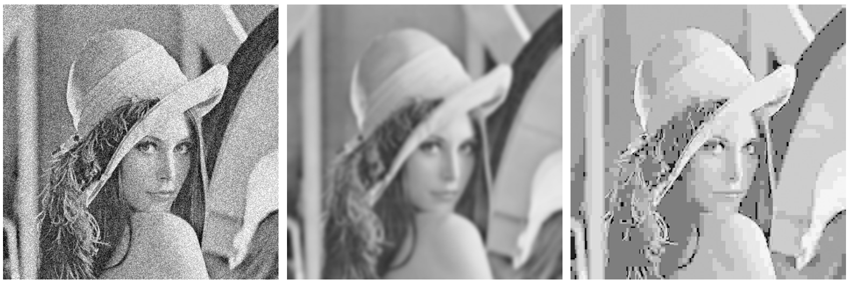


Figure 5.8: *Sample images from our validation test set. We consider random pixel noise (left), gaussian blur (center) and JPEG compression (right) as distortion types.*

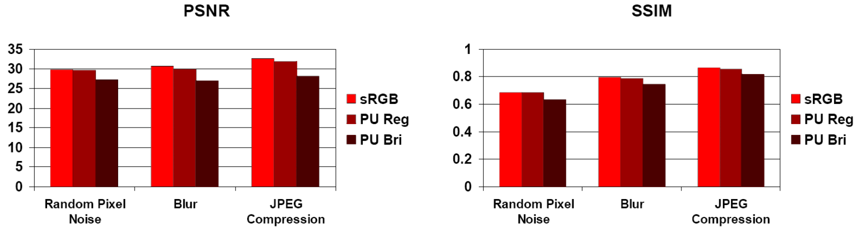


Figure 5.9: *Backward-compatibility with sRGB encoding. The average PSNR (left) and SSIM (right) responses of PU encoded images for different distortion types provides a good match to corresponding sRGB encoded images.*

5.7 Validation of Backwards Compatibility

We validate the compatibility of the PU encoding with the sRGB nonlinearity by comparing the extended and non-extended metric responses for a set of images viewed on a CRT display ($0.1 - 80cd/m^2$). The test set of distorted images is generated by converting the reference images to display luminance values and applying a distortion which can be either of the following types: random pixel noise, gaussian blur or JPEG compression (Figure 8.1). Each type of distortion is applied to 3 reference images at 2 different levels. The image luminance is converted to pixel values using sRGB and PU encodings, and the quality of the distorted images in both cases are assessed by PSNR and SSIM. Figure 5.9 shows the average responses for both extended and non-extended metrics separately for each type of distortion. We observe that the match between the responses is not exact, since our optimization procedure does not result in a perfect fit of PU encoding to sRGB nonlinearity (Figure 5.7). Still, the difference between extended and non-extended metric responses are quite low (< 1 dB for PSNR and < 0.01 for SSIM), indicating that they can be used interchangeably for typical CRT dynamic range if small deviations in metric responses are acceptable.

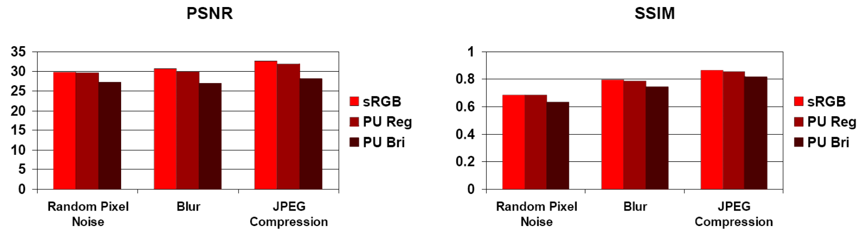


Figure 5.10: *Image quality on bright display. The pixel values of sRGB encoded images are the same for both regular* $(1 - 100cd/m^2)$ *and bright* $(10 - 1000cd/m^2)$ *displays. PU encoding successfully accounts for the effect of display brightness.*

5.8 Quality Assessment for Bright Displays

The subjective experiment in Section 5.2 revealed that distortions of the same type and magnitude appear more annoying on a bright display than a regular one. In this section we show that the extended metrics can correctly predict this effect, while non-extended metrics fail to do so. In parallel with the subjective study, we simulate the brightness of an LDR image on two hypothetical displays: a regular $(1 - 100\ cd/m^2)$ and a bright display $(10 - 1000\ cd/m^2)$, both with the same dynamic range $(1 : 100)$. The resulting luminance values from both display models are transformed to perceptually uniform pixel values with the proposed encoding.

In Figure 5.10, we compare the metric predictions for sRGB encoded images side by side with the extended metric responses for both display models. Note that the pixel values generated by sRGB nonlinearity are exactly the same for both displays, and consequently the quality estimates are also the same. On the other hand, quality of the PU encoded images viewed on the bright display are noticeably lower than the quality of the same images viewed on the regular display, in agreement with the outcome of our subjective experiment.

5.9 Quality assessment of HDR Images

Unlike 8-bit images that store gamma corrected code values tailored towards particular display devices, the content of an HDR image is related to the actual photometric characteristics of the scene it depicts, which in turn directly correspond to physical luminance. In order to get meaningful responses from PSNR and SSIM when comparing a pair of HDR images, physical luminance of both images need to be converted to perceptually uniform pixel values. The use of sRGB encoding for HDR images brings in an ambiguity in the choice of the white point value. The straightforward approach of setting the white point to the maximum luminance of the image generally results in suppression of details in dark image regions. Instead, the logarithmic function is a simple and often used approximation of the HVS response along the entire visible luminance

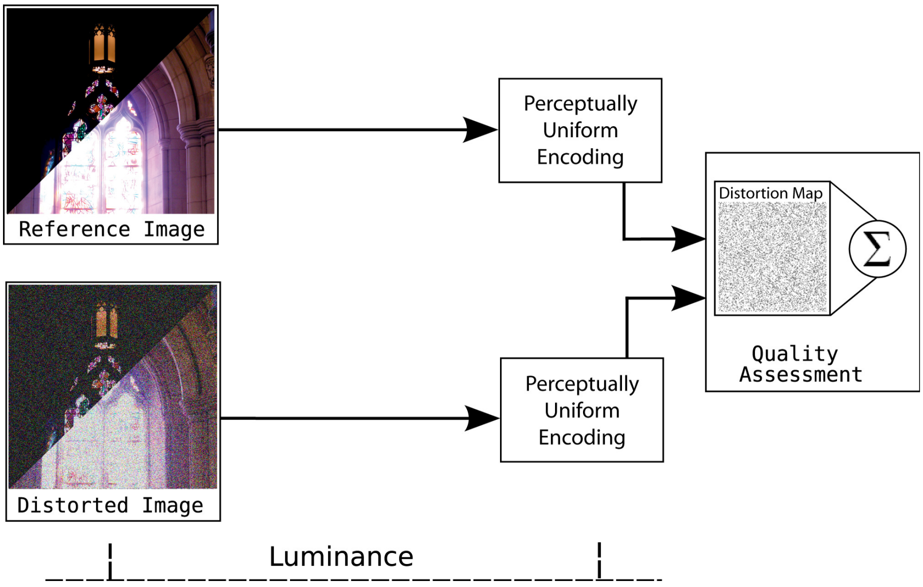


Figure 5.11: *Data flow diagram of the extended metrics for HDR images. HDR images are scene referred and store luminance, which is directly converted to percepually uniform pixel values by the PU encoding*

range. Although logarithmic encoding adheres to the Weber-Fechner law (Section 5.3), it provides a very coarse approximation and does not predict the loss of sensitivity for the low light conditions. These shortcomings can be avoided by employing the PU encoding to generate perceptually uniform pixel values for HDR images. The data flow of the extended "HDR metrics" is shown in Figure 5.11. Since HDR images already contain physical luminance information, the use of a diplay model is not necessary.

5.10 Conclusion

We proposed an extension to two popular image quality metrics, namely PSNR and SSIM, that makes them capable of handling all luminance levels visible to the human eye, without altering their response at the dynamic range of a typical CRT display. Our extension consists of transforming image luminance to perceptually uniform pixel values, that are optimized to fit gamma correction nonlinearity within the range from 0.1 to 80 cd/m^2 in a least squares sense. The proposed extension does not impose any changes on the quality metric part. Another consequence of this modularity is that it can potentially be applied to any quality metric that takes gamma corrected pixel values as input.

In the future, we would like to validate the metric responses for HDR images through subjective experiments. We are also interested in exploring the application of our extension to other quality metrics.

Chapter 6

Dynamic Range Independent Image Quality Assessment

In this section we present a second image quality assessment technique, focusing on comparing image pairs with different dynamic ranges motivated by the recent trends towards HDR imaging. In recent years we have witnessed a significant increase in the variation of display technology, ranging from sophisticated HDR displays [Seetzen *et al.*, 2004] and digital cinema projectors to small displays on mobile devices. In parallel to the developments in display technologies, the quality of electronic content quickly improves. For example luminance and contrast values, which are encoded in the so-called HDR images [Reinhard *et al.*, 2005], correspond well with real world scenes. HDR images are already being utilized in numerous applications because of their extra precision, but reproduction of these images is only possible by adjusting their dynamic range to the capabilities of the display device using tone mapping operators (TMO) [Reinhard *et al.*, 2002; Durand and Dorsey, 2002; Fattal *et al.*, 2002; Pattanaik *et al.*, 2000]. With the proliferation of new generation display devices featuring higher dynamic range the problem of enhancing legacy 8-bit images arises, which requires the so-called inverse tone mapping operators (iTMO) [Rempel *et al.*, 2007; Meylan *et al.*, 2007]. An essential, but yet unaddressed problem is *how to measure the effect of a dynamic range modification on the perceived image quality.*

Typical image quality metrics commonly assume that the dynamic range of compared images is similar [Daly, 1993; Lubin, 1995a; Wang and Bovik, 2002]. They predict visible distortion using the measures based on the magnitude of pixel intensity or normalized contrast differences between both input images, which become meaningless when input images have significantly different contrast or luminance ranges. However, when we look at images on a computer screen or even on traditional photographs we often have an impression of plausible real world depiction, although luminance and contrast ranges are far lower than in reality. So, *the key issue in image reproduction is not obtaining an optical match, but rather plausible reproduction of all important image features and* *preserving overall image structure*. Such features improve the discrimination and identification of objects depicted in the image, which are important factors in image quality judgment [Janssen, 2001]. The processed image structure can be affected by introducing visible artifacts such as blur, ringing, ghosting, halo, noise, contouring and blocking, which distort structure of the original image and may degrade the overall impression of image quality.

In this chapter we present a novel image quality metric that can compare a pair of images with significantly different dynamic ranges. We call this metric as *dynamic range independent visible differences predictor*, or *DRIVDP* for short. Our metric employs a model of the HVS, and its main contribution is a new concept of visible distortions based on the visibility of image features and the integrity of image structure (Section 6.2). DRIVDP generates a distortion map, which signalizes the loss of visible features, the amplification of invisible features, and reversal of contrast polarity (Section 6.3). All these distortions are considered at various scales and orientations, which correspond to the visual channels in the HVS (Chapter 2.4). Novel features of DRIVDP are tested (Section 8.2), and the overall metric performance confirmed in a psychophysical study (Section 6.5). We demonstrate application examples of our metric to predict distortions in feature visibility introduced by the state-of-the-art TMOs (Section 6.6.1) and inverse-TMOs (Section 6.6.2). Also, we analyze the influence of display dynamic range on the visibility of such distortions for three different displays (Section 6.6.3).

6.1 Background

Image quality evaluation is important in many applications such as image acquisition, synthesis, compression, restoration, enhancement, reproduction, and is relatively well covered in a number of textbooks [Winkler, 2005; Wang and Bovik, 2006; Wu and Rao, 2005]. Three important metric categories can be distinguished: metrics measuring contrast distortions, changes in the image structure, and judging visual equivalence between images. In this section we discuss all these metric categories from the standpoint of their ability to handle image pairs of significantly different dynamic ranges.

The most prominent contrast distortion metrics such as the *visible difference predictor* (VDP) [Daly, 1993] and the *Sarnoff visual discrimination model* (VDM) [Lubin, 1995a] are based on advanced models of the HVS and are capable of capturing just visible (near threshold) differences or even measuring the magnitude of such differences and scale them in JND (just noticeable difference) units. While these metrics are designed for LDR images, Mantiuk et al. [2005] proposed an HDR extension of VDP, that can handle the full luminance range visible to the human eye. Similar capabilities demonstrates also iCAM06 [Kuang *et al.*, 2007a], which additionally models important aspects of color appearance. While, the iCAM06 framework has been mostly applied in tone mapping applications, it has clear potential to compute HDR image difference statistics and to derive from them image quality metrics. Recently, Smith et al. [2006] proposed an objective tone mapping evaluation tool, which is focused on measuring suprathreshold contrast distortions between the source HDR image and its tone mapped LDR version. The problem with this metric is that it is based on the contrast measure for neighboring pixels only, which effectively means that its sensitivity is limited to high frequency details. Also, the metric calibration procedure has not been reported, while it may be expected that the metric may be excessively sensitive for small near-threshold distortions because the peak sensitivity is assumed for each luminance adaptation level instead of using contrast sensitivity function.

With the development of *structural similarity index metric* (SSIM) by Wang and Bovik [2002], an important trend in quality metrics has been established. Since the HVS is strongly specialized in learning about the scenes through extracting structural information, it can be expected that by measuring structural similarity between images, the perceived image quality can be well approximated. The SSIM proved to be extremely successful in many image processing applications, it is easy to implement, and very fast to compute. As the authors admit [Wang *et al.*, 2003], a challenging problem is to calibrate the SSIM parameters, which are quite "abstract" and thus difficult to derive from simple-stimulus subjective experiments as it is typically performed for contrast-based metrics. For this reason it is difficult to tell apart visible and non-visible (just below threshold) structure changes, even for multi-scale SSIM incarnations [Wang *et al.*, 2003]. The SSIM is sensitive for local average luminance and contrast values, which makes it inadequate for comparing LDR and HDR images. Recently, Wang and Simoncelli [2005] proposed the CW-SSIM metric, which in its formulation uses complex wavelet coefficients instead of pixel intensities employed in the SSIM. Since in CW-SSIM bandpass wavelet filters are applied, the mean of the wavelet coefficients is equal to zero in each band, which significantly simplifies the metric formulation with respect to the SSIM and makes it less sensitive for uniform contrast and luminance changes. However, this reduced sensitivity concerns rather small changes of the order 10–20%, which are not adequate for comparing HDR and LDR images.

An interesting concept of *the visual equivalence predictor* (VEP) has been recently presented by Ramanarayanan et al. [2007]. The VEP is intended to judge whether two images convey the same impression of scene appearance, which is possible even if clearly visible differences in contrast and structure are apparent in a side-by-side comparison of the images. The authors stress the role of higher order aspects in visual coding, but developing general computational model for the VEP is a very difficult task. The authors show successful cases of the VEP models for different illumination map distortions, which also requires some knowledge about the scene geometry and materials. While the goals of VEP and our metric are different, both approaches tend to ignore certain types of visual differences, which seem to be unimportant both for the scene appearance and image structure similarity judgements.

DRIVDP can be considered as a hybrid of contrast detection and structural similarity metrics. Careful HVS modeling enables precise detection of only visible contrast changes, but instead of reporting such changes immediately as VDM, HDRVDP, and VDM metrics, we use the visibility information to analyze only visible structure changes. We distinguish three classes of structure changes, which provides with additional insight into the nature of structural changes compared to SSIM. Finally, what makes our approach clearly different from existing solutions is the ability to compare images of drastically different dynamic ranges, which broadens the range of possible applications.

6.2 Image Distortion Assessment

Instead of detecting contrast changes, our metric is sensitive to three types of structural changes:

Loss of visible contrast is signalized when a contrast that was visible in a reference image becomes invisible in a test image. This typically happens when a TMO compresses details to the level that they become invisible.

Amplification of invisible contrast is signalized when a contrast that was invisible in a reference image becomes visible in a test image. For instance, it can happen when contouring artifacts starts to appear due contrast stretching in the inverse TMO application.

Reversal of visible contrast is signalized when a contrast is visible in both a reference and a test images, but it has different polarity. This can be observed for strong image distortions, such as clipping or salient compression artifacts. An intuitive illustration of the three types of distortions is shown in Figure 6.1

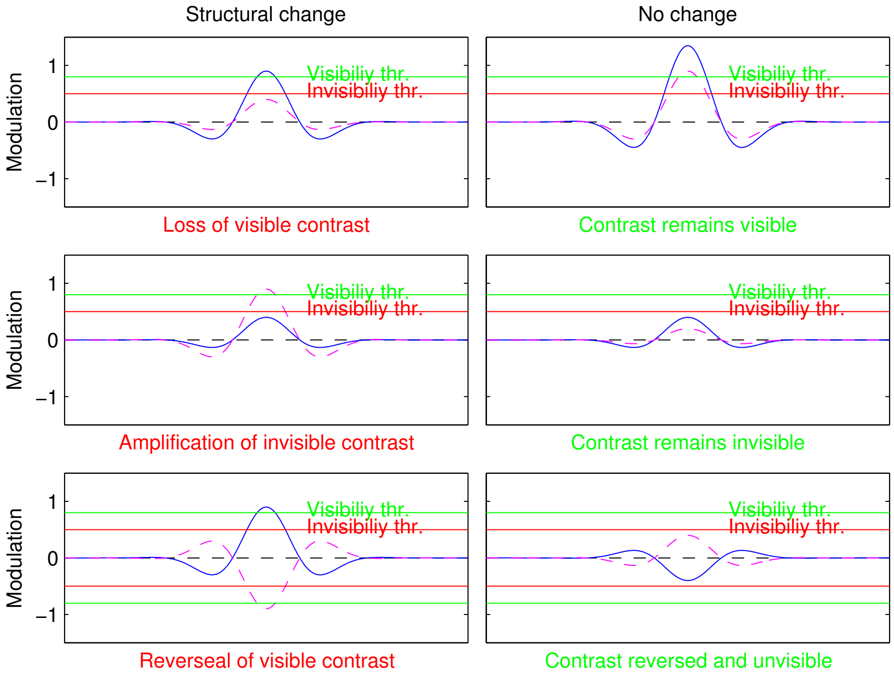


Figure 6.1: *Several cases of contrast modification, that DRIVDP classifies as a structural change (left) or a lack of structural change (right). Blue continuous line – reference signal; magenta dashed line – test signal. For the explanation of visibility and invisibility threshold (50% probability) refer to the text and Figure 6.4.*

Note that this formulation makes DRIVDP invariant to differences in dynamic range or to small changes in the tone-curve.

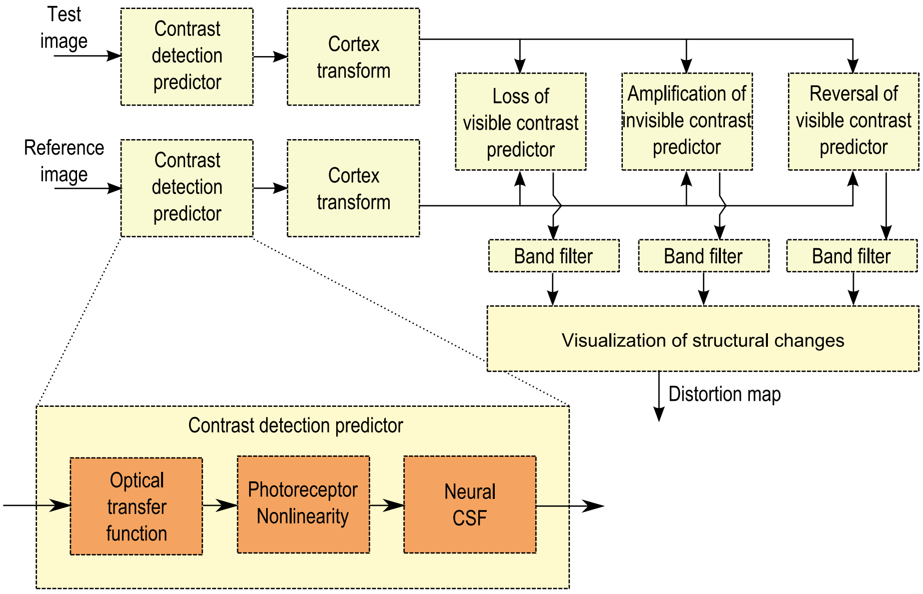

Test image
Contrast detection predictor
Cortex transform
Reference image
Contrast detection predictor
Cortex transform
Loss of visible contrast predictor
Amplification of invisible contrast predictor
Reversal of visible contrast predictor
Band filter
Band filter
Band filter
Visualization of structural changes
Distortion map
Contrast detection predictor
Optical transfer function
Photoreceptor Nonlinearity
Neural CSF

Figure 6.2: *The data flow diagram of our metric.*

Before we can detect any of the three types of distortions, we need to predict whether a contrast is visible or not. This is achieved with the metric that is outlined in Figure 6.2. The input to DRIVDP are two luminance maps, one for a reference image (usually an HDR image), and one for a test image (usually an image shown on the display). 8-bit images must be transformed using the display luminance response function to give actual luminance values shown on a screen. In the first step we predict detection thresholds and produce a perceptually normalized response map, in which the amplitudes equal to 1 correspond to the detection threshold at $P_{det} = 75\%$ (1 JND). Although several such predictors has been proposed in the literature, we found the HDRVDP detection model [Mantiuk *et al.*, 2005], designed especially for HDR images, the most appropriate. The predictor takes into account light scattering in the eye's optics (Chapter 2.1), nonlinear response of the photoreceptors (Chapter 10.1)and spatial-sensitivity changes due to local adaptation (Chapter 2.2).

To ensure accurate predictions, we calibrated the HDRVDP detection model with the ModelFest [Watson, 2000] measurements. The ModelFest data set was collected in a number of different laboratories to enhance both the generality and accuracy, and was especially designed to calibrate and validate vision models. Figure 6.3 shows a few examples of the detection probability maps for stimuli below, at and above the detection threshold. All results were generated by setting the pixels per visual degree to 120, and observer distance to $2m$. The model fitting error for 0.25% peak sensitivity was below 2dB contrast units. The errors were the largest for the stimuli "GaborPatch14" and "Dipole32", for which our predictor was too sensitive.

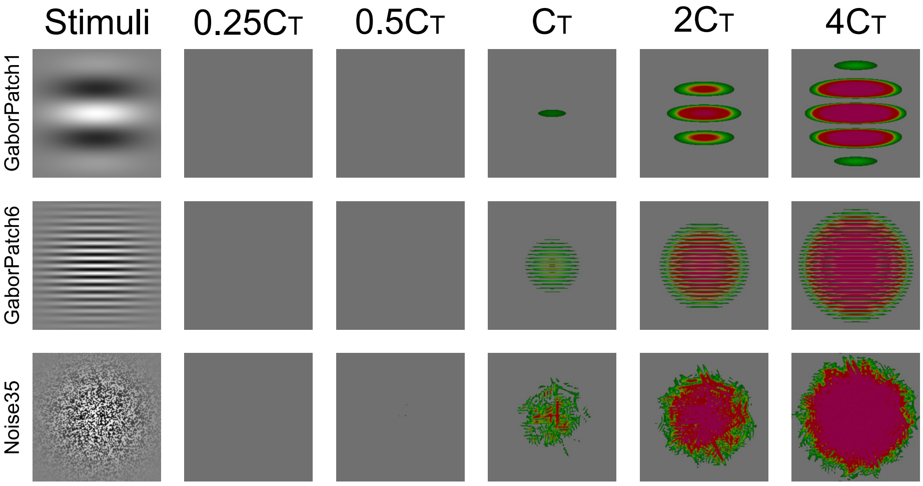


Figure 6.3: *The output of the detection predictor for the selected ModelFest stimuli at 0.25, 0.5, 1, 2 and 4 times the detection threshold, C_T. The first column shows the original stimuli at high contrast. The predictor is well calibrated if the visible contrast starts to be signalized in the C_T column.*

In the second step, we split the perceptually normalized response into several bands of different orientation and spatial bandwidth. We employ the cortex transform [Watson, 1987] with the modifications from [Daly, 1993] (Chapter 10.4). Then, to predict three types of distortions separately for each band, we compute conditional probabilities of

$$\begin{array}{ll} \text{loss of visible contrast:} & P^{k,l}_{loss} = P^{k,l}_{r/v} \cdot P^{k,l}_{t/i}, \\ \text{amplification of invisible contrast:} & P^{k,l}_{ampl} = P^{k,l}_{r/i} \cdot P^{k,l}_{t/v}, \\ \text{and reversal of visible contrast:} & P^{k,l}_{rev} = P^{k,l}_{r/v} \cdot P^{k,l}_{t/v} \cdot R^{k,l}, \end{array} \tag{6.1}$$

where k and l are the spatial band and orientation indices, the subscript $r/\cdot$ denotes reference and $t/\cdot$ test image, the subscript $\cdot/v$ visible and $\cdot/i$ invisible contrast. R equals 1 if the polarity of contrast in the reference and test images differ:

$$R^{k,l} = \left[C^{k,l}_r \cdot C^{k,l}_t < 0 \right]. \tag{6.2}$$

For simplicity we omit the pixel indices (x, y). The above formulation assumes that that contrast detection process is performed in the visual system separately for each visual channel.

The probabilities $P_{\cdot/v}$ and $P_{\cdot/i}$ are found from the detection probabilities, as shown in Figure 6.4. The visual models commonly assume that a contrast is visible when it is detectable ($P_{det}{\geq}75\%$), as in the two alternative forced choice (2AFC) experiments. We found this assumption to be too conservative, since complex images are never as scrutinously observed as stimuli in such experiments. Therefore, we require a contrast to be detected with a higher probability, to be regarded as visible. From our empirical study on a series of simplified stimuli, we found that a reliable predictor of visible contrast is given by shifting the psychophysical function, so that a contrast magnitude is *visible* with 50%

probability, if it can be *detected* by our predictor with 95% probability (about 2 JND), as shown in Figure 6.4. The probability of invisible contrast is given by the negation of the probability of detection.

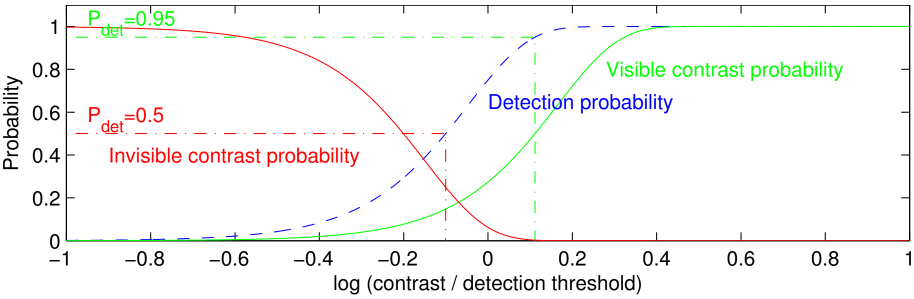


Figure 6.4: *Probability functions for a normalized contrast magnitude being visible (green) and invisible (red).*

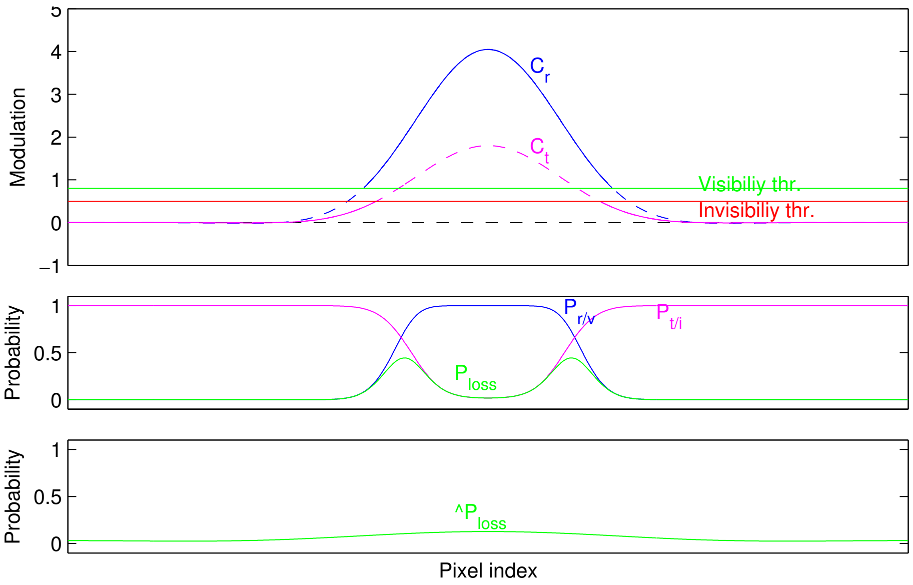


Figure 6.5: *The probability rules may produce response that do not belong to a particular frequency band. Top pane: although a contrast magnitudes are well above visibility threshold, there is a small part in which contrast is visible in the reference image (C_r) but not visible in a test image (C_t). Center pane: this triggers higher values of the P_{loss} in these regions. Bottom pane: the spurious responses can be eliminated with a band-pass filter.*

The rules from Equation 6.1 contain the nonlinear operators, therefore the resulting probability map $P^{k,l}_{\cdot}$ can contain features of spatial frequency that do not belong to a particular subband. This leads to spurious distortions, as shown in Figure 6.5. To avoid this problem, each probability map is once more filtered using the corresponding cortex filter $B^{k,l}$:

$$\hat{P}^{k,l}_{loss} = \mathscr{F}^{-1}\left\{\mathscr{F}\{P^{k,l}_{loss}\}\cdot B^{k,l}\right\}, \tag{6.3}$$

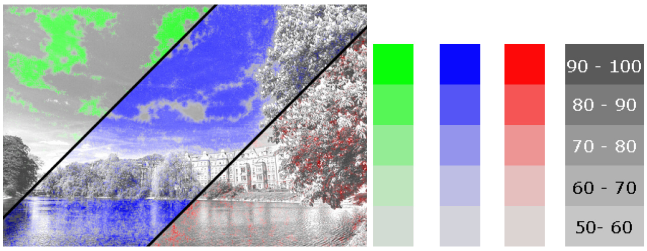


Figure 6.6: *Three distortion maps shown partially (left). We arbitrarily chose green for loss of invisible contrast, blue for amplification of invisible contrast, and red for reversal of visible contrast. The saturation of each color indicates the magnitude of detection probability, as shown in the respective scales.*

where $\mathscr{F}$ and $\mathscr{F}^{-1}$ are the 2D Fourier and inverse Fourier transforms, and the formulas for $B^{k,l}$ can be found in the Appendix.

Assuming that detection of each distortion in each band is an independent process, the probability that a distortion will be detected in any band is given by:

$$P_{loss} = 1 - \prod_{k=1}^{N} \prod_{l=1}^{M} \left(1 - \hat{P}_{loss}^{k,l}\right). \tag{6.4}$$

The probability maps P_{ampl} and P_{rev} are computed in a similar way.

Unlike typical HVS-based contrast difference predictors, DRIVDP does not model visual masking (decrease in sensitivity with increase of contrast amplitude). Since our metric is invariant to suprathreshold contrast modifications, visual masking does not affect its result. If we compare two visible contrast stimuli, as the ones shown in top-right pane of Figure 6.1), the visual masking can predict by how many JNDs their amplitudes differ. The contrast difference is not relevant for our metric, therefore there is no need to estimate the magnitude of suprathreshold contrast in JND units.

6.3 Visualization of Distortions

The multitude of distortion types detected by DRIVDP makes visualization of the outcome on a single image a challenging task. We employ an in-context distortion map [Daly, 1993] approach to provide with an overview of distortions, but also introduce a custom viewer application for more detailed inspections.

To generate the in-context map, luminance of the distorted image is copied to all three RGB channels, and each channel is scaled by the detection probabilities of corresponding distortion type. We observed that using multiple colors for each type of distortion makes it is hard to memorize the association of each color to the correct distortion type. We also found that in regions where multiple distortions overlap, the simple approach of blending the colors makes the final map less intuitive by increasing the number of colors. We therefore show only the distortion with the highest detection probability at each pixel location. We

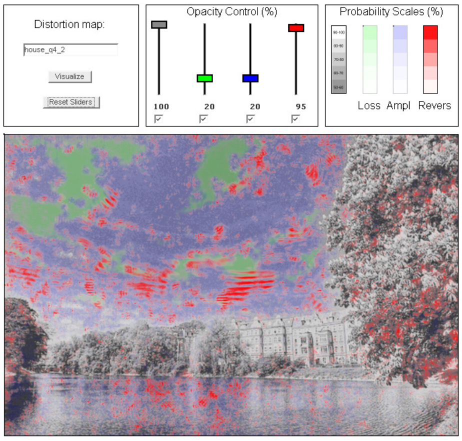


Figure 6.7: *Our distortion viewer. Users can adjust opacities of distortion maps and background image. The respective scales (top right) are adjusted accordingly by the tool. In this example setting, the user emphasizes contrast reversal, while keeping the other distortions barely visible.*

arbitrarily chose **green** for loss of invisible contrast, **blue** for amplification of invisible contrast, and **red** for reversal of visible contrast (Figure 6.6).

In cases where the test image is heavily distorted the in-context map representation may become too cluttered, and there may be significant overlaps within different distortion types. On the other hand, one may simply be interested in a closer examination of each distortion type present in the image. Using the viewer application one can dynamically set the opacity values of distortion types and the background image to a legible configuration, that allows to investigate distortions separately (Figure 6.7). In the rest of this chapter, all metric responses are presented as in-context maps. The viewer application can be used for any further investigation of the results.

6.4 Evaluation and Results

In the following sections, we present results that demonstrate advantages of our metric to previous work

6.4.1 Dynamic Range Independence

We claim that DRIVDP generates meaningful results even if the input images have different dynamic ranges, in addition to the case where both have the same dynamic range. In Figure 6.8, we show the distortion maps resulting from the comparison of all variations of an HDR and LDR image. The LDR image is generated by applying a compressive power function to the HDR reference (more sophisticated tone-mapping operators are discussed in Section 6.6.1). We always distort the test image by locally adding random pixel noise, whose magnitude

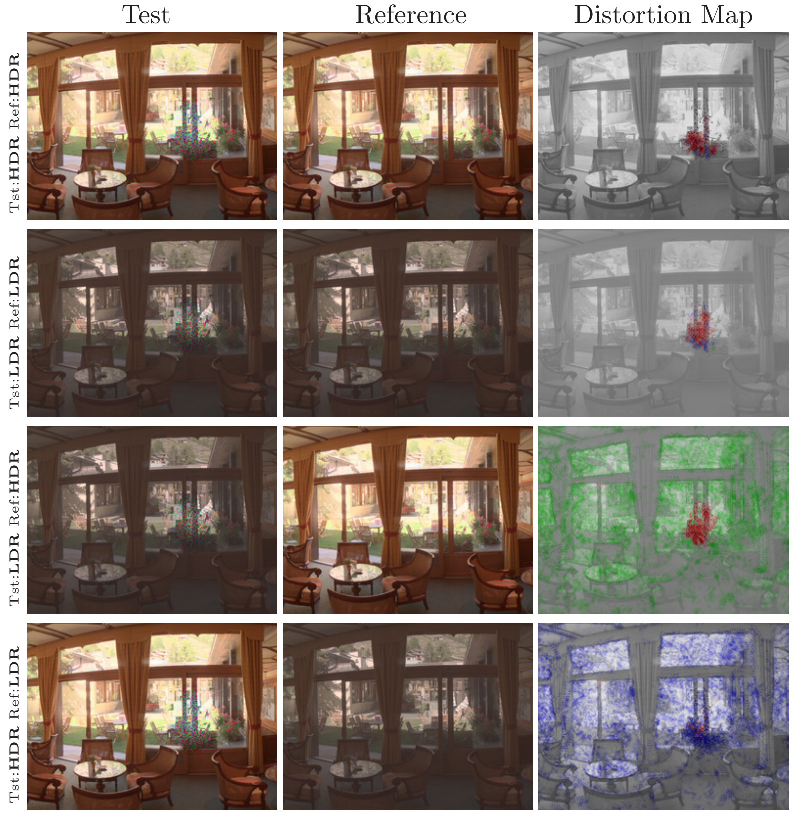


Figure 6.8: *Comparing images with different dynamic ranges. While distortions caused by the local distortion are visible in all results, in the LDR-HDR and HDR-LDR cases, additional visible contrast loss and invisible contrast amplification can be observed due to the contrast lost through dynamic range compression. HDR images are tone-mapped using Reinhard's photographic tone reproduction for printing purposes.*

is modulated with a gaussian that has its peak at the center of the distorted region.

Randomly distributed pixels in the distorted region both introduce previously non-existent contrast and invert the polarity of the contrast proportional to the magnitude of the distortion. Consequently, for both HDR-HDR and LDR-LDR cases (first two rows) our metric reports visible contrast reversal and amplification of invisible contrast confined in the distorted region. Similar responses are also observed in LDR-HDR and HDR-LDR cases. Additionally, a comparison of the distorted LDR image with an HDR reference yields to an overall loss of visible contrast spread across the entire image, indicating the effect of contrast compression applied to the test image (third row). When we compare the HDR test image with the LDR reference, visible contrast of the reference lost during compression manifests itself this time as amplification of invisible contrast in the distortion map (last row).

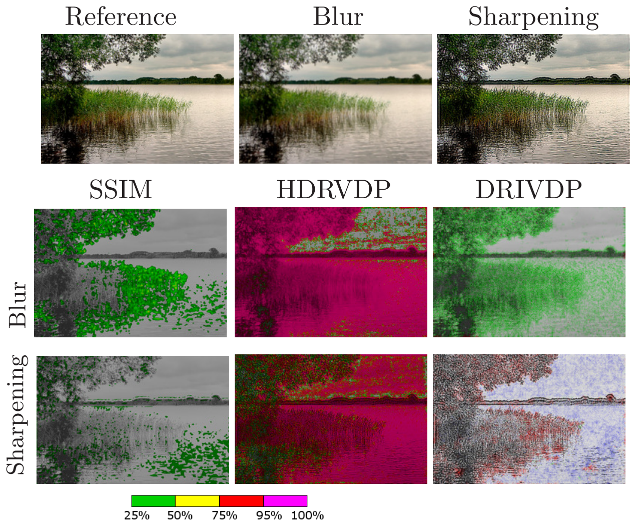


Figure 6.9: *The reference, blurred and sharpened test images (top row), and metric responses to blurring (middle row) and sharpening (bottom row). Color coding for SSIM and HDRVDP are given in the scale. Our metric is visualized as discussed in Section 6.3*

6.4.2 Comparison with Other Metrics

DRIVDP has two major advantages to the previous work: classification of distortion types, and dynamic range independence. In this section, we compare responses of our metric with a pair of state-of-the-art metrics, namely SSIM [Wang and Bovik, 2002] that predicts changes in the image structure, and HDRVDP [Mantiuk *et al.*, 2005] that is explicitly designed for HDR images. Figure 6.9 shows a side-by-side comparison of the three metrics where a blurred and a sharpened version of the reference was used as test image. The reference is an 8-bit image, which is linearized and converted to luminance for HDRVDP and our metric. The outcome of SSIM is a simple matrix of probability values with the same size as the input images, to which we applied HDRVDP's visualization algorithm to make it legible. The spatial distribution of the responses from all three metrics to blurring and sharpening is similar, with the overall tendency of HDRVDP's response being stronger (due to reporting all visible differences) and SSIM's response being weaker (due to the difficulty of calibration) than that of our metric.

The important difference between the proposed metric and others is the classification of distortion types. That is, in case of blurring DRIVDP classifies all distortions as a loss of visible contrast, confirming the fact that high frequency details are lost. On the other hand, in the sharpening case we observe contrast reversal and amplification of invisible contrast, both of which are expected effects of unsharp masking. Such a classification gives insight about the nature of the image processing algorithm and enables distortion-type-specific further processing.

The second major advantage of our metric is that it enables a meaningful comparison of images with different dynamic ranges (Section 6.4.1). We ran all three metrics on a test set, that is generated using a similar procedure as used for Figure 6.8, with the only difference being the use of gaussian blur as the distortion type. HDR images in the test set were calibrated to absolute luminance values of the scene, and were directly passed to both our metric and HDRVDP. For SSIM, we took the 10-base logarithm of the HDR images to compensate for the Weber law, and mapped them to pixel values within 0-255 to prevent an ambiguity in the dynamic range parameter of the metric.

Figure 6.10 shows a comparison of images with same dynamic range results in all three metrics reporting distortions in the blurred region with slightly different magnitudes (first two rows). One important difference between our metric's and HDRVDP's responses is, that the distorted area reported by HDRVDP is larger than that of our metric's. HDRVDP simply reports all visible differences of the blurred test images with respect to their references, while DRIVDP ignores the differences in the periphery of the gaussian, where the magnitude of the blur is weaker and details in the distorted image are still visible. This example shows a case where our metric provides complementary information to well established metrics. In the different dynamic range case, the distortion maps of SSIM and HDRVDP are entirely dominated by contrast change due to the dynamic range compression (last two rows). Similar to the results for different dynamic range case in Figure 6.8, DRIVDP reports an overall loss of visible contrast in the LDR-HDR case, and an overall amplification of invisible contrast in the HDR-LDR case, both due to the dynamic range compression. These responses, however, do not mask the response at the blurred region, as they do with the other metrics.

6.5 Validation

Validation of the metric is performed by comparing the metric responses to subjective distortion assessments. We generated a test set containing permutations of 3 images of natural scenes, 3 types of distortions and 3 levels of distortions. Each subject evaluated the entire test set twice to ensure reliability, leading to 54 images per subject. Gaussian blur that produces visible contrast loss, and unsharp masking that mostly produces invisible contrast amplification were chosen as distortions. Another type of distortion was considered to specifically produce contrast reversal, where we calculate a bandpass image pyramid, invert the signs of a number of layers proportional to desired distortion level, and recombine the pyramid to get the distorted image. All distorted images were generated to dominantly produce a metric response of the desired type.

We asked 14 subjects within the ages $23-48$, with all nearly perfect or corrected vision, to *identify the type of distortion* they see on a number of test images. Possible answers were *blur*, *sharpening*, *contrast reversal* or *no distortion*. We assumed no prior knowledge of the subjects about the distortion types. Therefore, a short training section preceded the actual experiment, where subjects were shown a series of images that contain strong distortions of all three types, together with the correct distortion labels.

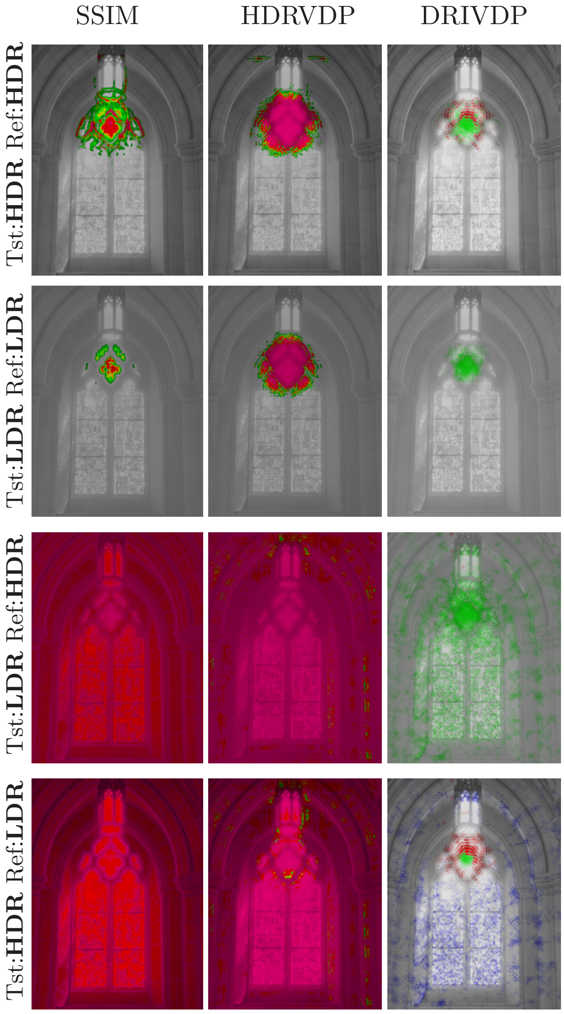


Figure 6.10: *A comparison of SSIM, HDRVDP and DRIVDP on all dynamic range combinations. Results for the same dynamic range case are comparable (first two rows), whereas in the different dynamic range case SSIM and HDRVDP responses are dominated by the dynamic range difference (last two rows).*

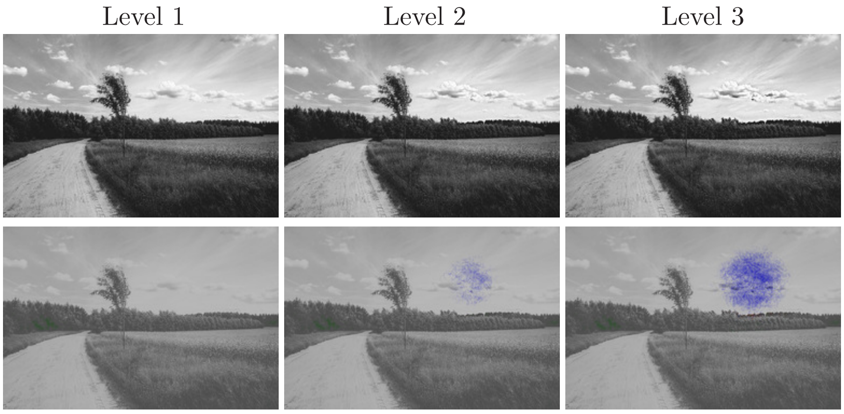


Figure 6.11: *A sample image from the validation set, showing three levels of sharpening (top row), and the corresponding metric responses (bottom row) increasing from left to right.*

In order to account for the variation of subject responses to different distortion magnitudes, we applied all distortions at three different levels, from which the first is selected to generate no metric response at all. The second level was chosen to generate a weak metric response of the desired type, where the detection probability at most of the distorted pixels is less than one. Similarly, the third level was chosen to generate a certain metric response in a noticeably large region. In our statistical analysis, we considered the first level as invisible, and the other two as equally visible. Since our metric is not intended to produce a single number, we restrained ourself from using an average of the detection probabilities within the distorted region.

First, we examined subject reliability by testing the stochastic independence of the consecutive iterations for each subject. Using the χ^2 test we obtained a $\chi^2(9)$ value of 739.105, where the value in parenthesis denotes the number of degrees of freedom. The corresponding $p - value$ was found to be $\ll 0.05$, indicating that the null-hypothesis can safely be rejected. The Cramer's V [Cramér, 1999], that measures the association between two categorical variables, is found to be 0.807 which is considered a large effect size. Next, we investigated the main effect of factors using the ANalysis Of VAriance (ANOVA) method (See [D'Agostino, 1972] for the use of ANOVA on nominal data). We found that distortion type and level to have a significant effect on the subject response ($F(2) = 179.96$ and $F(2) = 456.20$ respectively, and $p \ll 0.01$ for both). We also found that the test image factor ($F(2) = 4.97$ and $p = 0.02$) to have an effect on the final outcome, which is hard to avoid when experimenting with complex stimuli. Finally, we analyzed the statistical dependency between the subject and metric responses. For the null-hypothesis that these responses are independent, we found $\chi^2(9) = 1511.306$ and $p \ll 0.05$, showing that it is unlikely that the initial assumption holds. The corresponding Cramer's V of 0.816 signals a strong dependency between the metric and subject responses.

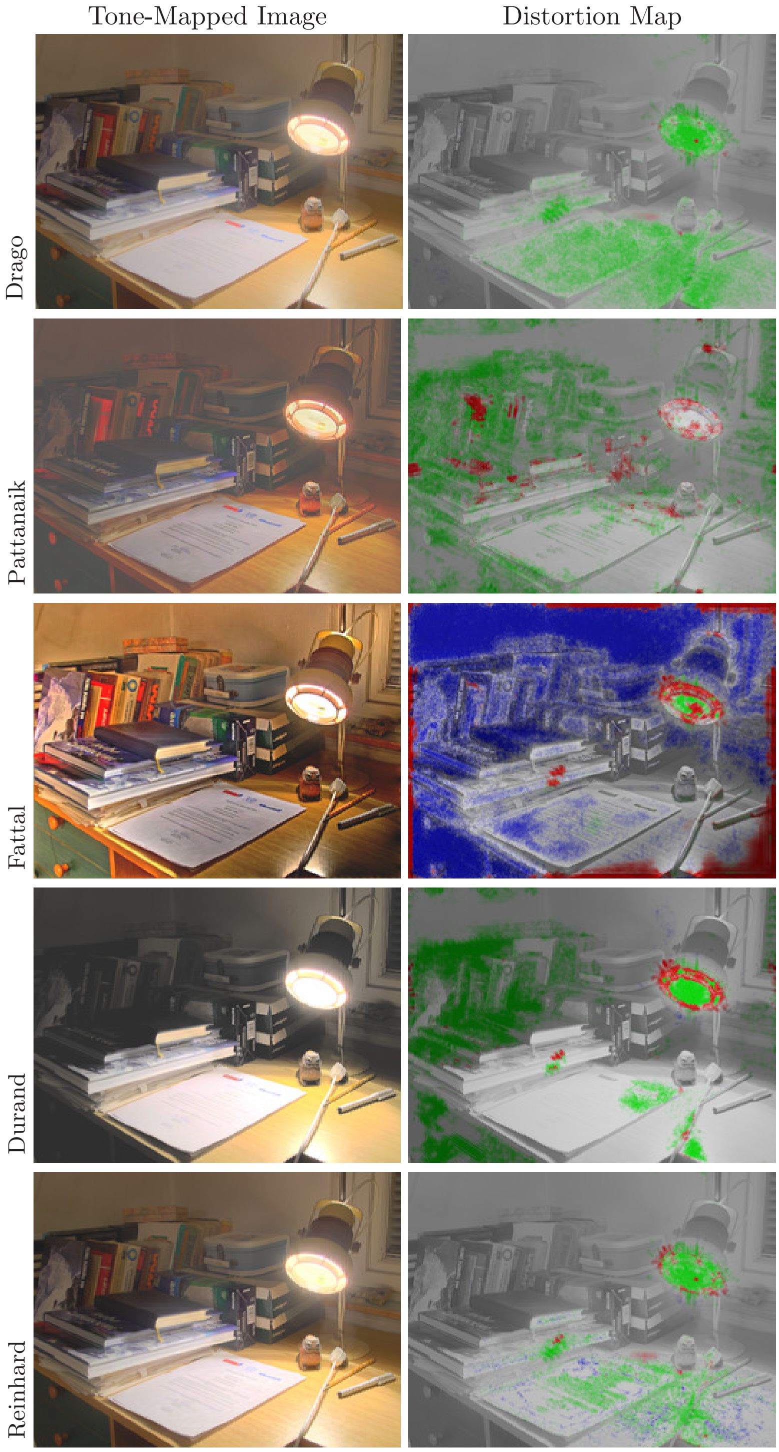


Figure 6.12: *Comparison of Tone-Mapping Operators*

6.6 Applications

In this section, we present several application areas of DRIVDP, where a comparison of images with different dynamic ranges is required.

6.6.1 Tone Mapping Operator Comparison

Tone mapping operators (TMO) are commonly used for contrast compression of HDR images to reproduce them properly on conventional media. This is a lossy process by definition. From a functional point of view, information reproduction capability of a TMO is a suitable measure of its performance. Figure 6.12 shows the comparison result of an HDR image with the corresponding tone mapped images. The luminance ranges of 0.24–89,300 and 0.1–80 cd/m^2 have been assumed for the original scene and displayed tone mapped image, respectively. Five TMOs (2 global and 3 local operators) have been considered: Drago's adaptive logarithmic mapping [2003a], Pattanaik's visual adaptation model [2000], Fattal's gradient domain compression [2002], Durand's bilateral filtering [2002], and Reinhard's photographic tone reproduction [2002].

For all studied TMOs certain detail loss can be observed in the brightest lamp region due to strong contrast compression. Pixel intensity clipping also causes visible contrast reversal in the lamp region, which is reported for some pixels as the strongest distortion. Drago's operator relatively well reproduces contrast in dark image regions and tends to wash out image details in bright regions due to logarithmic shape of the tone mapping curve. Pattanaik's operator, which is based on the sigmoid photoreceptor response (mostly adapted to the luminance levels at the illuminated table regions), tends to suppress strongly image details in dark regions, but also in very bright highlights. The detail amplification inherent for Fattal's operator can be seen in non-illuminated scene regions, which in real-world observation conditions are not visible due to insufficient HVS sensitivity. Our metric takes into account such sensitivity by modeling the dependence of contrast sensitivity function (refer to Equation 10.3) on luminance values in the HDR image. Durand's operator uniformly compresses lower spatial frequencies across the whole image, which means that resulting contrast loss will be more likely visible in dark display regions in which the HVS sensitivity is lower. The compression of low frequency features leads also to the reversal of visible contrast. The default parameters used for Reinhard's operator tend to excessively saturate bright image regions for this particular scene. Also, in the full size image it can be seen that contrast of certain pixels representing the table and paper page textures has been magnified due to local dodging and burning mechanism. Our results are consistent with the expected outcomes of the TMOs, indicating to the potential use of DRIVDP as a diagnostic tool for such algorithms.

6.6.2 Inverse Tone Mapping Evaluation

Recently, [Meylan *et al.*, 2007] and [Rempel *et al.*, 2007] attacked the problem of recovering the contrast in LDR images that has been clipped and/or compressed

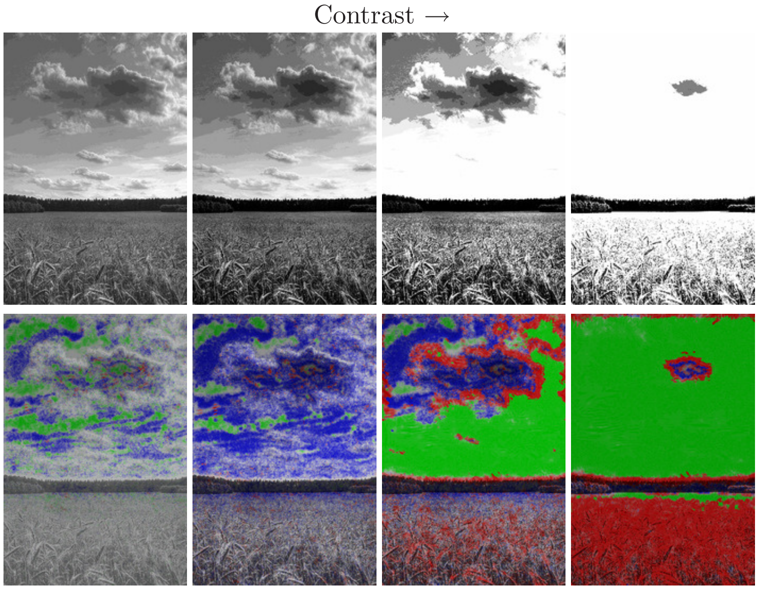


Figure 6.13: *Response of the metric to simple contrast stretching with clipping. Contrast is increased from left to right, which results in more clipping and generates stronger visible contrast loss and reversal responses.*

due to the limited dynamic range. These algorithms should be validated by costly subjective user studies to assess the plausibility of the results and the amount of visible artifacts [Akyüz *et al.*, 2007]. The latter task can be fulfilled much more efficiently by our metric.

The response of DRIVDP to simple contrast stretching with clipping is shown in Figure 6.13. To exaggerate the contouring artifacts, we use a 4-bit quantized version of the 8-bit reference as our test image. We observe that the more we increase image contrast, the more visible contrast in the bright sky region is lost, and invisible contrast in the darker horizon line is amplified, both due to clipping on both sides of the expanded image histogram. Our metric also reports contrast reversal on the boundaries within the visible and clipped contrast regions. In Figure 6.14, we show the comparison of an HDR image reconstructed by Ldr2Hdr [Rempel *et al.*, 2007] algorithm, with the reference LDR image image. The smooth contrast enhancement may result in loss of visible contrast around bright regions. However, the bright sun region of the reference image does not contain any high frequency details. Moreover, the visual glare caused by the sun effectively results in further smoothing of the region. Consequently, our metric does not report any loss of visible contrast. Finally, the increase in contrast due to stretching reveals some previously invisible details around the trees in the foreground, which is correctly reported by our metric. Contrast content amplified in bright regions, however, was already visible, and therefore is not interpreted as a structural change.

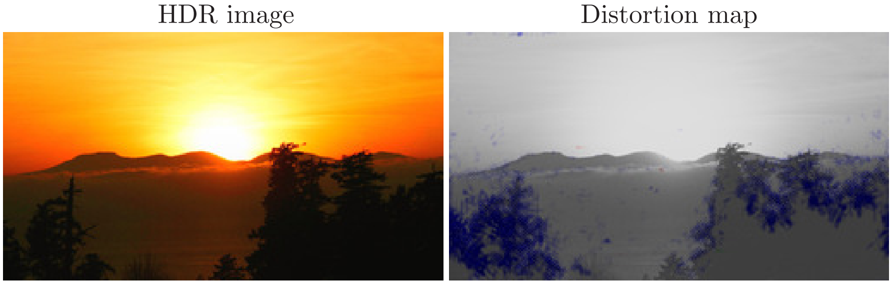


Figure 6.14: *HDR image generated by Ldr2Hdr algorithm (left), and the distortion map obtained by comparing the HDR image with the LDR reference (right). Both images are taken from the original author's website.*

6.6.3 Simulation of Displays

The highly diverse characteristics of today's display devices make an objective analysis of their reproduction capability an interesting problem. DRIVDP can be used as a measure on how well the structural information of the image is preserved, when it is viewed on different displays to ensure that important features of the image are preserved regardless to the display type.

In Figure 6.15 we show the distortion maps for an HDR reference image that is viewed on an BrightSide DR37-P HDR display ($2,005cd/m^2$), Barco Coronis 3MP LCD display ($400cd/m^2$), and a Samsung SGH-D500 cell phone display ($30cd/m^2$). To simulate the HDR and LCD displays, we apply the respective display response functions to image luminance values using a Minolta LS-100 luminance meter.

The results show that the HDR display faithfully reproduces most of the visible and invisible contrast. The small amount of distortion is expected, as even the dynamic range of the HDR display does not span the whole visible luminance range. The distortion map for the LCD display shows visible contrast loss in the outside region directly illuminated by sunlight. This luminance level exceeds the capabilities of the display device and therefore details are clipped. On the other hand, we observe invisible contrast amplification in parts of the darker interior region. This is because these regions in the reference image are so dark that the fine details at the chairs and floor are not visible. But since the LCD display is not capable of displaying such low luminance, those details are amplified above the visibility threshold. Finally, the cell phone display fails to reproduce most of the visible contrast, and hence we observe strong visible contrast loss in both the interior and exterior regions, as well as contrast reversal around the borders of the clipped regions.

6.7 Conclusion

We presented a quality assessment metric capable of handling image pairs with arbitrarily different dynamic ranges. The unique feature of DRIVDP is that it

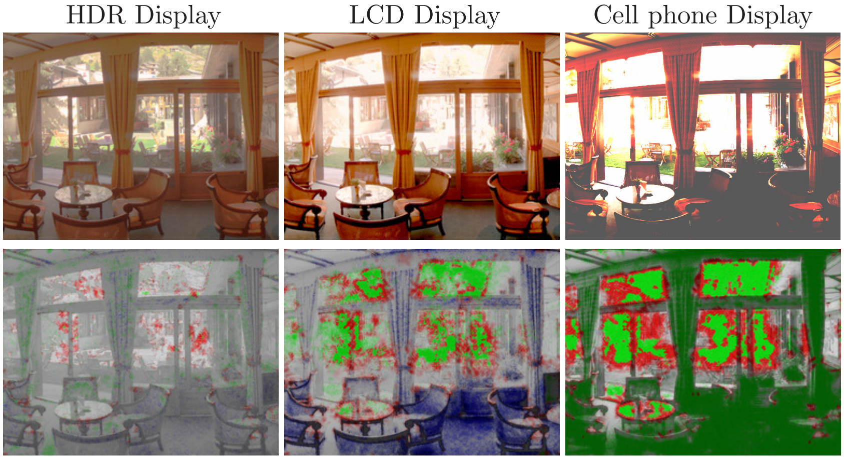


Figure 6.15: *Display Comparison. The brightness of the LCD (first row center) and Cell phone (first row right) display images are artificially enhanced for maximum detail visibility.*

detects distortions in the image structure and evaluates their visibility on any display device. This is achieved by applying models of the HVS that can handle the full visible range of luminance and their careful calibration. The metric gives also an insight into the nature of reported structural distortions by classifying them into three different categories, which can be conveniently visualized including image regions that are simultaneously affected by multiple distortion types. The metric predictions compare favorably with distortions perceivability by the human observers, and introduced distortion categorization has an intuitive meaning in terms of typical distortions introduced by image processing. In this work we have specifically chosen application examples that involve the structural difference evaluation between HDR and LDR images which has not been possible so far. However, due to guaranteed visibility of detected structural distortions and their unique categorization our metric may have many potential applications in evaluating image pairs of similar dynamic range as well.

As future work, we intend to test DRIVDP in medical applications which require faithful reproduction of details captured by HDR sensors in the displayed images. It would be also interesting to try our metric in watermarking applications, which require reproducing images on various media.

Part III

Video Quality Assessment

Chapter 7

Dynamic Range Independent Video Quality Assessment

In the final part of this dissertation we focus on video quality assessment. The treatment of video brings additional concerns from the HVS modeling perspective compared to image data, mainly because the spatiotemporal mechanisms are far more complicated than mere spatial mechanisms. In this chapter we present a video quality metric, based on our work presented in Chapter 6, that addresses these issues. We also show that the metric is useful in a number of computer graphics applications. We also present a comprehensive subjective validation study in Chapter 8.

The contributions of newly proposed Computer Graphics techniques are usually demonstrated through images, and more often through videos, in which the merit of the technique is apparent. The performance of, for example a new rendering method, can be assessed by comparing sequences rendered on one hand using the proposed method, and on the other hand a more precise, but slower reference method. The point of this comparison could be to show that the proposed method produces results comparable to the reference method, but much more efficiently. A similar evaluation process is also common in other subfields such as High Dynamic Range (HDR) Imaging. Evaluation of tone mapping operators, as well as compression methods for HDR video both involve a comparison of, respectively the tone mapped and compressed video, with the HDR reference sequence. In fact, assessment of the fidelity of a video sequence to a reference is a task common to numerous Computer Graphics techniques.

Formal subjective methods of video quality evaluation such as [ITU-T, 1999], where a Mean Opinion Score is computed by obtaining responses from multiple test subjects are often too laborious to be used on large sets of data. For the same reason the use of such methods in a feedback loop during development is not feasible; in fact most authors perform subjective evaluation only after the development of their algorithm is completed. Video Quality Metrics provide an objective means of comparing video sequences much faster than subjective methods by trading off accuracy of the prediction due to simplified modeling of visual perception. Simple metrics like PSNR, that rely solely on image pixel statistics fail to predict significant HVS properties like visual masking and contrast sensitivity. More sophisticated metrics [Winkler, 2005; Seshadrinathan and Bovik, 2010] on the other hand are not designed for HDR content. In the light of the recent trends towards HDR Imaging, the absence of HDR capable HVS models severely limits the use of these metrics in Computer Graphics context. Recently however, several *image* quality assessment metrics have been proposed, either designed specifically for HDR images [Mantiuk *et al.*, 2005], or that can compare image pairs with arbitrary dynamic range [Aydın *et al.*, 2008a]. However, simply using image quality metrics to evaluate each frame of a video sequence fails to reflect the temporal aspects of HVS mechanisms, typically resulting in underestimating the visibility of temporal artifacts such as flickering (Section 7.3, and Chapter 8).

A video quality metric specifically designed for Computer Graphics applications by addressing the aforementioned issues, could be used as a practical diagnostic tool and a quick alternative to subjective evaluation. We propose a *dynamic range independent* video quality metric that can compare a video pair of arbitrarily different dynamic ranges. The metric comprises a temporal HVS model, that accounts for major effects like luminance adaptation, contrast sensitivity dependency to both spatial and temporal frequencies, and similarly visual masking computed in spatiotemporal visual channels (Section 7.2). The results in Section 7.3 show that our metric predicts distortion visibility more accurately than previous video quality metrics and state-of-the-art image quality assessment methods applied to each video frame separately. The predictions of the proposed metric are also validated through a subjective study (Chapter 8). We show that our metric enables new applications of evaluating HDR video tone mapping and compression methods. We also demonstrate the comparison of videos rendered with different methods and quality settings, and assessment of the impact of dropped frames to perceived quality (Section 7.4).

7.1 Background

In this section we summarize previous work on objective video quality assessment and the use of video quality measures in Computer Graphics applications, and give some background on the temporal HVS mechanisms related to our metric.

7.1.1 Video Quality Assessment

The focus of the early work has been quality assessment of digitally coded video, mainly resulting from the observation that simple statistics like signal-to-noise ratio are not necessarily correlated with human vision. Van den Branden Lambrecht's Moving Picture Quality Metric (MPQM) [1996] utilizes a spatial decomposition in frequency domain using a filter bank of oriented Gabor filters, each with one octave bandwidth. Additionally two temporal channels, one low-pass (sustained) and another band-pass (transient) are computed to model visual masking. The output of their metric is a numerical quality index between $1 - 5$, similar to the Mean Opinion Score obtained through subjective studies. In a more efficient version of MPQM, the Gabor filter bank is replaced by the Steerable Pyramid [Lindh and van den Branden Lambrecht, 1996]. In later work targeted specifically to assess the quality of MPEG-2 compressed videos [van den Branden Lambrecht *et al.*, 1999], they address the space-time nonseparability of contrast sensitivity through the use of a spatiotemporal model. Another metric based on Steerable Pyramid decomposition aimed towards low bit-rate videos with severe artifacts is proposed by Masry and Hemani [2004], where they use finite impulse response filters for temporal decomposition.

Similarly, Watson et al. [2001] published an efficient Digital Video Quality metric (DVQ) based on the Discrete Cosine Transform. The DVQ models early HVS processing including temporal filtering and simple dynamics of light adaptation and contrast masking. Later they propose a relatively simple Standard Spatial Observer (SSO) based method [Watson and Malo, 2002], which, on the Video Quality Experts Group data set, is shown to make as accurate predictions as more complex metrics. Winkler [1999; 2005] proposed a perceptual distortion metric (PDM) where he introduced a custom multiscale isotropic local contrast measure, that is later normalized by a contrast gain function that accounts for spatiotemporal contrast sensitivity and visual masking.

Seshadrinathan and Bovik [2007] proposed an extension to the Complex Wavelet Structural Similarity Index (CW-SSIM [Wang and Simoncelli, 2005; Sampat *et al.*, 2009]) for images to account for motion in video sequences. The technique (called V-SSIM) incorporates motion modeling using *optical flow* and relies on a decomposition through 3D Gabor filter banks in frequency domain. V-SSIM is therefore able to account for motion artifacts due to quantization of motion vectors and motion compensation mismatches. Recently, the authors published the MOVIE index in a follow-up work [Seshadrinathan and Bovik, 2010], which outputs two separate video quality streams for every 16^{th} frame of the assessed video: *spatial* (closely related to the structure term of SSIM) and *temporal* (assessment of the motion quality based on optical flow fields). In Section 7.3 we compare our work with the MOVIE index and Winkler's PDM, along with a frame-by-frame evaluation by image quality metrics HDRVDP [Mantiuk *et al.*, 2005] and the dynamic range independent metric [Aydın *et al.*, 2008a] (henceforth referred as DRIVDP, refer to Chapter 6).

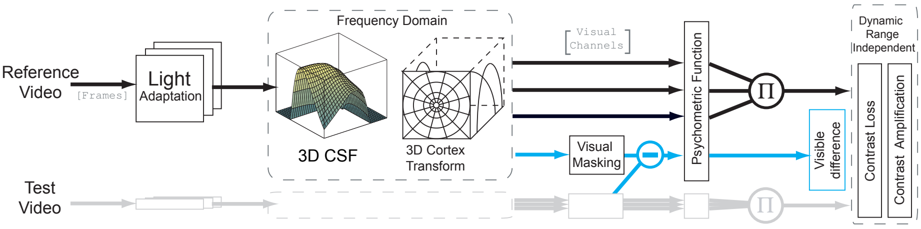


Figure 7.1: *The computational steps of our metric. Refer to text for details.*

7.1.2 Applications in Computer Graphics

The image quality evaluation with the use of HVS models has been an important topic in realistic image synthesis, particularly for static images [Rushmeier *et al.*, 1995; Bolin and Meyer, 1998]. More recently spatiotemporal models of visual perception have been considered for reducing the rendering time of animation sequences by exploiting limitations of the HVS. Myszkowski et al. [2000] proposed the use of an Animation Quality Metric (AQM), which utilizes image flow between a pair of subsequent frames to derive the retinal velocity, which is an input parameter for the spatiovelocity contrast sensitivity function (SVCSF) [Daly, 1998]. Yee et al. [2001] further extended this work by using a computational model of visual attention to predict which image regions are more likely to be consciously attended by the observer, resulting in even more precise retinal velocity estimation. Both those techniques lack explicit processing of intensities between subsequent images, which makes detection of temporal artifacts such as flickering impossible. Such temporal information has been implicitly accumulated by averaging photon density across frame sequences and then applying the AQM metric to the resulting animation frames [Myszkowski *et al.*, 2001]. However, in this case only temporal noise due to the photon density can be estimated, while other temporal artifacts such as flickering of improperly sampled textures or edge aliasing cannot be detected.

Schwarz and Stamminger [2009] propose a quality metric, which is targeted specifically for detection of popping artifacts due to level-of-detail (LOD) changes between frames. They assume the knowledge of the point in time when the LOD is changed and compare whether for that frame the differences for current and previous LOD (the latter image must be specifically re-rendered) are visible taking into account the SVCSF [Daly, 1998]. Since temporal processing over frames is ignored, the influence of the dynamically changing scene and camera on the LOD change cannot be modeled properly. Clearly, an explicit 3D space-time contrast sensitivity function (CSF) processing over a number of subsequent frames is required to account for all possible temporal artifacts in a general setup, which is one of the main goals of our work.

7.1.3 Temporal Aspects of Human Visual System

A significant area of interest of vision research is the Lateral Geniculate Nucleus (LGN), which is a portion of the brain inside the thalamus. It is estimated that 90% of monkey retinal ganglion cells send their axons to LGN layers, thus LGN is known as the primary processing center of visual information. In general, retinal ganglion cells can be divided into *midget* (smaller, majority of ganglion cells, sensitive to detail) and *parasol* (larger, faster output signals, sensitive to movement, only ∼10%) cells. LGN, in turn contains *magnocellular* (large cell bodies) and *parvocellular* (small cell bodies) layers. The axons of midget retinal ganglion cells terminate in the parvocellular layers, while the parasol cells terminate in magnocellular layers [Wandell, 1995, p.124]. This structure suggests the existence of separate *parvocellular* and *magnocellular visual streams.*

Experiments have shown that the destruction of the cells in the parvocellular layers of a monkey's LGN resulted in deteriorated performance for a variety of tasks such as pattern detection and color discrimination. Destroying the cells in the magnocellular layers, however, did not affect the performance in the same tasks, but it was observed that the animal became less sensitive to rapidly flickering targets [Wandell, 1995, p.126]. This leads to the conclusion that the magnocellular pathway is specialized to process high temporal frequency information [Watson, 1986]. Meanwhile, some work has been done to find models that fit psychophysical measurements of the temporal sensitivity of human subjects. While models with many narrow band mechanisms, as well as three channels have been proposed in the past, it is now believed that there is just one low-pass, and one band-pass mechanism [Winkler, 2005]. This theory is consistent with the biological structure of the LGN, moreover Friedericksen and Hess [1998] obtained a very good fit to large psychophysical data using only a *transient* and a *sustatined* mechanism.

Although the parvo– and magnocellular pathways carry different types of information to the brain, the receptive fields of neurons in the parvocellular pathway are not space-time separable [Wandell, 1995, p.143]. No clear anatomical separation between spatial and temporal frequencies supports the psychophysical finding that the contrast sensitivity is not separable along time and spatial dimensions. That leads to the **space-time nonseparability of the Contrast Sensitivity Function**. Thus, spatial CSFs measured for static stimuli cannot be extended linearly to account for the effect of temporal frequency to sensitivity. Another direct consequence of separate pathways for high and low temporal frequency contrast is the **spatiotemporal locality of inter-channel visual masking**. This suggests the use of 3D filter banks that span both spatial and temporal dimensions. Faithful modeling of temporal aspects of the HVS is vital in Computer Graphics applications, where flickering is an important source of visual artifacts. In Section 7.2 we describe how the proposed metric addresses these issues.

7.2 Video Quality Assessment

The recent proliferation of High Dynamic Range Imaging dictates that the HVS model employed in a video quality metric for Computer Graphics applications should be designed for all visible luminance levels. This requirement limits the use of earlier video quality metrics designed towards detecting compression artifacts in low dynamic range (LDR) videos. Moreover, applications such as tone mapping and compression of HDR video sequences require detecting structural distortions where the reference video is HDR and the test video is LDR. Consequently, in this work we use an HDR capable model that accounts for both major spatial and temporal aspects of the visual system, and employ the dynamic range independent distortion measures *contrast loss* and *amplification* introduced in DRIVDP in addition to simply computing the *visible differences* between reference and test videos. In Computer Graphics applications the main concern is often the existence of visible artifacts, rather than the magnitude of visibility, since methods that produce clearly visible artifacts are often not useful in practice. Consequently the HVS model we use trades off supra-threshold precision for accuracy near the detection threshold.

The computational steps of our metric are summarized in Figure 7.1. The input is a pair of videos V_{ref} and V_{tst} with arbitrary dynamic ranges, both of which should contain calibrated luminance values. The luma values of LDR videos should be inverse gamma corrected and converted to display luminance (In all examples we assumed a display device with the luminance range $0.1-100\ cd/m^2$ and gamma 2.2). The HVS model is then applied separately to both videos to obtain the normalized multichannel local contrast at each visual channel, where the first step is to model the nonlinear response of the photoreceptors to luminance, namely **Luminance adaptation** (Chapter 2.2). In our metric we apply the nonlinearity described in Section 10.1, which maps the video luminance to linear Just Noticeable Differences (JND) values, such that the addition or subtraction of the unit value results in a just perceivable change of relative contrast.

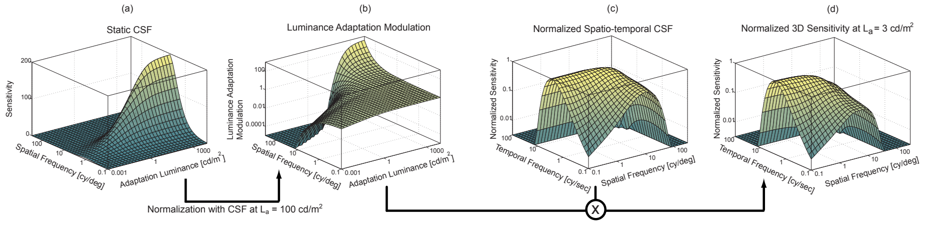


Figure 7.2: *Computation of the CSF^{3D}. The static $CSF^S(\rho, L_a)$ (a) is divided to $CSF^S(\rho, L_a = 100cd/m^2)$ to obtain scaling coefficients (b) that account for luminance adaptation in CSF^{3D}. The specific adaptation level is chosen to reflect the conditions where the spatiotemporal CSF^T was measured (c). The scaling coefficients are computed for the current L_a (3 cd/m^2 in this case), and multiplied with the normalized CSF^T to obtain the CSF^{3D} that accounts for spatial and temporal frequencies, as well luminance adaptation (d).*

Contrast sensitivity (Chapter 2.3) is a function of spatial frequency ρ and temporal frequency ω of a contrast patch, as well as the current adaptation luminance of the observer L_a. The spatiotemporal CSF^T plotted in Figure 7.2c shows the human contrast sensitivity for variations of ρ and ω at a fixed adaptation luminance (Equation 10.5). At a retinal velocity v of 0.15 deg/sec, the CSF^T is close to the static CSF^S [Daly, 1993] (Figure 7.2a) at the same adaptation level (the relation between spatio-temporal frequency and retinal velocity is $\omega = v\rho$ assuming the retina is stable). The formula for CSF^S is given in Equation 10.3. This particular retinal velocity corresponds to the lower limit of natural drift movements of the eye which are present even if the eye is intentionally fixating in a single position [Daly, 1998]. In the absence of eye tracking data we assume that the observer's gaze is fixed, but also the drift movement is present. Accordingly, a minimum retinal velocity is set as follows:

$$CSF^T(\rho, \omega) = CSF^T(\rho,\ max(v, 0.15) \cdot \rho). \tag{7.1}$$

On the other hand, the shape of the CSF depends strongly on adaptation luminance especially for scotopic and mesopic vision, and remains approximately constant over 1000 cd/m^2. Consequently, using a spatiotemporal CSF at a fixed adaptation luminance results in erroneous predictions of sensitivity at the lower luminance levels that can be encoded in HDR images. Thus, we derive a "3D" CSF (Figure 7.2d) by first computing a *Luminance Modulation Factor* (Figure 7.2b) as the ratio of CSF^S at the observer's current adaptation luminance (L_a) with the CSF^S at $L_a = 100\ cd/m^2$, which is the adaptation level at which the CSF^T is calibrated to the spatiotemporal sensitivity of the HVS. This factor is then multiplied with the normalized spatiotemporal CSF ($nCSF^T$), and finally the resulting CSF^{3D} accounts for ρ, ω and L_a:

$$CSF^{3D}(\rho, \omega, L_a) = \frac{CSF^S(\rho, L_a)}{CSF^S(\rho, 100)} nCSF^T(\rho, \omega). \tag{7.2}$$

Ideally the CSF^{3D} should be derived from psychophysical measurements in all three dimensions, since current findings suggest that the actual contrast sensitivity of the HVS is linearly separable in neither of its dimensions. In the absence of such measurements, we found that estimating luminance adaptation using a scaling factor is better than the alternatives that involve an approximation by linear separation of spatial and temporal frequencies (as discussed earlier in Section 7.1.3). The effect of luminance adaptation to spatiotemporal contrast sensitivity is approximately linear except for very low temporal frequencies [Wandell, 1995, p.233].

The perceptually scaled luminance contrast is then decomposed into *visual channels*, each sensitive to different temporal and spatial frequencies and orientations (Chapter 2.4). For this purpose we extend the **Cortex Transform** [Watson, 1987] (Section 10.4 that comprises 6 spatial frequency channels each further divided into 6 orientations (except the base band), by adding a sustained (low temporal frequency) and a transient (high temporal frequency) channel in the temporal dimension (total 62 channels). The time (t given in seconds) dependent impulse responses of the sustained and transient channels, plotted in Figure 7.3-left, are given as Equation 7.3 and its second derivative, respectively [Winkler, 2005]:

$$f(t) = e^{-\frac{\ln(t/0.160)}{0.2}}. \tag{7.3}$$

The corresponding frequency domain filters are computed by applying the Fourier transform to both impulse responses and are shown in Figure 7.3-right.

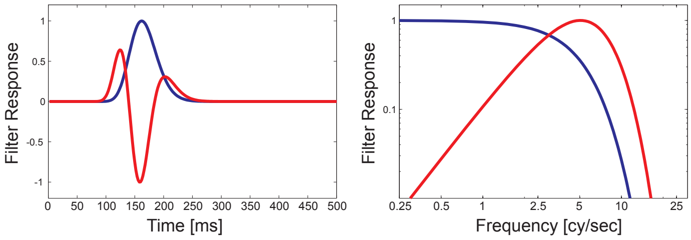


Figure 7.3: *Impulse (left) and frequency (right) responses of the transient (red) and sustained (blue) temporal channels. The frequency responses comprise the extended 3D Cortex Transform's channels in temporal dimension.*

Combining all models discussed so far, the computation of visual channels from the calibrated input video V is performed as follows:

$$\begin{aligned} C^{k,l,m} &= \mathscr{F}^{-1}\left\{V_{csf}\ cortex^{k,l} \times temporal^{m}\right\} \text{ and} \\ V_{csf} &= \mathscr{F}\{jnd(V)\}\ CSF^{3D}, \end{aligned}$$

where the 3D Cortex Filter for channel $C^{k,l,m}$ is computed from the corresponding 2D cortex filter $cortex^{k,l}$ at spatial frequency level k and orientation l, and the sustained and transient channel filters $temporal^{m}$. The function jnd denotes the light adaptation nonlinearity, and $\mathscr{F}$ is the Fourier Transform. The threshold elevation due to **visual masking** (Chapter 2.5) is computed using the following nonlinearity [Daly, 1993]:

$$Te^{k,l,m} = \left[1 + \left(0.0153\left(392.498|C_{pu}^{k,l,m}|\right)^{slope}\right)^{4}\right]^{\frac{1}{4}}, \tag{7.4}$$

where $C_{pu}^{k,l,m}$ indicates the channel with *phase uncertainty* and the *slope* is linearly interpolated between $0.7 - 1$ for visual channels from low to high spatial frequencies.

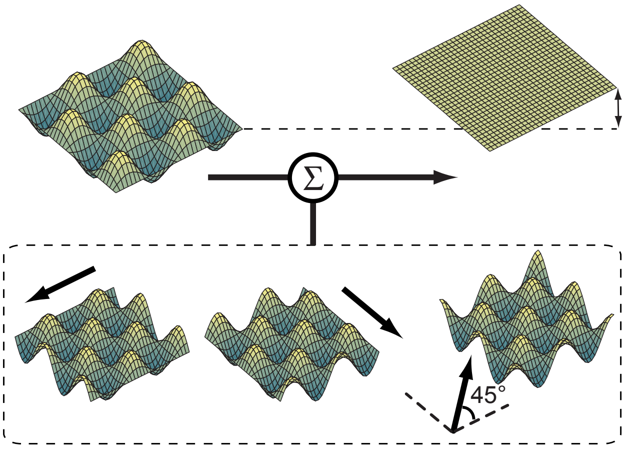


Figure 7.4: *Practical illustration of achieving phase uncertainty in 2D. The Hilbert transform should be applied in multiple orientations to obtain a phase independent signal.*

The dependency of the visual channels to signal phase contradicts with the observation that the phase sensitivity of the HVS is very limited. A common way of removing phase dependency of a 1D signal is to use a *quadrature pair* of filters where one filter is obtained by shifting the other's phase by 90 degrees. Although the phase shift can be computed in 1D by means of Hilbert transform, the extension of the Hilbert transform to higher dimensions is not trivial (Figure 7.4). Our implementation of phase uncertainty is an extension of the quadrature cortex filters [Lukin, 2009] to the temporal domain. The spatial phase-shift is computed using an oriented 2D Hilbert Transform:

$$h^{S}(\rho_x, \rho_y) = i\ sgn(p\ \rho_x + q\ \rho_y), \tag{7.5}$$

where i is the imaginary unit, and the line given by the equation $p\ \rho_x + q\ \rho_y = 0$ specifies the "direction" of the transform. Parameters p and q are selected such that the direction of the Hilbert Transform coincides with the spatial orientation of the cortex channel. In the temporal dimension the phase shift can be achieved using a 1D Hilbert Transform:

$$h^T(\omega) = i\ sgn(\omega). \tag{7.6}$$

The quadrature responses of spatiotemporal visual channels are then computed as follows:

$$H^{S|T}\{C^{k,l,m}\} = \mathscr{F}^{-1}\{h^{S|T}\ \mathscr{F}\{C^{k,l,m}\}\}. \tag{7.7}$$

The phase independent channel $C_{pu}^{k,l,m}$ used in the threshold elevation formula is computed by summing up the original signal with all phase shifted responses in spatial and temporal dimensions as illustrated in Figure 7.5.

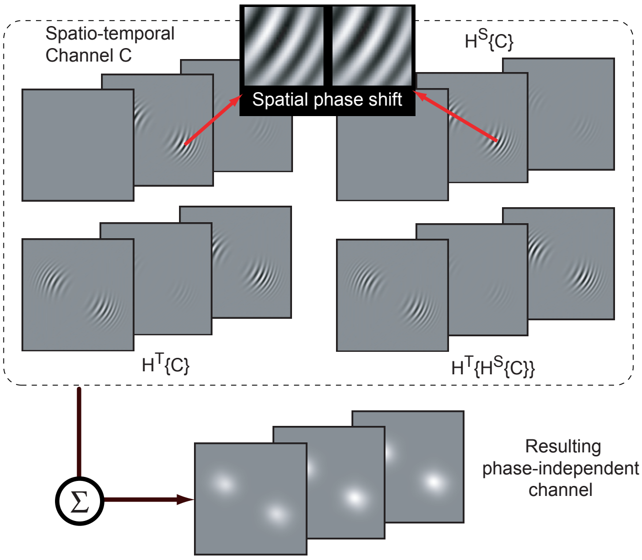


Figure 7.5: *3D phase uncertainty on a frequency plate image modulated in temporal domain using a sinusoid function. The spatiotemporal channel C obtained by 3D Cortex Transform is used to compute $H^S\{C\}$, $H^T\{C\}$ and $H^T\{H^S\{C\}\}$, the phase shifted response in spatial, temporal and both dimensions, respectively. The combination of all four responses yields a spatiotemporaly phase independent response constant along the entire sequence.*

The detection probability of the normalized contrast response C at each visual channel is computed using the following **psychometric function** (Chapter 2.7), separately for the reference and test images:

$$P(C) = 1 - \exp(-|C|^3). \tag{7.8}$$

The psychometric function relates the normalized contrast to detection probability. Using this function, we compute the detection probabilities of the following three types of distortions:

- **Visible Difference** $\left(P_{\Delta}^{k,l,m} = P(\frac{C_{tst}^{k,l,m}}{Te_{tst}^{k,l,m}} - \frac{C_{ref}^{k,l,m}}{Te_{ref}^{k,l,m}})\right)$
- **Contrast Loss** $\left(P_{\searrow}^{k,l,m} = P(C_{ref}^{k,l,m})(1 - P(C_{tst}^{k,l,m})\right)$
- **Contrast Amplification** $\left(P_{\nearrow}^{k,l,m} = P(C_{tst}^{k,l,m})(1 - P(C_{ref}^{k,l,m})\right)$

The visible differences between video sequences convey more information than the other two types of distortions, but especially if the input video pair has different dynamic ranges, the probability map is quickly saturated by the contrast difference that is not necessarily perceived as a distortion. In this case contrast loss and amplification are useful which predict the probability of a detail visible in the reference becoming invisible in the test video, and vice versa. While additionally contrast reversal proposed in DRIVDP can be easily computed within this framework, we found that this type of distortion did not convey further information in the examples we considered, and thus excluded from the metric output. Detection probabilities of each type of distortions are then combined using a standard probability summation function:

$$\hat{P}_{\Delta|\searrow|\nearrow} = 1 - \prod_{k=1}^{K} \prod_{l=1}^{L} \prod_{m=1}^{M} \left(1 - P_{\Delta|\searrow|\nearrow}^{k,l,m}\right). \tag{7.9}$$

The resulting three *distortion maps* $\hat{P}$ are visualized separately using an in-context distortion map approach where detection probabilities are shown in color over a low contrast grayscale version of the test video. We also found that an overall summary of the distortion information conveyed through a 3D visualization is useful in certain applications (Section 7.4.4).

7.3 Results

In this section we compare the predictions of our metric with the outcomes of the recent video quality metrics PDM [Winkler, 2005] and the MOVIE index [Seshadrinathan and Bovik, 2010]. Although not intended for videos, we also considered two recent HDR capable image quality metrics HDRVDP [Mantiuk *et al.*, 2005] and DRIVDP [Aydın *et al.*, 2008a] (discussed in Chapter 6), with which we evaluated each video frame separately. To ensure that our metric is calibrated to psychophysically measured detection thresholds, we computed the visible differences of the Modelfest data set at five different contrast levels with the background luminance. The video for a stimulus is generated by repeating it in all frames. As expected, the majority of the stimuli produced no response below the threshold, and a response with increasing magnitude for near– and above threshold. Figure 7.6 shows the outcome for selected stimuli relevant to our applications: a low and a high frequency noise, and a complex image. The worst results were obtained for “GaborPatch9” and “Gaussian26” for which our metric was too insensitive

The test video for this section is generated using an HDR image, to which we added spatiotemporal random noise filtered with a Gaussian to roughly mimic

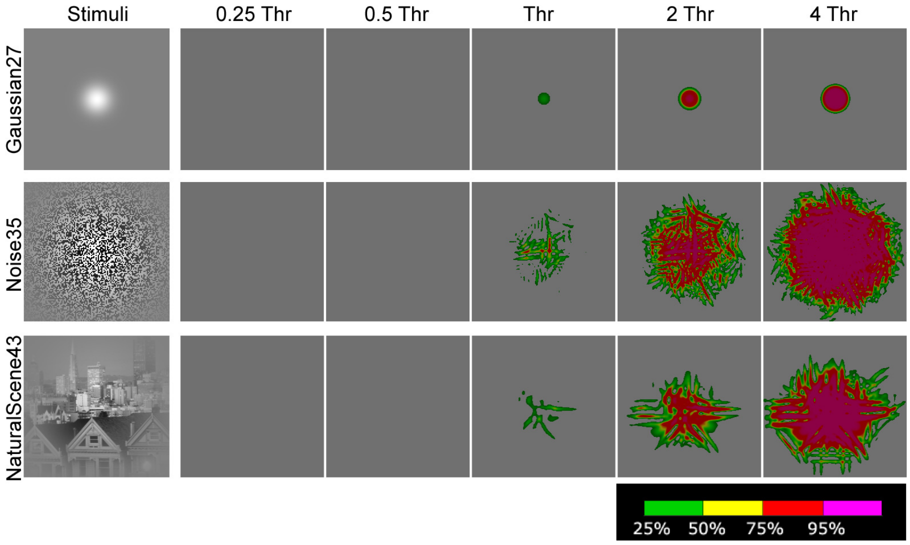


Figure 7.6: *Predicted visible differences between selected stimuli from the Modelfest data set and the background luminance, where the stimuli is scaled at* $\frac{1}{4}, \frac{1}{2}, 1, 2$ *and* 4 *times the threshold contrast (The same color coding is used throughout this chapter for visualizing distortion detection probabilities, unless noted otherwise).*

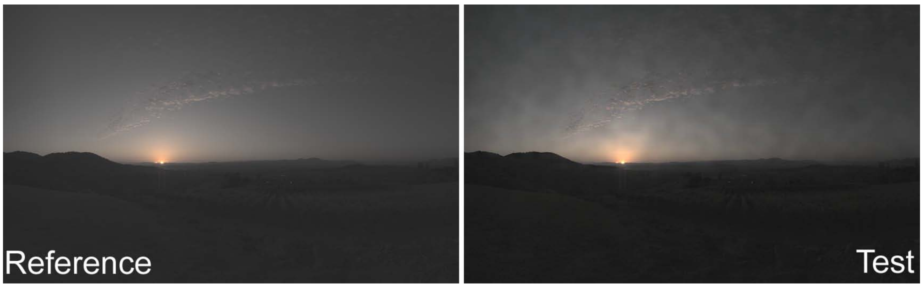


Figure 7.7: *Approximate perception of the reference and test scenes*

the artifacts that appear in rendered videos in the absence of temporal coherency. The magnitude of the noise has been modulated with the luminance levels of the relatively dark image that depicts a sunset. The reference video is generated similarly by repeating the same HDR image in all frames. The frames in Figure 7.7, tone mapped using Pattanaik's operator [2000], depict the approximate appearance of the scene.

First, we compare the distortion visibility prediction of our metric with PDM and MOVIE index on this tone mapped LDR image pair. Due to the random nature of the distortion, the frames of the distortion maps in this section are very similar, and thus we arbitrarily choose a single representative frame. In this case the outcome of our metric and the PDM are similar (Figure 7.8).

The output of the MOVIE index on the other hand are a series of spatial and a

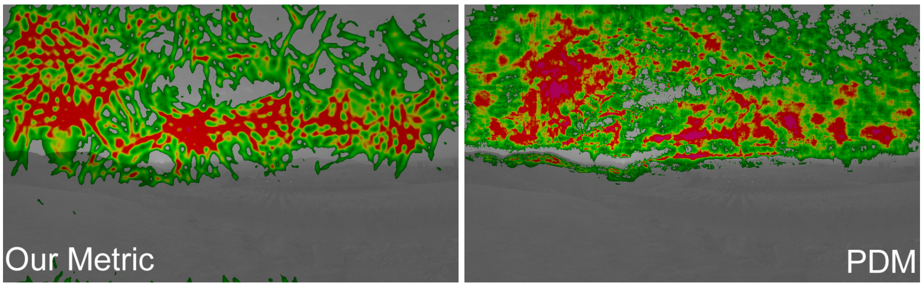


Figure 7.8: *Metric comparison for LDR test and reference videos*

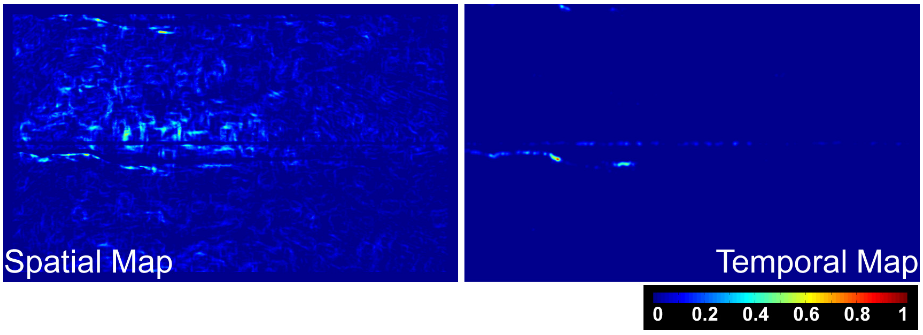


Figure 7.9: *MOVIE index for LDR videos. Note the different color coding*

temporal distortion maps that are computed at every 16^{th} frame. In Figure 7.9 we show the spatial distortion map at the 3^{rd} scale along with the temporal distortion map. While the output format of the MOVIE index is not directly comparable with other metrics discussed in this section, one can see that the spatial map of structural distortions (Figure 7.9-left) closely correlates to the distortions in the video sequence. However, due to the lack of a mechanism to estimate threshold contrast, distortions are detected even at the darker bottom half of the video.

Next, we test the metrics on the HDR test and reference videos. Note that the HDR format is capable of encoding the actual scene luminance unlike display-referred LDR videos in the previous case. The MOVIE index is excluded from the remaining comparisons since its extention to HDR is not trivial. The difference in predictions of our metric and PDM in this case is because the latter does not model luminance adaptation. Consequently distortion visibility is underestimated due to artificially high thresholds in this low luminance scene (Figure 7.10). The visible difference and contrast amplification predicted by frame-by-frame evaluation of HDRVDP and DRIVDP are also noticeably lower than ours due to the absence of a temporal model that accounts for the higher sensitivity to flickering distortions compared to static distortions.

An even more striking difference can be observed in the final setup where the distorted video tone mapped with Pattanaik's operator is compared with the reference HDR video (Figure 7.11). Here, both PDM and and HDRVDP's distortion maps are dominated by the contrast difference due to the different dynamic

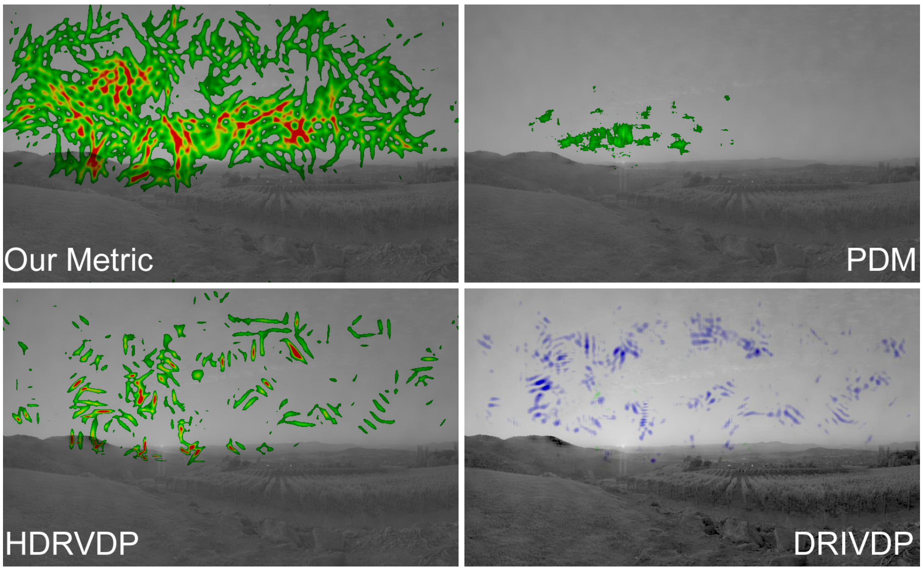


Figure 7.10: *Metric comparison for HDR test and reference videos. The contrast amplification in DRIVDP is color coded with blue.*

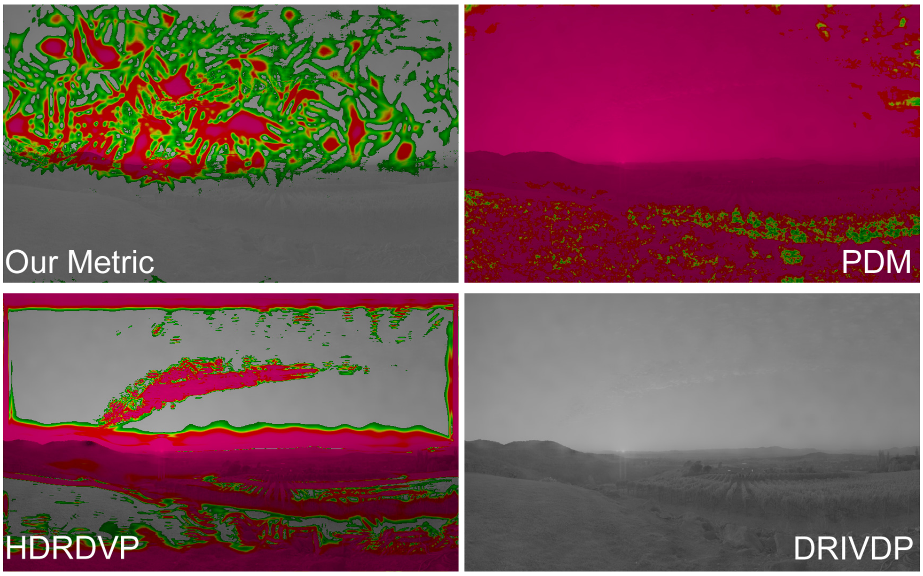


Figure 7.11: *Metric comparison for HDR reference and LDR test videos*

ranges of the input video pair. This is especially evident in HDRVDP's prediction where the spatiotemporal distortion appears to be completely ignored. Moreover, DRIVDP predicts no visible detail amplification at all, since it does not detect the distortion and is also not affected by the different dynamic ranges of the input videos. The contrast amplification predicted by our metric on the other hand correctly identifies distortions where they are visible, and similar to

DRIVDP also ignores the changes due to dynamic range difference. Note also that the predictions of our metric in all three scenarios are fairly consistent.

7.4 Applications

The proposed method for objective quality assessment of a test video with respect to a reference without any constraints on the dynamic range provides a faster alternative to subjective evaluation of rendering methods, and also enables a computational comparison of HDR video compression and tone mapping techniques. We also show that our metric gives insight on the effect of dropped frames to overall quality.

7.4.1 HDR Video Compression

While HDR content is becoming more commonplace, since it offers higher fidelity compared to traditional media, it does so at the cost of significantly increased file sizes. This is often not a problem for images due to cheaply available storage. However, working with long, high resolution videos quickly becomes prohibitively expensive. Incidentally HDR video compression has become an active topic of research. Figure 7.12 shows that our metric can be used to detect compression artifacts in a video sequence compressed [Mantiuk *et al.*, 2004] at various quality settings.

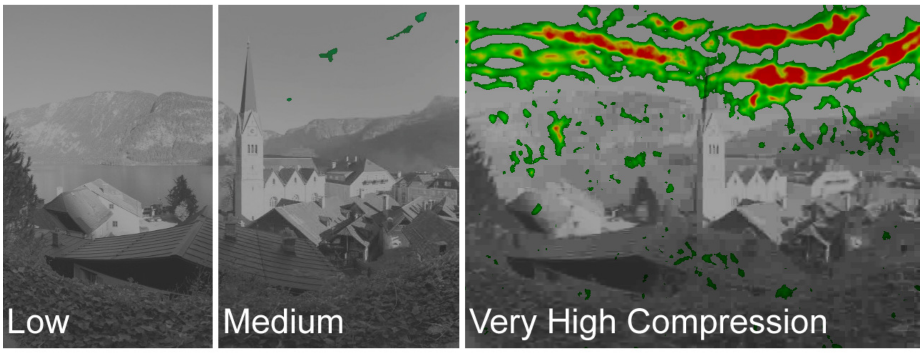


Figure 7.12: *Visible differences between frames from the HDR video and the corresponding compressed frames shown in three compression settings (Low – q=1, Medium – q=5, Very High – q=31). The banding artifacts become clearly visible under extreme compression. Near the foliage at the bottom, banding artifacts are present but not visible due to the low luminance*

7.4.2 Temporal Tone Mapping

HDR display technology is still early in its development, thus it is often necessary to reduce the dynamic range of the HDR content such that it can be viewed on current display hardware. While the goal of tone mapping is considered to be subjective, the fidelity of the tone mapped video to the reference HDR is often a good indicator of quality. In Figure 7.13 we show the results from selected frames of a tone mapped HDR sequence computed with global [Drago *et al.*, 2003b] and gradient based [Fattal *et al.*, 2002] tone mapping methods.

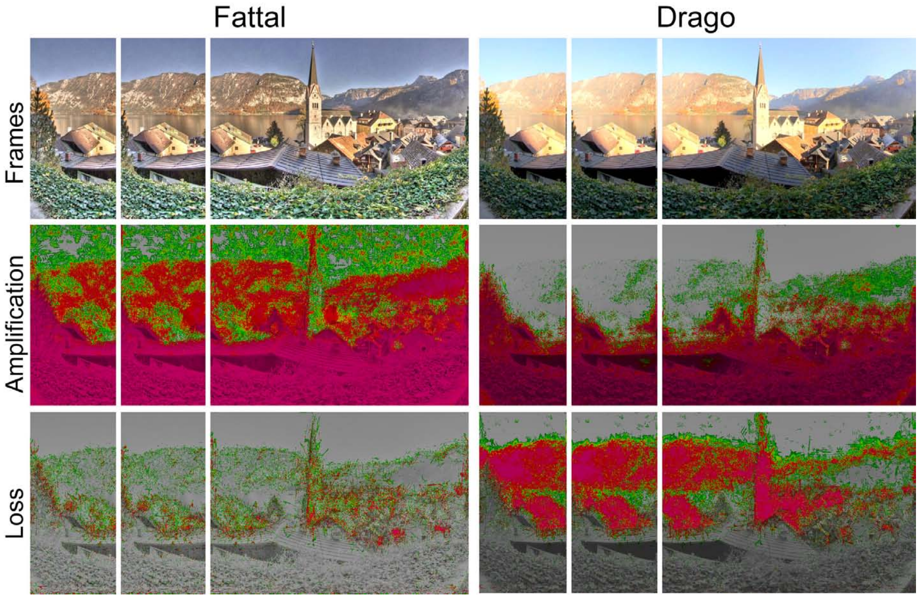


Figure 7.13: *Selected frames from the tone mapped HDR sequences and corresponding contrast amplification and loss maps. Each frame of the reference HDR video is tone mapped separately. Fattal's gradient based operator enhances perceived contrast notably, thus leading to highly detectable contrast amplification but little contrast loss. Drago's global operator on the other hand produces a more "flat" image by amplifying contrast near the dark foliage in the foreground and clipping brighter details near the horizon line.*

Another interesting practical problem involves both temporal tone mapping and compression. Consider a scenario where visual content is stored in a centralized media server in compressed HDR format. One may require to perform on-the-fly tone mapping to reduce the video's dynamic range to be suitable for the client machine's display device, which may range from an high-end LCD panel to a limited CRT monitor. An obvious consideration in this case is to make sure that tone mapping does not amplify previously invisible compression artifacts. In Figure 7.14 we show such an example where tone mapping adversely affects perceived quality of the compressed HDR video, which is correctly detected by our metric.

7.4.3 Rendering

Our metric can be used to compare different rendering approaches. Figure 7.15 shows the visible differences of a dynamic scene walkthrough rendered with indirect lighting using reflective shadow maps [Dachsbacher and Stamminger,

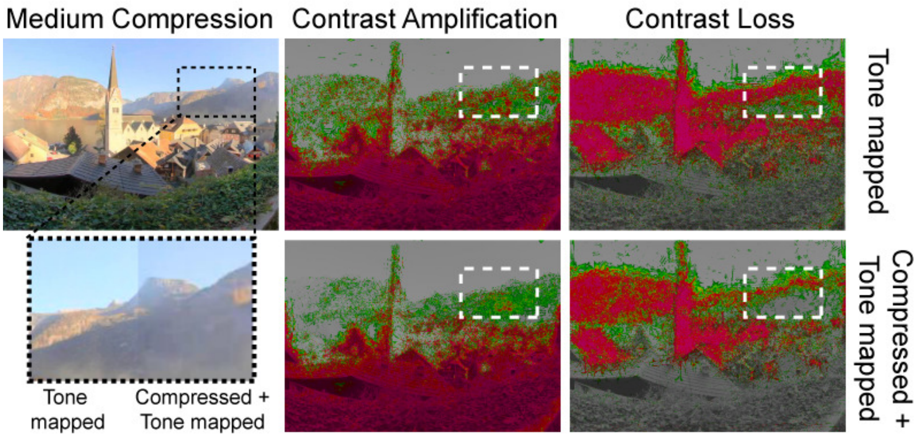


Figure 7.14: *Contrast amplification and loss predicted with respect to the reference HDR sequence for the compressed (at medium quality) and then tone mapped sequence using Drago's operator. Note the slightly increased contrast amplification and loss in the tone mapped version of the compressed HDR video. As shown in Figure 7.12, the artifacts generated in medium compression setting for this scene are mostly not detectable in the HDR video, but they become visible due to tone mapping applied later.*

2005] with 1000 virtual point light (VPL) sources, with respect to the reference sequence obtained with the same amount of VPLs, however using a recent technique [Herzog *et al.*, 2010] that utilizes spatio-temporal filtering. Due to this filtering, there are virtually no visible artifacts in the reference sequence, while the test technique produces visible flickering during the entire sequence.

To complement the previous scene with mostly temporal distortions, we show another example with artifacts of spatiotemporal nature (Figure 7.16). Here, the squences are rendered using an image-space horizon based ambient occlusion technique [Bavoil *et al.*, 2008] augmented with the screen space directional occlusion (SSDO) [Ritschel *et al.*, 2009a] (48 × 32 and 12 × 10 polar samples on the hemisphere for the reference and test sequences, respectively) with directional light source sampled from an environment map (128 and 96 samples, respectively) and percentage closer filtering (PCF) shadow maps [Reeves *et al.*, 1987] (64 and 16 samples, respectively). Visible differences are predicted mostly near the boundaries of the elephant's shadow.

7.4.4 Variable Frame Rate

Maintaining a high enough frame rate is desirable in applications like rendering and video streaming, but at the same time is not always possible due to hardware or bandwidth limitations. In this case, the visible differences between the low FPS video and the full FPS reference is a good measure for the loss in perceived quality due to low frame rate. Figure 7.17 shows that our metric can be used to predict the perceived distortions caused by dropped frames in a rendered walkthrough scene. The reference sequence was generated by Coherent

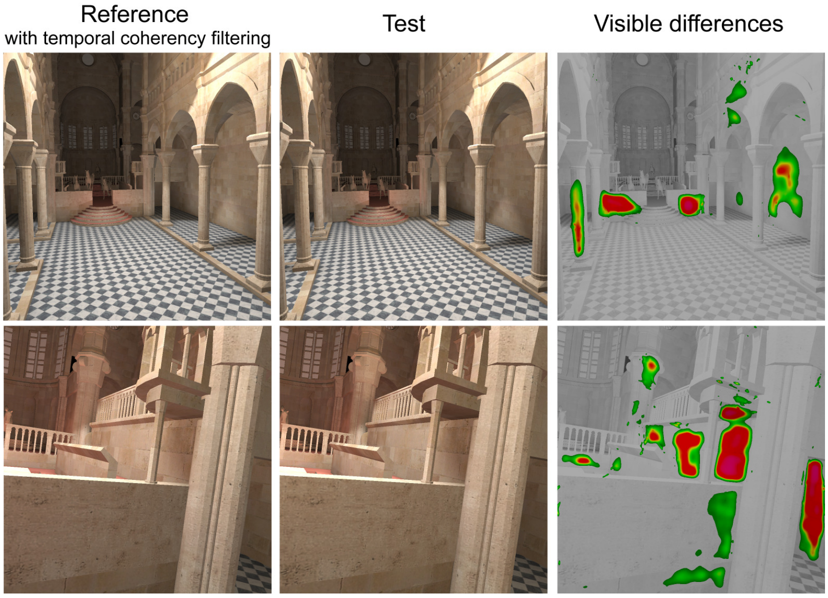


Figure 7.15: *Visible differences between rendering techniques. Even though the rendered frames are visually indistinguishable when viewed side-by-side, the test method produces significantly visible flickering artifacts, which is not the case for the reference method with temporal coherency filtering. Our metric also detects the non-uniform perception of these flickering artifacts, such as the perception of the artifacts on the ground masked by the moving checkerboard pattern*

Hierarchical Culling technique [Bittner *et al.*, 2004] which never falls below 60 FPS for this scene. On the other hand, the performance of the traditional view frustum culling drops below 1 FPS at times. We also show an alternative 3D visualization of this scene utilizing volume rendering that gives an overview of the distortion data (Figure 7.18).

7.5 Discussion

The running time of the proposed metric depends highly on the resolution and length of the input videos, however in its current state is intended to work offline ($\sim$ 5 minutes for $512 \times 512 \times 64$ sequence). In our experience, the main bottleneck in performance is computing the 3D Fourier Transform of an 64 frames portion of the video, where that specific number is chosen because the sensitivity to temporal frequencies higher than 32 cy/sec is significantly low. This approach also requires that the portions of the video being processed should be kept in memory.

While our implementation runs in a standard workstation hardware without problems, another approach that trades off efficiency for prediction accuracy is to approximate the frequency domain Cortex Transform with the Steerable Pyramid decomposition performed in the spatial domain through polynomial

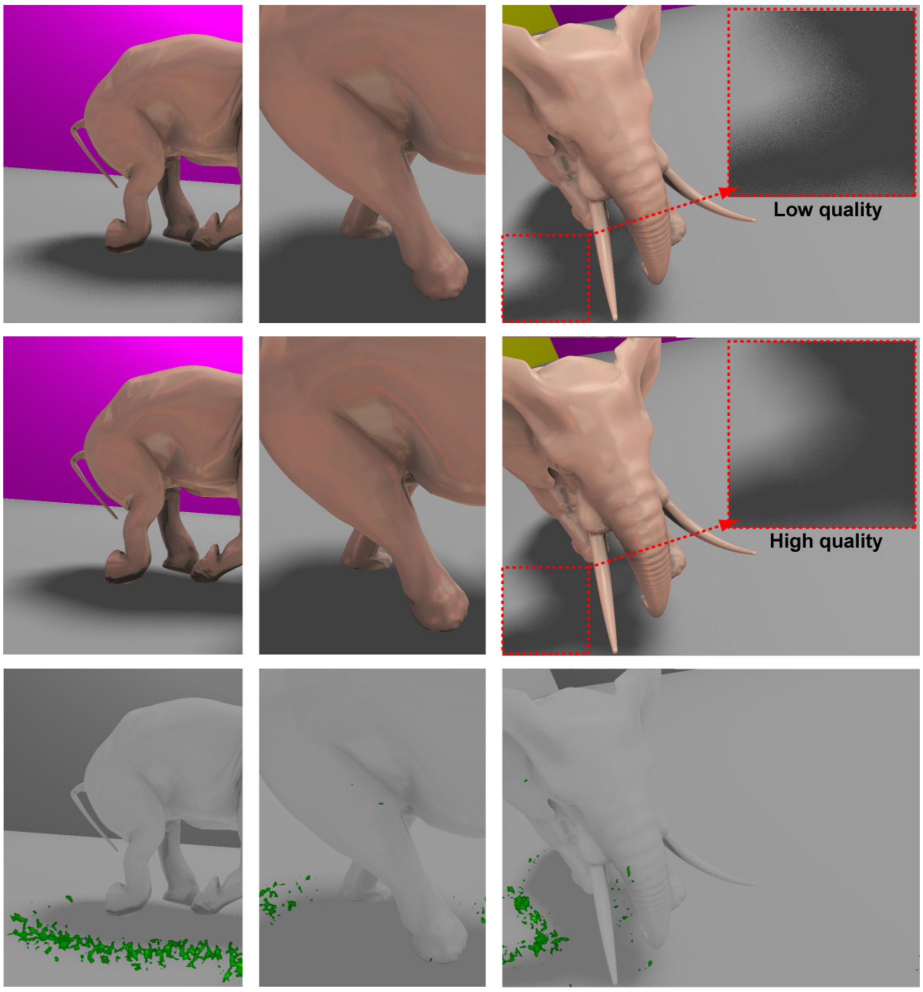


Figure 7.16: *Visible differences (bottom row) between the high (top row) and low quality (middle row) renderings are focused mostly near shadow boundaries.*

approximations of the second derivative Gaussian filters [Freeman and Adelson, 1991]. The filters that compute transient and sustained temporal channels can also be approximated by 9-tap filters corresponding to the impulse responses given in Figure 7.3 as described in Winkler's book [2005]. As a result, the memory requirement can be reduced by a factor of nearly 7, and the overall computation can be accelerated by efficiently computing convolution operations in graphics hardware. The downside is the metric's reduced prediction performance since second derivative Gaussian filters are not perceptually justified and our pilot implementation also indicated difficulties in calibration.

A limitation of our metric is the lack of a mechanism to model visual attention. In the absence of either a computational model, or eye tracking data to predict the observer's gaze direction, our metric's predictions are conservative in the sense that the possibility of the observer focusing her attention to some other

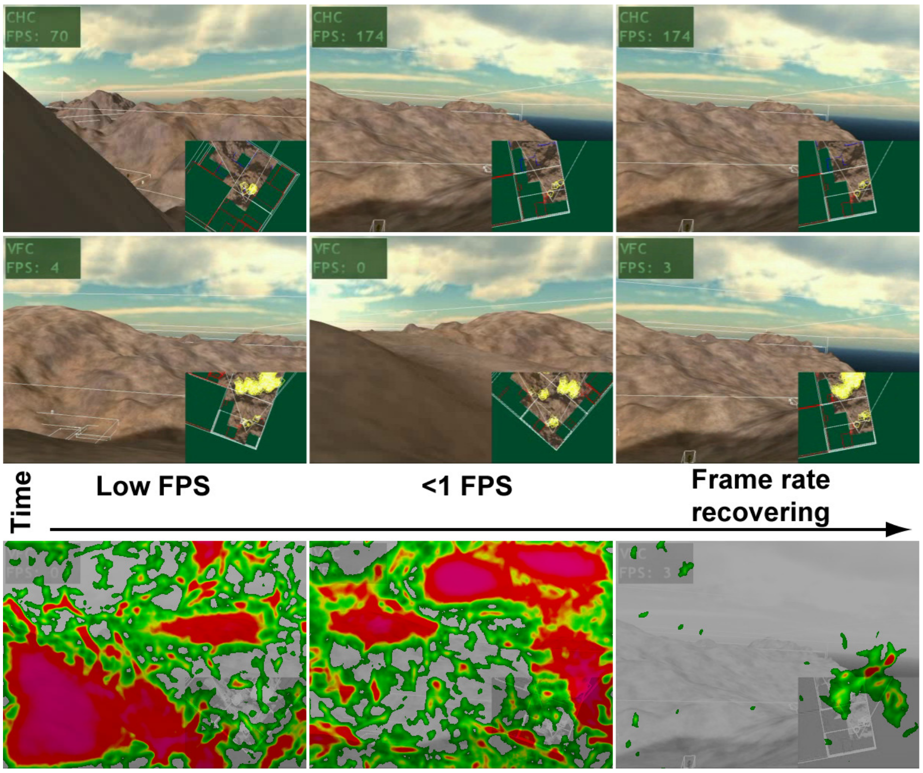


Figure 7.17: *The effect of dropped frames to perceived quality. One should note, however, that our method does not compensate for camera movements and assumes frames are perfectly aligned with each other.*

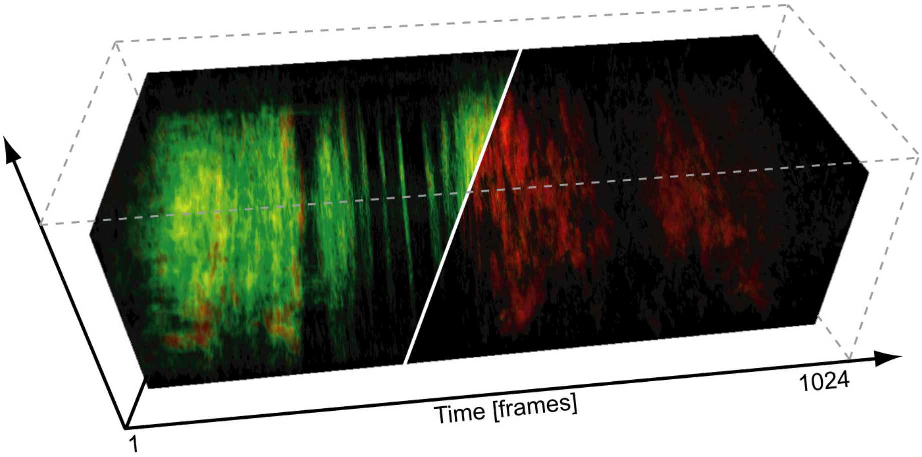


Figure 7.18: *An alternative 3D visualization. The left slice shows a volume rendering of the entire visible differences data. The right slice shows only the differences with detection probability above 75% where the locations of the missing frames along the time axis are better visible.*

region than where the sought artifact appears is not considered. Another limitation of our metric is the requirement of a reference video for quality evaluation, which may not be available in some applications. No reference metrics, however, have limited utility since they are often geared toward detecting a single type of distortion, and are generally not as accurate as full reference metrics.

7.6 Conclusion

We presented a video quality metric specifically designed for Computer Graphics applications. Our method comprises an HVS model built with spatiotemporal components that are designed for HDR luminance levels. The capability of comparing video pairs with different dynamic ranges enables applications such as objective evaluation of HDR video compression and tone mapping, as well as comparison of different rendering methods and predicting the effect of dropped frames to perceived quality.

The validation of video quality metrics is often performed by comparing the metric responses to standard image quality databases. In the absence of such a collection of video pairs and corresponding spatial distortion maps comprising stimuli with different dynamic ranges and multitude of artifact types relevant to Computer Graphics, we created a modest data set for validation purposes. A future direction is to extend our initial effort to a standardized data set. Another possible extention to our work is the inclusion of color channels utilizing a color appearance model designed for HDR luminance levels. Temporal inverse tone mapping evaluation is a natural application area of our metric, but it was not included in this work since from the metric's point of view, the difference between forward and inverse tone mapping is merely swapping reference (HDR) and test (LDR) videos. Nevertheless, the metric's detection performance of application specific banding artifacts deserves further investigation.

Chapter 8

Video Quality Metric Validation

In this chapter we discuss the subjective validation study performed for the video quality metric presented in Chapter 7. The goal of the study was to examine the correlation between the objective quality predictions computed by the proposed video quality metric, and the subjective responses obtained by the experimental procedure described below in Section 8.1. The calibration procedure (described in Section 7.3) and the validation study are complementary, in the sense that the former involves simple stimuli at near threshold visibility to match the sensitivity of the metric to that of an average observer, and the latter involves complex, application oriented stimuli for validating that the individual components of the metric work well in concert.

Two important properties of the proposed metric were influential while designing the validation study: (i) the capability of assessing the quality of HDR videos, as well as comparing HDR videos with LDR videos and vice versa, and (ii) the outcome of the metric in the form of distortion maps that show quality prediction as a function of spatial position which is especially important for applications in computer graphics. To that end the subjective study has the following novelties over previous studies on video quality assessment:

- The test set includes LDR-LDR, HDR-HDR, and HDR-LDR reference-test video pairs with various types of distortions.
- A BrightSide DR37-P HDR display (max. luminance $\approx 3000\ cd/m^2$) was used for displaying the videos.
- The subjects are not asked to assess only an overall quality of the video, but to mark the regions where they see differences between test and reference videos, resulting in distortion maps similar to the metric outcome.

In the remainder of this document we will describe our experimental setup and procedure (Section 8.1), present (Section 8.2) and discuss (Section 8.3) the results based on the correlation between the outcome of the subjective study and corresponding predictions of our metric, PDM, HDRVDP and DRIVDP, and conclude with final remarks and future directions (Section 8.4).

8.1 Experimental Methods

The set of 9 reference-test video pairs (1 LDR-LDR, 2 HDR-LDR, and 6 HDR-HDR) used in the experiment are listed in Table 8.1. The video stimuli were generated by imposing temporally varying visual artifacts to HDR scenes (Figure 8.1), such as HDR video compression artifacts and temporal random noise along with temporal luminance modulation and tone mapping. The magnitudes of the visual artifacts were carefully selected so that there were *sub-*, *near-* and *supra-threshold* distortions present in the experimental videos. The temporal random noise was generated by filtering a three dimensional array of random values between -0.5 and 0.5 by a Gaussian with standard deviations 20 (high) and 5 (low) pixels along each dimension. The magnitude of noise was adjusted by multiplying with two constants separately, such that the artifacts are barely visible in one setting (low), and clearly visible in the other (high). HDR compression [Mantiuk *et al.*, 2004] was similarly applied at two levels to the HDR scenes, where the luminance was globally modulated over time by 0.5% of the maximum scene luminance to vary the visibility of image details over time. Videos generated by applying tone mapping operators [Fattal *et al.*, 2002; Pattanaik *et al.*, 2000] to each input HDR video frame were used in the dynamic range independent comparisons.

#	Source	Ref. DR	Test DR	Artifact Type of Test Video
1	Cars	HDR	HDR	Noise - high magnitude, low stddev
2	Lamp	HDR	HDR	Noise - high magnitude, low stddev
3	Desk	HDR	HDR	Noise - low magnitude, low stddev
4	Tree	HDR	HDR	Noise - high magnitude, high stddev
5	Cafe	HDR	HDR	HDR compression - high quality, luminance mod.
6	Tower	HDR	HDR	HDR compression - low quality, luminance mod.
7	Cafe	HDR	LDR	Luminance modulation, Pattanaik's tone mapping
8	Lamp	HDR	LDR	Luminance modulation, Fattal's tone mapping
9	Lamp	LDR	LDR	Noise

Table 8.1: *List of the experimental stimuli. Refer to text for details.*

All test videos consisted of 60 frames, and were presented at 24 fps. In order to faithfully reproduce the luminance values on the HDR display, the response function of the display was measured using a Minolta LS-100 luminance meter. The measurements consisted of 32 samples taken from the displayable luminance range with equal logarithmic spacing. The sample points were then fitted to a 3^{rd} degree polynomial function, from which 100 points were resampled and stored as a lookup table. Finally, the pixel values for the HDR videos were determined by cubic spline interpolation between nearest two luminance levels. Furthermore, the displayed luminance of the HDR videos were measured again at various regions, and whenever necessary, the scenes were slightly recalibrated to ensure that the displayed luminance values match the actual scene luminance.

The participants of the study were 16 subjects between ages of 23 and 50.

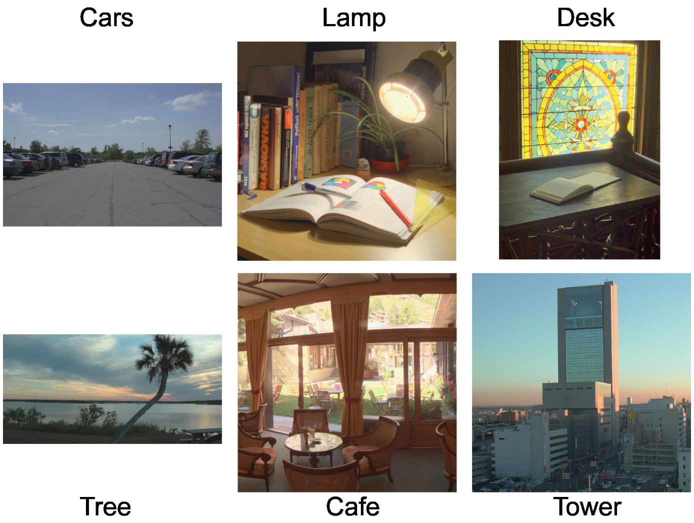


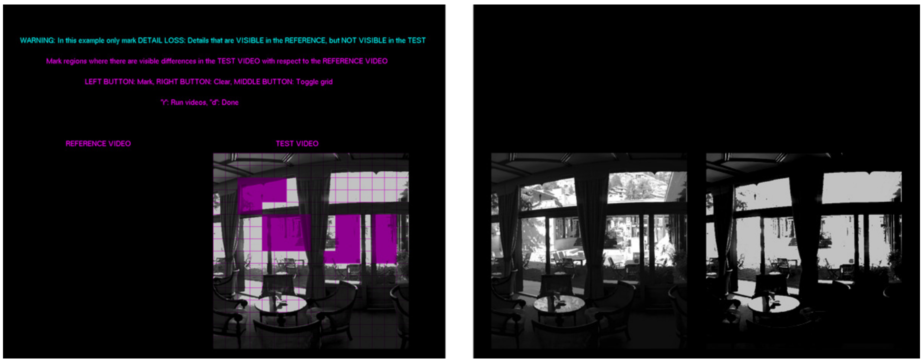
Figure 8.1: *The video test set is generated from 6 calibrated HDR scenes (tone mapped for presentation purpose [Reinhard* et al.*, 2002]). The scene luminance was clipped where it exceeded the maximum display luminance. The displayed luminance of the videos resulting from the scenes were between 0.1 and 3000 cd/m^2.*

They all had near-perfect or corrected to normal vision, and were naïve for the purposes of the experiment. Each subject evaluated the quality of the whole test set through a graphical user interface displayed on a BrightSide DR37-P HDR display (Figure 8.2). In the HDR-HDR, and LDR-LDR comparisons, the task was to mark the regions in the test video where visible differences were present with respect to the reference video. In the HDR-LDR comparisons on the other hand, the subjects were asked to assess the contrast loss and amplification.

In the instruction phase before the experiment, the subjects were asked to mark a grid tile even if visible differences were present only in a portion of that grid's area. They were also encouraged to mark a grid tile in the case they cannot decide whether it contains a visible difference or not. The subjects were placed 0.75 meters away from the display so that a 512 × 512 image spanned 16 visual degrees and the grid cell size was approximately 1 visual degree. The environment illumination was dimmed and controlled, and all subjects were given time to adapt to the room illumination. There were no time limitations set for the experiment, but the majority of the subjects took $15 - 30$ minutes for the entire test set.

8.2 Results

The marked regions for each trial were stored as distortion maps with 16 × 16 resolution, which were then averaged over all subjects to find the mean subjective response. Next, the metric prediction for the corresponding stimulus was computed, averaged over the whole 60 frames, and downsampled to the



WARNING: In this example only mark DETAIL LOSS: Details that are VISIBLE in the REFERENCE, but NOT VISIBLE in the TEST

Mark regions where there are visible differences in the TEST VIDEO with respect to the REFERENCE VIDEO

LEFT BUTTON: Mark, RIGHT BUTTON: Clear, MIDDLE BUTTON: Toggle grid

"r": Run videos, "d": Done

Figure 8.2: *The experiment was performed through a graphical user interface shown on the HDR display. Subjects were shown reference and test videos side by side in a randomized order (right), and were asked to mark the relevant image locations on a 16 × 16 grid according to the instructions (left). The interface and messages were disabled while the videos were being shown. The interface allowed the subjects to watch the videos for an unlimited amount of iterations.*

same resolution as the mean subjective response. For each video pair, we computed the 2D correlation between the mean subjective response and the metric prediction (Table 8.3) and used the results to evaluate the performance of our metric.

The resulting correlations for our metric vary from 0.733 to 0.883. The first two columns of Figure 8.3 show the mean subjective distortion maps along with the corresponding metric predictions for visual inspection. Furthermore, the descriptive statistics of these maps are summarized in Table 8.2. While not optimal, we believe that the presented correlations, along with the fact that the maps obtained by the metric's predictions and the subjective experiment look visually similar, clearly show that our metrics predictions are accurate for practical purposes. Highest correlations were obtained for the #2 HDR-HDR Lamp stimulus with high magnitude, low standard deviation noise, and the #7 HDR-LDR Cafe stimulus with luminance modulation and Pattanaik's tone mapping (0.883 and 0.879, respectively). For these two cases, the magnitude of the probability of detection predicted by the metric, and the average of the binary maps over subjects obtained experimentally are also very similar. In other cases, either the magnitudes of the mean subjective maps were lower than the corresponding detection probability magnitude predictions (such as #4 Tree HDR-HDR stimulus with high magnitude, high standard deviation noise, and #9 Lamp LDR-LDR stimulus with noise), or a certain region with visible distortions was missed out (#1 Cars HDR-HDR stimulus with high magnitude, low standard deviation noise). For the remaining stimuli, a combination of both deviations can be observed in the metric predictions and subjective responses. However, even in the worst case (#8, 0.733), the correlation was at an acceptable level.

Figure 8.4 shows the standard deviation for each stimulus over the test subjects, separately for each grid tile. Over all images, the minimum and maximum values are obtained as 0 and 0.51, the former indicating the tiles on which all subjects gave the same response, and the latter indicating the tiles where approximately half of the subjects have marked.

8.3 Discussion

A problem we experienced during the experiment was the extreme brightness of the sky region of the *Tower* scene, reaching the maximum displayable luminance level ($\approx 3000cd/m^2$). We observed that subjects were disturbed by the displayed luminance level and rushed to the next scene. We also found that the subjects had difficulties understanding the concept of contrast amplification. We believe the reason for that might be that contrast amplification often improves quality, unlike other distortions that were employed in the experiment. As a result, the correlation results in these two cases are slightly worse compared to the others.

We also computed the predictions of PDM [Winkler, 2005], HDRVDP [Mantiuk *et al.*, 2005], and DRIVDP [Aydın *et al.*, 2008a]. The latter two metrics are designed for image quality evaluation, thus, as in Chapter 7, the video stimuli was evaluated for each frame separately. HDRVDP, while capable of evaluating the quality of HDR images, lacks any temporal processing and is geared towards comparing images with the same dynamic range. The DRIVDP addresses the latter limitation, but still suffers from the former. Consequently, DRIVDP's predictions for the HDR-LDR stimuli (numbers 7 and 8) is slightly better than HDRVDP. PDM, on the other hand, is designed for the video stimuli, but lacks the HDR and dynamic range independent mechanisms of HDRVDP and DRIVDP, producing the least average correlation with the subjective responses. As shown in Table 8.3, our metric significantly outperforms others in most cases. The significant difference in average correlations over the entire test set (last row of Table 8.3) shows that overall our metric's predictions are clearly more accurate than others. The corresponding distortion maps predicted by PDM, HDRVDP and DRIVDP are shown in Figure 8.3 columns $3-5$ (averaged and downsampled to 16×16 after the computation).

While the relation between the correlation values and distortion maps is obvious in most cases, the high correlation of PDM for stimulus #3 deserves further explanation. While PDM correctly detects the distorted regions in that stimulus in a spatial sense, the magnitude of detection probabilities are very low (refer to Table 8.2), to the point that they are quantized by the visualization. Thus the map appears to be blank, but since the relation with the subjective data is linear, the correlation is high.

For the purposes of generating the maps in Figure 8.3, in cases of PDM and HDRVDP we simply used the distortion maps produced by those metrics. In the DRIVDP case however, the output of the metric is three separate maps for contrast loss, amplification and reversal. Thus, it is not clear how to produce a single distortion map for HDR-HDR and LDR-LDR stimuli. After experimenting with various methods for combining the distortion maps predicted by DRIVDP, we found that the combined map defined as:

$$P^{k,l,m}_{combined} = 1 - (1 - P^{k,l,m}_{loss}) \cdot (1 - P^{k,l,m}_{ampl}), \qquad (8.1)$$

gives the best correlation with subjective data. Here, $P^{k,l,m}_{loss|ampl}$ refer to the detection probability of contrast loss and amplification at scale k, orientation l, and temporal channel m. The resulting map $P^{k,l,m}_{combined}$ corresponds to the probability of detecting either contrast loss or amplification at a visual channel. Leaving contrast reversal resulted in slightly improved correlations.

8.4 Conclusion

The high correlations between the metric predictions and subjective responses over a diverse test set including HDR and LDR stimuli with distortions of various type and magnitude indicate that the metric proposed in Chapter 7 provides a reliable estimate of the video quality as a function of spatial location.

We believe the establishment of a public, standardized test set containing video pairs with diverse dynamic ranges and types of artifacts, coupled with spatially varying corresponding subjective responses, is essential for this line of research. As future work, we would like to extend our data set and make it publicly available as a first step in that direction.

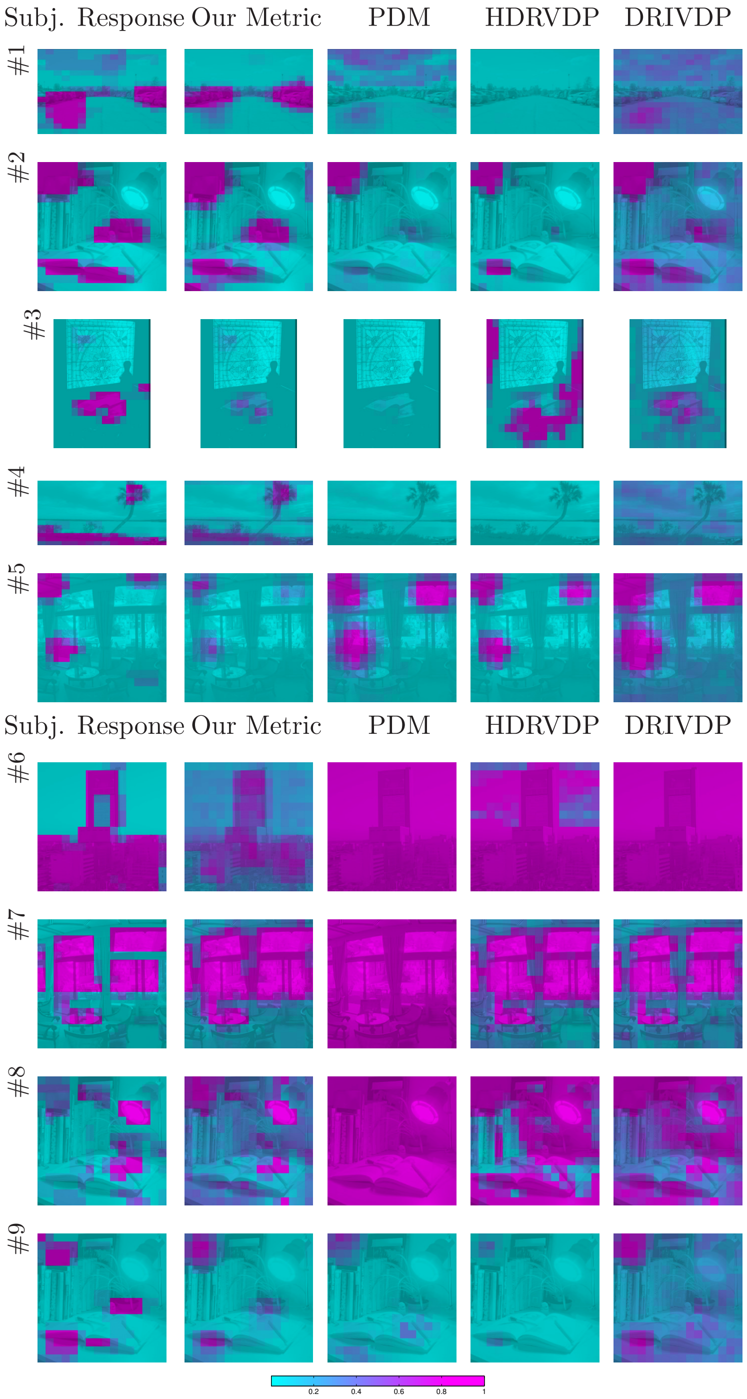


Figure 8.3: *Mean subjective response maps and corresponding metric predictions pairs.*

St. #	**Subjective Response**	**Our Metric**	**PDM**	**HDRVDP**	**DRIVDP**
	[min, max]; avg; std	[min, max]; avg; std	[min, max]; avg; std	[min, max]; avg; std	[min, max]; avg; std
1	[0.000, 1.000]; 0.177; 0.276	[0.000, 0.850]; 0.128; 0.230	[0.000, 0.301]; 0.082; 0.079	[0.000, 0.019]; 0.001; 0.002	[0.075, 0.417]; 0.194; 0.058
2	[0.000, 1.000]; 0.201; 0.347	[0.000, 0.954]; 0.185; 0.282	[0.000, 0.813]; 0.061; 0.138	[0.000, 0.893]; 0.050; 0.157	[0.072, 0.799]; 0.218; 0.155
3	[0.000, 1.000]; 0.082; 0.242	[0.000, 0.307]; 0.015; 0.045	[0.000, 0.052]; 0.003; 0.008	[0.000, 0.889]; 0.163; 0.247	[0.006, 0.440]; 0.090; 0.078
4	[0.000, 1.000]; 0.124; 0.250	[0.001, 0.457]; 0.094; 0.115	[0.000, 0.024]; 0.007; 0.006	[0.000, 0.000]; 0.000; 0.000	[0.067, 0.240]; 0.137; 0.039
5	[0.000, 1.000]; 0.066; 0.186	[0.000, 0.420]; 0.026; 0.063	[0.000, 0.952]; 0.146; 0.207	[0.000, 0.866]; 0.074; 0.166	[0.040, 0.873]; 0.241; 0.199
6	[0.000, 1.000]; 0.399; 0.389	[0.072, 0.468]; 0.232; 0.103	[0.810, 0.984]; 0.965; 0.026	[0.180, 0.942]; 0.657; 0.202	[0.626, 0.928]; 0.789; 0.058
7	[0.000, 1.000]; 0.312; 0.392	[0.037, 0.984]; 0.451; 0.342	[0.838, 0.984]; 0.980; 0.018	[0.002, 0.953]; 0.448; 0.327	[0.031, 0.953]; 0.374; 0.288
8	[0.000, 0.812]; 0.108; 0.180	[0.041, 0.942]; 0.225; 0.146	[0.606, 0.984]; 0.971; 0.043	[0.005, 0.953]; 0.509; 0.274	[0.148, 0.884]; 0.406; 0.172
9	[0.000, 1.000]; 0.105; 0.238	[0.000, 0.502]; 0.054; 0.104	[0.000, 0.396]; 0.032; 0.066	[0.000, 0.211]; 0.006; 0.025	[0.067, 0.577]; 0.176; 0.097

Table 8.2: *Descriptive statistics of distortion maps (depicted in Figure 8.3) for each input stimulus. Abbreaviations used: min=minimal value, max=maximal value, avg=average value, std=standard deviation.*

Stimulus #	Our Metric	PDM	HDRVDP	DRIVDP
1	**0.765**	-0.0147	0.591	0.488
2	**0.883**	0.686	0.673	0.859
3	0.843	**0.886**	0.0769	0.865
4	**0.815**	0.0205	0.211	-0.0654
5	**0.844**	0.565	0.803	0.689
6	**0.761**	-0.462	0.709	0.299
7	0.879	0.155	0.882	**0.924**
8	**0.733**	0.109	0.339	0.393
9	**0.753**	0.368	0.473	0.617
Average	**0.809**	0.257	0.528	0.563

Table 8.3: *Correlations of subjective responses with predictions of our metric, PDM, HDRVDP, and DRIVDP. The last row shows the average correlations over the test set, the best correlations for each stimulus are printed in bold text.*

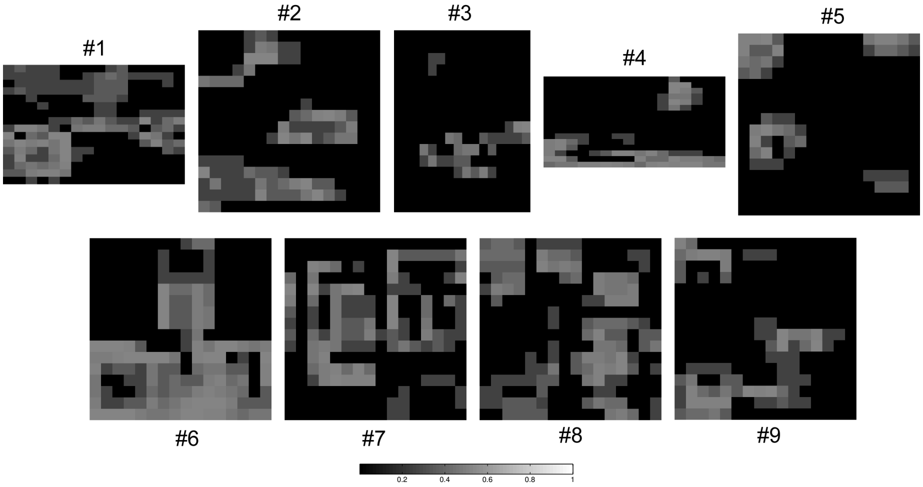


Figure 8.4: *Maps showing the standard deviations over subjects for each stimulus. The numbers refer to the first column of Table 8.1.*

Chapter 9

Conclusions and Future Work

In this final section of this dissertation we will state our conclusions and give directions for future research.

9.1 Conclusions

The main motivation of this dissertation was to explore the use of human visual system models in computer graphics context. We presented human visual system models with various scopes and complexities designed for specific types of applications, and demonstrated their merit in terms of extending the functionality of the state of the art, and improving application performance.

We have shown that the strength of image edges can be more accurately predicted using simple perceptual models compared to purely mathematical measures such as the gradient magnitude (Chapter 3). We also showed that such a perceptual edge strength measure can be integrated into a second generation wavelet-based image decomposition without a prohibitive computational cost. A more complex human visual system model, also accounting for the maladaptation of the observer due to the dynamically changing lighting conditions was presented in Chapter 4. Using this model we predicted the visibility of images shown on a desktop computer display, and a simulated car interior display.

We have also investigated the detection problem in the context of image and video quality assessment. An important limitation of previous quality assessment methods has been the lack of HDR support. We addressed this issue in Chapter 5, where we introduced an extension to popular simple image quality measures that enables them to process HDR images. An important property of this method is the backwards compatibility with LDR content, that is: the extended metrics would still predict the same quality for LDR images, in addition to their ability to predict the quality of HDR images. However merely supporting HDR content is not enough for tasks such as tone mapping operator evaluation, where the reference and test images have different dynamic ranges. To that end, we proposed a dynamic range independent image quality assessment method in Chapter 6, that for the first time enabled objective comparison of tone mapping operators.

Similarly, we proposed a dynamic range independent video quality assessment method (Chapter 7), where we extended the steady-state human visual system model with temporal mechanisms. We have shown that objective quality assessment is possible for comparison of different rendering qualities, HDR video compression and temporal tone mapping.

In many instances throughout this work we calibrated and validated the human visual system models we used. While the psychophysical experiments have been discussed in the corresponding chapters, in Chapter 8 we present a novel method of spatial evaluation of video sequences, which enables the assessment of local distortions.

9.2 Future Work

The results of this thesis suggest that exploiting the properties of the human visual system is beneficial while processing visual data . In hindsight, this may seem an obvious statement, since the ultimate receiver of the visual data are humans. That said we have presented solid scientific evidence to back up this claim, and presented working solutions to computer graphics related problems.

An immediate future direction of our research is to further integrate physiological and psychophysical findings on human visual perception into the methods that have been used in the subfields of computer science dealing with visual data. This task has two aspects, on one hand it is desirable to make the human visual system models generally more accurate, but on the other hand one should identify the specific application needs of the target method and design more constrained but efficient models of human vision.

A more specific problem we encountered during the course of this work was the absence of a standardized set of images and videos of various dynamic ranges, containing a variety of distortion types along with their subjectively determined quality estimates. Often such data sets have a limited scope, for example the Video Quality Experts Group's (VQEG) set of distorted videos is limited compression artifacts in LDR videos. Moreover, most of these sets represent the quality of an entire image or video sequence with a single number. A spatially variant quality estimate, similar to the in-context maps produced by the quality metrics discussed in Chapters 6 and 7, would be far more informative. To that end we created a small data set of both HDR and LDR videos containing a multitude of distortions (Chapter 8). The quality values were subjectively measured on a 16×16 grid, adding a certain level of spatial variation. A similar, but more comprehensive data set would potentially enable standardization and meaningful comparison between various human visual system models and quality assessment methods.

In this dissertation we also presented models that take solely luminance as their input. Obviously we don't see the world in grayscale; even though one could arguably perform most visual tasks even in the absence of chrominance information, color perception is still an important and very interesting aspect of human vision. A future direction of our work is to utilize the recent developments in HDR color perception [Kim *et al.*, 2009] in the human visual system models.

Chapter 10

Appendix

In this appendix we present fundamental formulas that were left out from the dissertation for brevity. Note that all formulas can be found in referenced articles, we merely recollected them for completeness and ease of implementation.

10.1 JND Space

The JND space nonlinearity accounts for lower sensitivity of the photoreceptors at low luminance, where the luminance L is transformed using a transfer function constructed from the peak detection thresholds [Mantiuk *et al.*, 2005]. One can construct such a transfer function from the following recursive formula:

$$T_{inv}[i] = T_{inv}[i-1] + cvi(T_{inv}[i-1])\, T_{inv}[i-1] \quad \text{for } i = 2..N, \tag{10.1}$$

where $T_{inv}[1]$ is the minimum luminance we want to consider (10^{-5} cd/m^2 in our case). The actual photoreceptor response R is found by linear interpolation between the pair of i values corresponding to particular luminance L.

The contrast versus intensity function cvi used in the recursive formula above estimates the lowest detection threshold at a particular adaptation level:

$$cvi(L_a) = \left(\max_{\mathbf{x}} \left[CSF^S(\mathbf{x}, L_a)\right]\right)^{-1}, \tag{10.2}$$

where CSF^S is the static contrast sensitivity function and $\mathbf{x}$ are all its parameters except adaptation luminance. If perfect local adaptation is assumed, then $L_a = L$.

10.2 Static Contrast Sensitivity Function

The static contrast sensitivity function (CSF^S) describes the sensitivity of the visual system as a function of spatial frequency and adaptation luminance. In our implementation we use the CSF proposed by Daly [1993]:

$$CSF^S(\rho, L_a, \theta, i^2, d, c) = P \cdot \min\left[S_1\left(\frac{\rho}{r_a \cdot r_c \cdot r_\theta}\right), S_1(\rho)\right], \tag{10.3}$$

where

$$\begin{array}{rl} r_a = & 0.856 \cdot d^{0.14} \\ r_c = & \frac{1}{1+0.24c} \\ r_\theta = & 0.11\cos(4\theta) + 0.11 \\ S_1(\rho) = & \left[\left(3.23(\rho^2 i^2)^{-0.3}\right)\right)^5 + 1\right]^{-\frac{1}{5}} \cdot \\ & A_l \epsilon \rho e^{-(B_l \epsilon \rho)} \sqrt{1 + 0.06 e^{B_l \epsilon \rho}} \\ A_l = & 0.801\left(1 + 0.7\, L_a^{-1}\right)^{-0.2} \\ B_l = & 0.3\left(1 + 100\, L_a^{-1}\right)^{-0.15}. \end{array} \tag{10.4}$$

The parameters are:

- ρ – spatial frequency in cycles per visual degree,
- L_a – light adaptation level in cd/m^2,
- θ – orientation,
- i^2 – stimulus size in deg^2 ($i^2 = 1$),
- d – distance in meters,
- c – eccentricity ($c = 0$),
- ϵ – constant ($\epsilon = 0.9$),
- P – absolute peak sensitivity ($P = 250$).

Note that the formulas for A_l and B_l contain the corrections found after the correspondence with the author of the original publication (Scott Daly).

Since the filter function depends on the local luminance of adaptation, the same kernel cannot be used for the entire image. To speed up computations, the response map R is filtered six times assuming $L_a = \{$ 0.001, 0.01, 0.1, 1, 10, 100 $\}$ cd/m^2 and the final value for each pixel is found by the linear interpolation between the two filtered maps closest to the L_a for a given pixel.

10.3 Spatiotemporal Contrast Sensitivity Function

The spatiotemporal contrast sensitivity function (CSF^T), on the other hand, models the variation of contrast sensitivity at a fixed adaptation luminance (100 cd/m^2 in this case) as a function of spatial and temporal frequencies. In our work we derive spatiotemporal CSF from the following spatiovelocity CSF formula [Daly, 1998]:

$$\begin{array}{l} CSF^T(\rho, v) = \\ c_0\left(6.1 + 7.3|log\left(\frac{c_2 v}{3}\right)|^3\right) c_2 v \left(2\pi c_1 \rho\right)^2 exp\left(-\frac{4\pi c_1 \rho (c_2 v + 2)}{45.9}\right), \end{array} \tag{10.5}$$

where

- ρ is the spatial frequency in cycles per visual degree,
- v is the retinal velocity in degrees per second,
- $c_0 = 1.14$,
- $c_1 = 0.67$,
- $c_2 = 1.7$.

The last three coefficients ensure that the CSF^T for $v = 0.15$ is close to the CSF^S for $L_a = 100\ cd/m^2$. The relation between spatiotemporal frequency ω and retinal velocity is $\omega = v\rho$ assuming the retina is stable.

10.4 Cortex Transform for Images

The 2D Cortex Transform [Daly, 1993] is a collection of the band-pass and orientation selective filters. The band-pass filters are computed as:

$$dom_k = \begin{cases} mesa_{k-1} - mesa_k & \text{for } k = 1..K-2 \\ mesa_{k-1} - base & \text{for } k = K-1 \end{cases} \tag{10.6}$$

where K is the total number of spatial bands and the low-pass filters $mesa_k$ and $baseband$ have the form:

$$\begin{array}{ll} mesa_k = & \begin{cases} 1 & \text{for } \rho \le r - \frac{tw}{2} \\ \frac{1}{2}\left(1 + \cos\left(\frac{\pi(\rho - r + \frac{tw}{2})}{tw}\right)\right) & \text{for } r - \frac{tw}{2} < \rho \le r + \frac{tw}{2} \\ 0 & \text{for } \rho > r + \frac{tw}{2} \end{cases} \\ base = & \begin{cases} e^{-\frac{\rho^2}{2\sigma^2}} & \text{for } \rho < r_{K-1} + \frac{tw}{2} \\ 0 & \text{otherwise,} \end{cases} \end{array} \tag{10.7}$$

where

$$r = 2^{-k}, \quad \sigma = \frac{1}{3}\left(r_{K-1} + \frac{tw}{2}\right) \text{ and } \quad tw = \frac{2}{3}r. \tag{10.8}$$

The orientation selective filters are defined as:

$$fan_l = \begin{cases} \frac{1}{2}\left(1 + cos\left(\frac{\pi|\theta - \theta_c(l)|}{\theta_{tw}}\right)\right) & \text{for } |\theta - \theta_c(l)| \le \theta_{tw} \\ 0 & \text{otherwise,} \end{cases} \tag{10.9}$$

where $\theta_c(l)$ is the orientation of the center, $\theta_c(l) = (l-1) \cdot \theta_{tw} - 90$, and θ_{tw} is the transitional width, $\theta_{tw} = 180/L$. The cortex filter is formed by the product of the dom and fan filters:

$$B^{k,l} = \begin{cases} dom_k \cdot fan_l & \text{for } k = 1..K-1 \text{ and } l = 1..L \\ base & \text{for } k = K. \end{cases} \tag{10.10}$$

Bibliography

[Adrian, 1989] W. Adrian. Visibility of targets: Model for calculation. In *Lighting Res. Technol.*, volume 21, pages 181–188, 1989.

[Ahumada *et al.*, 2006] A. J. Ahumada, M. T San-Martin, and J. Gille. Symbol discriminability models for improved flight displays. In *Proc. of SPIE: Human vision and electronic imaging XI*, volume 6057, 2006.

[Akyüz *et al.*, 2007] A. O. Akyüz, E. Reinhard, R. Fleming, B. E. Riecke, and H. H. Bülthoff. Do HDR displays support LDR content? a psychophysical evaluation. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 26(3), 2007. Article 38.

[Avidan and Shamir, 2007] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. *ACM Transactions on Graphics*, 26(3):10, 2007.

[Aydın *et al.*, 2008a] T. O. Aydın, R. Mantiuk, K. Myszkowski, and H-P. Seidel. Dynamic range independent image quality assessment. In *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, volume 27(3), 2008. Article 69.

[Aydın *et al.*, 2008b] T. O. Aydın, R. Mantiuk, and H-P. Seidel. Extending quality metrics to full luminance range images. In *Proc. of SPIE: Human Vision and Electronic Imaging XIII*, volume 6806, 2008.

[Aydın *et al.*, 2009] T. O. Aydın, K. Myszkowski, and H-P. Seidel. Predicting display visibility under dynamically changing lighting conditions. *Computer Graphics Forum (Proc. of EUROGRAPHICS)*, 28(3), 2009.

[Aydın *et al.*, 2010a] T. O. Aydın, M. Čadík, K. Myszkowski, and H-P. Seidel. Video quality assessment for computer graphics applications. In *To Appear in: ACM Transactions on Graphics (Proc. of SIGGRAPH Asia)*, 2010.

[Aydın *et al.*, 2010b] T. O. Aydın, M. Čadík, K. Myszkowski, and H-P. Seidel. Visually significant edges. *ACM Transactions on Applied Perception*, 7(4):1–15, 2010.

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

[Bavoil *et al.*, 2008] L. Bavoil, M. Sainz, and R. Dimitrov. Image-space horizon-based ambient occlusion. In *SIGGRAPH '08: ACM SIGGRAPH 2008 talks*, 2008.

[Bittner *et al.*, 2004] J. Bittner, M. Wimmer, H. Piringer, and W. Purgathofer. Coherent hierarchical culling: Hardware occlusion queries made useful. *Computer Graphics Forum (Proc. of EUROGRAPHICS*, 23(3):615–624, 2004.

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

[Boynton G M, 1999] Foley J M Boynton G M. Temporal sensitivity of human luminance pattern mechanisms determined by masking with temporally modulated stimuli. *Vision Research*, 39(9):1641–1656, 1999.

[Burt and Adelson, 1983] P. J. Burt and E. H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, COM-31(4):532–540, 1983.

[Čadík *et al.*, 2008] M. Čadík, M. Wimmer, L. Neumann, and A. Artusi. Evaluation of HDR tone mapping methods using essential perceptual attributes. *Computers & Graphics*, 32:330–349, 2008.

[Canny, 1986] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, 1986.

[Cole *et al.*, 2008] F. Cole, A. Golovinskiy, A. Limpaecher, H. Stoddart Barros, A. Finkelstein, T. Funkhouser, and S. Rusinkiewicz. Where do people draw lines? *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 27(3), 2008.

[Cramér, 1999] H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1999.

[Dachsbacher and Stamminger, 2005] C. Dachsbacher and M. Stamminger. Reflective shadow maps. In *I3D '05: Proc. of Symposium on Interactive 3D Graphics and Games*, pages 203–231, 2005.

[D'Agostino, 1972] R.B. D'Agostino. Relation between the chi-squared and ANOVA test for testing equality of k independent dichotomous populations. *The American Statistician*, 26:30–32, 1972.

[Daly, 1993] S. Daly. The Visible Differences Predictor: An algorithm for the assessment of image fidelity. In *Digital Images and Human Vision*, pages 179–206. MIT Press, 1993.

[Daly, 1998] S. Daly. Engineering observations from spatiovelocity and spatiotemporal visual models. In *Proc. of SPIE: Human Vision and Electronic Imaging III*, volume 3299, 1998.

[DeCarlo and Santella, 2002] D. DeCarlo and A. Santella. Stylization and abstraction of photographs. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 21(3):769–776, 2002.

[Deeley *et al.*, 1991] R. J. Deeley, N. Drasdo, and W. N. Charman. A simple parametric model of the human ocular modulation transfer function. *Ophthalmology and Physiological Optics*, 11:91–93, 1991.

[DICOM, 2001] Part 14: Grayscale standard display function. In *Digital Imaging and Communications in Medicine (DICOM)*. 2001.

[Dmitriev *et al.*, 2004] K. Dmitriev, T. Annen, G. Krawczyk, K. Myszkowski, and H-P. Seidel. A CAVE system for interactive modeling of global illumination in car interior. In *ACM Symposium on Virtual Reality Software and Technology (VRST 2004)*, pages 137–145, 2004.

[Drago *et al.*, 2003a] F. Drago, K. Myszkowski, T. Annen, and N. Chiba. Adaptive logarithmic mapping for displaying high contrast scenes. In *Proc. of Eurographics*, pages 419–426, 2003.

[Drago *et al.*, 2003b] F. Drago, K. Myszkowski, T. Annen, and N.Chiba. Adaptive logarithmic mapping for displaying high contrast scenes. *Computer Graphics Forum (Proc. of EUROGRAPHICS)*, 22(3), 2003.

[Dreyer, 2007] D. Dreyer. *Kontrastschwellensimulation fuer Sichtbarkeitsuntersuchungen an Displays*. Dissertation, Technische Universitaet Muenchen, 2007.

[Durand and Dorsey, 2002] F. Durand and J. Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 21(3):257–266, 2002.

[Elder and Goldberg, 2001] J. H. Elder and R. M. Goldberg. Image editing in the contour domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):291–296, 2001.

[Farbman *et al.*, 2008] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski. Edge-preserving decompositions for multi-scale tone and detail manipulation. In *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, volume 27, 2008.

[Fattal *et al.*, 2002] R. Fattal, D. Lischinski, and M. Werman. Gradient domain high dynamic range compression. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 21(3):249–256, 2002.

[Fattal, 2009] R. Fattal. Edge-avoiding wavelets and their applications. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 28(3), 2009.

[Ferwerda *et al.*, 1996] J. A. Ferwerda, S. Pattanaik, P. Shirley, and D. P. Greenberg. A model of visual adaptation for realistic image synthesis. In Holly Rushmeier, editor, *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, pages 249–258, 1996.

[Ferwerda *et al.*, 1997] J. A. Ferwerda, S. N. Pattanaik, P. S. Shirley, and D. P. Greenberg. A model of visual masking for computer graphics. In *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, pages 143–152, 1997.

[Fredericksen, 1998] R. F. Fredericksen, R. E. Estimating multiple temporal mechanisms in human vision. In *Vision Research*, 1998.

[Freeman and Adelson, 1991] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 13(9):891–906, 1991.

[Georgeson and Sullivan, 1975] M. A. Georgeson and G. D. Sullivan. Contrast constancy: deblurring in human vision by spatial frequency channels. *The Journal of Physiology*, 252(3):627–656, 1975.

[Georgeson *et al.*, 2007] M. A. Georgeson, K. A. May, T. C. Freeman, and G. S. Hesse. From filters to features: scale-space analysis of edge and blur coding in human vision. *Journal of Vision*, 7(13), 2007.

[Herzog *et al.*, 2010] R. Herzog, E. Eisemann, K. Myszkowski, and H-P. Seidel. Spatio-temporal upsampling on the GPU. In *I3D '10: Proc. of Symposium on Interactive 3D Graphics and Games*, 2010.

[Irawan *et al.*, 2005] P. Irawan, J. A. Ferwerda, and S. R. Marschner. Perceptually based tone mapping of high dynamic range image streams. In *Eurographics Symposium on Rendering*, pages 231–242, 2005.

[ITU-T, 1999] ITU-T. Subjective video quality assessment methods for multimedia applications. 1999.

[Jansen and Oonincx, 2005] M. H. Jansen and P. J. Oonincx. *Second Generation Wavelets and Applications*. Springer, 2005.

[Janssen, 2001] R. Janssen. *Computational Image Quality*. SPIE Press, 2001.

[Kelly, 1983] D. H. Kelly. Spatiotemporal variation of chromatic and achromatic contrast thresholds. *Journal of the Optical Society of America*, 73(6), 1983.

[Kim *et al.*, 2009] M. H. Kim, T. Weyrich, and J. Kautz. Modeling human color perception under extended luminance levels. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3):27:1–9, 2009.

[Krantz and Silverstein, 1992] J.H. Krantz and L.D. Silverstein. Visibility of transmissive liquid crystal displays under dynamic lighting conditions. *Human Factors*, 34(5):615–632, 1992.

[Kuang *et al.*, 2007a] J. Kuang, G. M. Johnson, and M. D. Fairchild. icam06: A refined image appearance model for hdr image rendering. *Journal of Visual Communication and Image Representation*, 18(5):406–414, 2007.

[Kuang *et al.*, 2007b] J. Kuang, H. Yamaguchi, C. Liu, G. M. Johnson, and M. D. Fairchild. Evaluating HDR rendering algorithms. *ACM Transactions on Applied Perception*, 4(2):9, 2007.

[Legge and Foley, 1980] G.E. Legge and J.M. Foley. Contrast masking in human vision. *Journal of the Optical Society of America*, 70(12):1458–1471, December 1980.

[Lindeberg, 1996] T. Lindeberg. Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116, 1996.

[Lindh and van den Branden Lambrecht, 1996] P. Lindh and C. van den Branden Lambrecht. Efficient spatio-temporal decomposition for perceptual processing of video sequences. In *Proceedings of International Conference on Image Processing ICIP'96*, volume 3 of *Proc. of IEEE*, pages 331–334. IEEE, 1996.

[Lubin, 1993] J. Lubin. The use of psychophysical data and models in the analysis of display system performance. In *Digital Images and Human Vision*, pages 163–178. MIT Press, 1993.

[Lubin, 1995a] J. Lubin. *Vision Models for Target Detection and Recognition*, chapter A Visual Discrimination Model for Imaging System Design and Evaluation, pages 245–283. World Scientific, 1995.

[Lubin, 1995b] J. Lubin. A visual discrimination model for imaging system design and evaluation. *Vision Models for Target Detection and Recognition*, pages 245–283, 1995.

[Lukin, 2009] A. Lukin. Improved visible differences predictor using a complex cortex transform. *GraphiCon*, pages 145–150, 2009.

[Mantiuk *et al.*, 2004] R. Mantiuk, G. Krawczyk, K. Myszkowski, and H-P. Seidel. Perception-motivated high dynamic range video encoding. *ACM Transactions on Graphics (Proc. of SIGGRAPH*, 23(3):733–741, 2004.

[Mantiuk *et al.*, 2005] R. Mantiuk, S. Daly, K. Myszkowski, and H-P. Seidel. Predicting visible differences in high dynamic range images - model and its calibration. In *Proc. of SPIE: Human Vision and Electronic Imaging X*, volume 5666, pages 204–214, 2005.

[Mantiuk *et al.*, 2006a] R. Mantiuk, K. Myszkowski, and H-P. Seidel. Lossy compression of high dynamic range images and video. In *Proc. of SPIE: Human Vision and Electronic Imaging XI*, page 60570V, 2006.

[Mantiuk *et al.*, 2006b] R. Mantiuk, K. Myszkowski, and H-P. Seidel. A perceptual framework for contrast processing of high dynamic range images. *ACM Transactions on Applied Perception*, 3(3):286–308, 2006.

[Mantiuk *et al.*, 2008] R. Mantiuk, S. Daly, and L. Kerofsky. Display adaptive tone mapping. In *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, volume 27(3), 2008. Article 68.

[Marimont and Wandell, 1994] David H. Marimont and Brian A. Wandell. Matching color images: the effects of axial chromatic aberration. *J. Opt. Soc. Am. A*, 11(12):3113–3122, 1994.

[Marr and Hildreth, 1980] D. Marr and E. Hildreth. Theory of edge detection. *Proc. Royal Soc. Lond.*, B(207):187–217, 1980.

[Masry and Hemami, 2004] M. A. Masry and S. S. Hemami. A metric for continuous quality evaluation of compressed video with severe distortions. *Signal Processing: Image Communication*, 19(2):133 – 146, 2004.

[Meylan *et al.*, 2007] L. Meylan, S. Daly, and S. Susstrunk. Tone mapping for high dynamic range displays. In *Human Vision and Electronic Imaging XII*, SPIE, volume 6492, 2007.

[Myszkowski *et al.*, 2000] K. Myszkowski, P. Rokita, and T. Tawara. Perception-based fast rendering and antialiasing of walkthrough sequences. *IEEE Transactions on Visualization and Computer Graphics*, 6(4):360–379, 2000.

[Myszkowski *et al.*, 2001] K. Myszkowski, T. Tawara, H. Akamine, and H-P. Seidel. Perception-guided global illumination solution for animation rendering. In *SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 221–230, 2001.

[Naka and Rushton, 1966] K. I. Naka and W. A. H. Rushton. S-potentials from luminosity units in the retina of fish (cyprinidae). *Journal of Physiology*, 185:587–599, 1966.

[Nakamae *et al.*, 1990] E. Nakamae, K. Kaneda, T. Okamoto, and T. Nishita. A lighting model aiming at drive simulators. In *Computer Graphics (Proceedings of SIGGRAPH)*, pages 395–404, 1990.

[Orzan *et al.*, 2007] A. Orzan, A. Bousseau, P. Barla, and J. Thollot. Structure-preserving manipulation of photographs. In *Int. Symposium on Non-Photorealistic Animation and Rendering*, 2007.

[Pattanaik *et al.*, 1998] S. N. Pattanaik, J. A. Ferwerda, M. Fairchild, and D. P. Greenberg. A multiscale model of adaptation and spatial vision for realistic image display. In *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, pages 287–298, 1998.

[Pattanaik *et al.*, 2000] S. N. Pattanaik, J. Tumblin, H. Yee, and D. P. Greenberg. Time-dependent visual adaptation for fast realistic image display. In *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 47–54, 2000.

[Peli, 1990] E. Peli. Contrast in complex images. *J. Opt. Soc. Am.*, 7(10):2032–2040, 1990.

[Pellegrino *et al.*, 2004] F. A. Pellegrino, W. Vanzella, and V. Torre. Edge detection revisited. *Systems, Man, and Cybernetics, Part B, IEEE Transactions on*, 34(3):1500–1518, 2004.

[Perez *et al.*, 2003] P. Perez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Transactions on Graphics*, 22(3):313–318, 2003.

[Perona and Malik, 1990] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:629–639, 1990.

[Ramanarayanan *et al.*, 2007] G. Ramanarayanan, J. Ferwerda, B. Walter, and K. Bala. Visual Equivalence: Towards a new standard for Image Fidelity. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 26(3), 2007. Article 76.

[Ramasubramanian *et al.*, 1999] M. Ramasubramanian, S. N. Pattanaik, and D. P. Greenberg. A perceptually based physical error metric for realistic image synthesis. In *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, pages 73–82, 1999.

[Rea and Ouellette, 1991] M.S. Rea and M.J. Ouellette. Relative visual performance: a basis for application. *Lighting Research and Technology*, 23(3):135–144, 1991.

[Reeves *et al.*, 1987] W. T. Reeves, D. H. Salesin, and R. L. Cook. Rendering antialiased shadows with depth maps. In *SIGGRAPH '87: Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 283–291, 1987.

[Reinhard *et al.*, 2002] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda. Photographic tone reproduction for digital images. In *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, pages 267–276. ACM Press, 2002.

[Reinhard *et al.*, 2005] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. Morgan Kauffman, 2005.

[Rempel *et al.*, 2007] A. G. Rempel, M. Trentacoste, H. Seetzen, H. D. Young, W. Heidrich, L. Whitehead, and G. Ward. Ldr2Hdr: On-the-fly reverse tone mapping of legacy video and photographs. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 26(3), 2007. Article 39.

[Ritschel *et al.*, 2009a] T. Ritschel, T. Grosch, and H-P. Seidel. Approximating dynamic global illumination in image space. In *I3D '09: Proc. of Symposium on Interactive 3D Graphics and Games*, pages 75–82, 2009.

[Ritschel *et al.*, 2009b] T. Ritschel, M. Ihrke, J. R. Frisvad, J. Coppens, K. Myszkowski, and H.-P. Seidel. Temporal glare: Real-time dynamic simulation of the scattering in the human eye. *Computer Graphics Forum*, 28(2):183–192, April 2009.

[Rubinstein *et al.*, 2009] M. Rubinstein, A. Shamir, and S. Avidan. Multi-operator media retargeting. In *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 2009.

[Rushmeier *et al.*, 1995] H. Rushmeier, G. Ward, C. Piatko, P. Sanders, and B. Rust. Comparing real and synthetic images: some ideas about metrics. In *Rendering Techniques '95*, pages 82–91. Springer, 1995.

[Ruzon and Tomasi, 1999] M. A. Ruzon and C. Tomasi. Color edge detection with the compass operator. In *Computer Vision and Pattern Recognition*, volume 2, page 166 Vol. 2, 1999.

[Sampat *et al.*, 2009] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey. Complex wavelet structural similarity: A new image similarity index. *Image Processing, IEEE Transactions on*, 18(11):2385–2401, 2009.

[Schwarz and Stamminger, 2009] M. Schwarz and M. Stamminger. On predicting visual popping in dynamic scenes. In *Proceedings of Symposium on Applied Perception in Graphics and Visualization (APGV)*, pages 93–100, 2009.

[Seetzen *et al.*, 2004] H. Seetzen, W. Heidrich, W. Stuerzlinger, G. Ward, L. Whitehead, M. Trentacoste, A.t Ghosh, and A. Vorozcovs. High dynamic range display systems. In *Proc. of ACM SIGGRAPH 2004*, volume 23, pages 760–768, 2004.

[Seshadrinathan and Bovik, 2007] K. Seshadrinathan and A.C. Bovik. A structural similarity metric for video based on motion models. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1, pages I–869–I–872, 2007.

[Seshadrinathan and Bovik, 2010] K. Seshadrinathan and A. C. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Transactions on Image Processing*, 2010.

[Sharpe *et al.*, 2003] R. Sharpe, C. M. Cartwright, W. A. Gillespie, K. Vassie, and W. C. Christopher. Sunlight readability of displays: a numerical scale. In *Proc. of SPIE*, volume 4826, 2003.

[Smith *et al.*, 2006] K. Smith, G. Krawczyk, K. Myszkowski, and H-P. Seidel. Beyond tone mapping: Enhanced depiction of tone mapped HDR images. *Computer Graphics Forum (Proc. of EUROGRAPHICS)*, 25(3):427–438, 2006.

[Spencer *et al.*, 1995] G. Spencer, P. S. Shirley, K. Zimmerman, and Greenberg D. P. Physically-based glare effects for digital images. In *Proceedings of SIGGRAPH 95*, Computer Graphics Proceedings, Annual Conference Series, pages 325–334, 1995.

[Sweldens, 1997] W. Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.*, 29(2):511–546, 1997.

[Taylor and Creelman, 1967] M. M. Taylor and C. D. Creelman. PEST: Efficient estimates on probability functions. *The Journal of the Acoustical Society of America*, 41(4A):782–787, 1967.

[Tumblin *et al.*, 1999] J. Tumblin, J. K. Hodgins, and B. K. Guenter. Two methods for display of high contrast images. *ACM Transactions on Graphics*, 18(1):56–94, January 1999. ISSN 0730-0301.

[Ueno *et al.*, 1985] T. Ueno, J. Pokorny, and C. S. Vivianne. Reaction times to chromatic stimuli. *Vision Research*, 25(11):1623–1627, 1985.

[Uytterhoeven *et al.*, 1997] G. Uytterhoeven, D. Roose, and A. Bultheel. Wavelet transforms using the lifting scheme, 1997.

[van den Branden Lambrecht and Verscheure, 1996] C. van den Branden Lambrecht and O. Verscheure. Perceptual Quality Measure using a Spatio-Temporal Model of the Human Visual System. In *Proc. of SPIE*, 1996.

[van den Branden Lambrecht *et al.*, 1999] C. van den Branden Lambrecht, D.M. Costantini, G.L. Sicuranza, and M. Kunt. Quality assessment of motion rendition in video coding. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(5):766–782, August 1999.

[Walraven *et al.*, 1990] J. Walraven, C. Enroth-Cugell, D.C. Hood, D.I.A. MacLeod, and J.L. Schnapf. The control of visual sensitivity: receptoral and postreceptoral processes. *Visual Perception: The Neurophysiological Foundations*, pages 53–101, 1990.

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

[Wang and Bovik, 2002] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, March 2002.

[Wang and Bovik, 2006] Z. Wang and A. C. Bovik. *Modern Image Quality Assessment.* Morgan & Claypool Publishers, 2006.

[Wang and Simoncelli, 2005] Z. Wang and E. Simoncelli. Translation insensitive image similarity in complex wavelet domain. In *Acoustics, Speech, and Signal*

Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on, volume 2, pages 573–576, 2005.

[Wang *et al.*, 2003] Z. Wang, E. Simoncelli, and A. Bovik. Multi-scale structural similarity for image quality assessment, 2003.

[Wang *et al.*, 2008] Y-S. Wang, C-L. Tai, O. Sorkine, and T-Y. Lee. Optimized scale-and-stretch for image resizing. In *ACM SIGGRAPH Asia*, 2008.

[Ward, 2006] G. Ward. Hiding seams in high dynamic range panoramas. In *APGV '06: Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization*, page 150. ACM, 2006.

[Watson and Malo, 2002] A. B. Watson and J. Malo. Video quality measures based on the standard spatial observer. In *ICIP (3)*, pages 41–44, 2002.

[Watson and Solomon, 1997] A. B. Watson and J. A. Solomon. A model of visual contrast gain control and pattern masking. *J. Opt. Soc. Am.*, A(14):2379–2391, 1997.

[Watson *et al.*, 2001] A.B. Watson, J. Hu, and J.F. McGowan III. Digital video quality metric based on human vision. *Journal of Electronic Imaging*, 10:20, 2001.

[Watson, 1986] A. B. Watson. Temporal sensitivity. In *Handbook of Perception and Human Performance*, pages 6–1– 6–43. John Wiley and Sons, New York, 1986.

[Watson, 1987] A. B. Watson. The Cortex transform: rapid computation of simulated neural images. *Comp. Vision Graphics and Image Processing*, 39:311–327, 1987.

[Watson, 2000] Andrew B. Watson. Visual detection of spatial contrast patterns: Evaluation of five simple models. *Optics Express*, 6(1):12–33, 2000.

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

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

[Winkler, 1999] S. Winkler. A perceptual distortion metric for digital color video. In *Proceedings of the SPIE: Human vision and electronic imaging*, volume 3644 of *Controlling Chaos and Bifurcations in Engineering Systems*, pages 175–184, 1999.

[Winkler, 2005] Stefan Winkler. *Digital Video Quality: Vision Models and Metrics*. Wiley, 2005.

[Wu and Rao, 2005] H.R. Wu and K.R. Rao. *Digital Video Image Quality and Perceptual Coding*. CRC Press, 2005.

[Yee *et al.*, 2001] H. Yee, S. Pattanaik, and D. P. Greenberg. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Transactions on Graphics*, 20(1):39–65, 2001.

[Yoshida *et al.*, 2008] A. Yoshida, M. Ihrke, R. Mantiuk, and H-P. Seidel. Brightness of the glare illusion. In *Proc. of Symposium on Applied perception in Graphics and Visualization (APGV)*, 2008.

[Zeng *et al.*, 2000] W. Zeng, S. Daly, and S. Lei. Visual optimization tools in JPEG 2000. In *Proc. of Inter. Conf. on Image Processing*, volume 2, pages 37 –40 vol.2, sept. 2000.

[Ziou and Tabbone, 1997] D. Ziou and S. Tabbone. Edge detection techniques - an overview. Technical report, International Journal of Pattern Recognition and Image Analysis, 1997.