Brno University of Technology, Faculty of Information Technology Department of Computer Graphics and Multimedia

HABILITATION THESIS

Computational Photography

Ing. Martin Čadík, Ph.D.

October 2014

Preface

This thesis presents thirteen research papers on computational photography which I have published, with my colleagues, since 2007. Three articles address *high dynamic range* (HDR) imaging. An analysis and evaluation of HDR tone mapping methods is shown, and two novel methods for image and video tone mapping are proposed. Two papers concern themselves with *color-to-grayscale image conversions*. A new adaptive perception-based conversion is presented, and a thorough evaluation of existing color-to-grayscale image conversions is performed. Five of the presented articles contribute to *image and video quality assessment*. A novel dynamic range independent video quality metric is proposed, and an HDR video dataset for validation of such metrics is published. New full-reference and a no-reference metrics for assessing image quality are proposed, and an analysis of state-of-the-art image metrics is performed. Finally, three papers enable *advanced image editing*. The concept of visually significant edges is advocated and implemented, and a new approach to multiscale image contrast editing is proposed. A newly developed automatic photo-to-terrain registration method makes advanced model-based image enhancements and manipulations possible.

Eight of the presented papers were published in impacted international journals including ACM Transactions on Graphics and Computer Graphics Forum, and one paper was accepted to the prestigious oral track at the Computer Vision and Pattern Recognition conference. The other articles were published at established conferences in the field.

The thesis contains an introductory part, followed by a brief overview of the contributions presented in the articles, and it concludes with possible avenues for future work. Reprints of the mentioned papers are enclosed in appendices.

Brno, October 6th, 2014 Martin Čadík

Acknowledgments

First of all, I would like to thank my colleagues who contributed to the papers presented in this thesis (in alphabetical order): Alessandro Artusi, Tunç O. Aydın, Lionel Baboud, Elmar Eisemann, Robert Herzog, Kwang In Kim, Rafał Mantiuk, Radosław Mantiuk, Karol Myszkowski, Antal Nemcics, László Neumann, Makoto Okabe, Dawid Pająk, Hans-Peter Seidel, and Michael Wimmer. I would like to express my special appreciation to Karol Myszkowski and Hans-Peter Seidel, senior researcher and head of the Computer Graphics Group at Max-Planck Institute Informatik Saarbrücken, respectively, for creating an extremely stimulating environment for research and education, and for sharing their invaluable experience. I would also like to thank to Jiří Žára and Pavel Slavík for leading the Department of Computer Graphics and Interaction at the CTU in Prague and for keeping the CG lab running. I am indebted to people from the CG lab at TU Vienna, namely Werner Purgathofer, Michael Wimmer, Alessandro Artusi, and Attila Neumann, for being great hosts during my research visits to Vienna. Thanks also go to Rafael García and László Neumann for their courteousness during my visits at the UdG in Girona. I also owe thanks to Honza Černocký, Adam Herout, and Pavel Zemčík for running the Department of Computer Graphics and Multimedia at Brno University of Technology and for their encouragement to finalize this thesis. I should not forget to mention Lionel Baboud, Robert Herzog, and Rafał Mantiuk, who besides being co-authors of my papers are truly great people and excellent friends. Many other people influenced my work. I wish to thank all my colleagues and friends I met in research groups in Brno, Prague, Pilsen, Vienna, Saarbrücken, Dresden, Zürich, Bangor, Girona, and Rennes for their ideas and comments, and for their friendship. Thanks to Martin Kolář, Pavel Dostal, and Ronan Boitard for proofreading the manuscript. Finally, my greatest thanks go to my family and to my girlfriend Jana, without whose infinite patience the work presented in this thesis would never have been finished, and thanks to God for all the mistakes I have made so far.

Parts of the work presented in this thesis have been supported by the MSMT CZ, res. prog. No.: MSM-212300014, MSM-6840770014 (Research in the area of information technologies and communications) and LC-06008 (Center for Computer Graphics), by the CTU in Prague, grant No. CTU-0408813, by the Aktion Kontakt OE/CZ grant No. 48p11, by European Cooperation in Science and Technology EU RTD (ECOST-STSM-IC1005-100312-015783), and by SoMoPro II grant (financial contribution from the EU 7 FP People Programme Marie Curie Actions, REA 291782, and from the South Moravian Region).

Contents

Chapter 1

Introduction

If we knew what it was we were doing,
it would not be called research, would it?
Albert Einstein

Visual communication is ubiquitous in today's world. With the advent of smart cell phones and hand-held devices equipped with integrated cameras, today virtually everyone is a photographer. Every day, we are taking photographs in larger quantities and often of higher technical qualities than ever before. We share our photos, edit them, search them, archive them, enhance them, capture them for some specific purpose, or we simply want to make our shots look nice.

Current digital cameras almost completely surpass traditional "chemical" photography. They do not only capture light, they in fact *compute* pictures [Haye08]. That said, there is practically no image that would not be computationally processed to some extent today. Visual computing is ubiquitous. Unfortunately, images taken by amateur photographers often lack the qualities of professional photos and some image editing is necessary. The main topic of this thesis is *computational photography* (CP), see Figure 1.1, which develops methods to enhance or extend the capabilities of the current digital imaging chain.

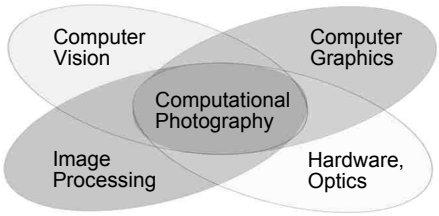


Figure 1.1: Computational photography methods use results of computer graphics, image processing, computer vision, and optics to enhance or extend the capabilities of current photography.

Image development and image processing have been extremely tedious in the past. The dark-

room image processing based on photo-chemistry of silver halide crystals came with frequent trial and error processes, and required a lot of experience. Professional photographers developed substantial skill in print processing. An example of such processes is dodging and burning [Adam05], a completely manual and analog way to manipulate image brightness and contrasts locally. Accordingly, in the past, image processing was devoted to a limited number of dedicated professional photographers.

Digital image processing took the editing processes out from the darkness of the darkroom to modern office desktops. The alterations of photographs are instantly visible and often reversible. However, the use of current photo-editing software, like GIMP or Photoshop [GIMP, Phot], still remains to a large extent the domain of experts. The path to an image to one's liking is still long, tedious and often manual. The aim of computational photography is to equip professionals with handy tools to make this process quicker, smoother, and more intuitive.

On the other hand, amateur photographers and ordinary users require powerful and easy to use solutions that can be used immediately, without any previous knowledge or training. In many cases, the user does not even need to know that some processing is going on. This is the domain of semi or fully automatic computational photography methods, which often involve advanced computations to make images look more realistic, natural, or simply nicer.

As the number of pictures taken is often quite high and their quality varies significantly, it is very important to be able to estimate it automatically. An indispensable area of computational photography research is therefore quality assessment, where automated metrics for predicting perceptual qualities of images and videos are developed. Furthermore, since the number of proposed CP methods grows quickly, automatic evaluations of algorithms, and experimentation in general, gains in importance. Particularly important are experiments involving real observers, because many CP methods aim to mimic human vision and to reproduce the perception of human observers.

The goal of computational photography is also to narrow down, or even to completely eliminate limitations of existing cameras. New approaches which use computational power to alleviate constraints imposed by physics are emerging every day. Moreover, many techniques dealing with such limitations are already widely known, e.g. extending limited dynamic range, widening field of view, extending depth of field, or image refocusing. Last but not least, computational photography seeks out completely new applications and novel approaches to imaging.

This thesis presents several contributions to the field of computational photography research. First, a summary of work on high dynamic range image and video processing (Chapter 2) is presented, followed by the efforts on color-to-grayscale conversions (Chapter 3). Then, image and video quality assessment methods (Chapter 4) are described, and new advanced image editing methods for automatic, as well as for manual image enhancements (Chapter 5), are shown. The thesis concludes with prospects of future research.

Chapter 2

HDR Image Processing

There is a crack in everything.
That's how the light gets in.
Leonard Cohen

One of the most developed areas of computational photography is high dynamic range imaging (HDR), which is concerned with overcoming the limited dynamic range of a sensor or a display device. A number of solutions have been published on HDR capture, storage, and reproduction [Rein10], so the complete HDR image chain is available today, see Figure 2.1. However, HDR displays are still rare and expensive, while printouts of photos are, and will be, a popular medium in the future. Therefore, most of the work in this research was devoted to *tone mapping methods* (TM), which aim to reproduce HDR images and videos on ordinary (low dynamic range, LDR) display devices. The main problem of HDR tone mapping resides in the fact that an HDR image can comprise a vast range of luminances, typically the whole range of a real-world scene, which ordinary (LDR) devices cannot reproduce.

Tone mapping methods (sometimes called tone mapping operators, TMO) convert an HDR image to an ordinary image while reducing (transforming) its dynamic range, see Figure 2.2. The goal of many tone mapping methods is to perform the transformation in a way that corresponds to human perception of the scene captured by the HDR image. Accordingly, many tone mapping methods mimic the behavior of the human visual system (HVS). One can classify existing tone mapping approaches according to the performed transformation into two main categories: *Global tone mapping methods* apply the tone reproduction curve (TRC), i.e. a transfer function. Therefore, they transform a particular value of the input luminance to one specific output value. *Local tone mapping methods* (TMOs) may on the other hand reproduce a particular input luminance to different output values, depending on the surrounding pixels.

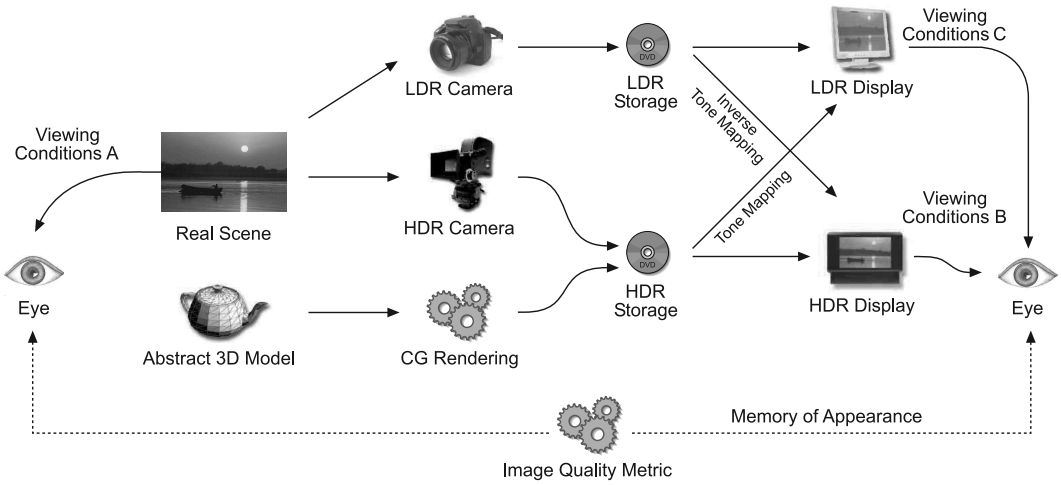


Figure 2.1: Traditional and high dynamic range image and video processing pipelines [Mysz08].

2.1 Evaluation of HDR Tone Mapping Methods Using Essential Perceptual Attributes

Several dozens of tone mapping methods have been proposed over the last decade [Rein10]. However, since their merits and shortcomings were not immediately clear, their experimental validation and evaluation was urgently needed. Consequently, a study was conducted on the effects of basic *image attributes for HDR tone mapping*, and a survey was made of how different methods reproduce these attributes [Cadi08b] (Appendix A). A scheme was proposed to detail the *relationships between essential image attributes*, leading to the concept of an overall image quality measure for HDR content. Two different subjective psychophysical experiments were performed, and the results showed that the proposed relationship between image attributes correlates with the choice or preference of the human observer. Finally, an *evaluation of fourteen existing tone mapping methods* was presented, with regard to these image attributes.

An interesting and important outcome of the two conducted experiments is that almost all of the

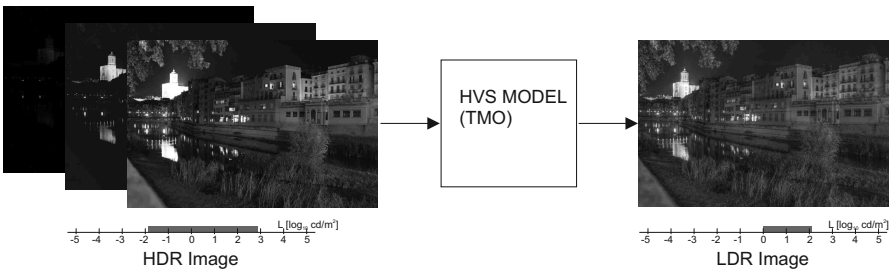


Figure 2.2: High dynamic range tone mapping (TMO) often involves a model of human visual system (HVS).

studied image quality attributes can be evaluated without comparison to a real HDR reference. With 14 tested methods and three real-world HDR scenes, the presented study was one of the most comprehensive evaluations of tone mapping methods. Another important result of this study is the finding that the proper global part of a tone mapping method is essential to obtain good perceptual results for typical real world scenes.

2.2 Perception Motivated Hybrid Approach to Tone Mapping

The results of several subjective experiments, which are summarized in Section 2.1, have shown high importance of preservation of global image attributes. Motivated by these findings, a concept of a simple yet powerful *general hybrid approach to tone mapping* [Cadi07] (Appendix B) has been proposed. In this approach, outputs of arbitrary global and local tone mapping methods are combined as follows: the global method is applied in order to reproduce overall image attributes correctly. Simultaneously, an enhancement map is constructed to guide a local operator to the critical areas of an image that deserve enhancement. Instead of inventing a new and complex TM method, we propose a general framework that utilizes already known ideas and combines existing and potentially forthcoming methods to obtain perceptually justifiable results. Moreover, based on the choice of involved methods and on the manner of construction of the enhancement map, it was shown that this approach is adaptable, and can easily be tailored to miscellaneous goals of tone mapping.

Subsequently, Artusi et al. [Artu07] published the concept of a *selective tone mapper*, which relies on a model of visual attention to direct local TMOs to perceptually important parts of an image, while a global TM method is used for the remainder. Artusi et al. proposed a generic GPU-aware implementation that can utilize any existing GPU TM method. In practice, this is a nice implementation and a verification of the hybrid approach presented above: the authors utilize a Canny edge detection to construct the enhancement map and then, in accord with the hybrid approach, they apply a local TM method only to the identified important parts of the image. This work also resulted in a patent [Artu10].

2.3 Temporal Tone Mapping: Visual Maladaptation in Contrast Domain

Intense changes in illumination may cause loss in visual sensitivity, which is usually recovered over a period of time. In fact, in the context of highly variant and temporally changing real-world illumination, the human visual system (HVS) itself is virtually never fully adapted. Due to this *maladaptation*, the visibility of some scene regions is reduced, although they would otherwise be perfectly visible. Some tasks such as driving a car, piloting a ship or an airplane, etc., require quick reaction times and undiverted attention. This may be simulated in safe conditions by including the temporal HVS model to perceptually tone mapped HDR videos.

Unfortunately, unlike tone mapping HDR images, there are only a few *temporal tone mapping* methods available [Patt00, Iraw05, Boit14]. Additionally, existing methods either completely neglect maladaptation, or they simulate only extremely simplistic cases while ignoring most aspects of the HVS. An advanced HDR tone mapping, which renders an HDR video as seen by a maladapted eye, was presented [Pajk10a] (Appendix C). The course of adaptation over time is modeled by considering both neural mechanisms and pigments bleaching and regeneration. This framework operates in the multi-scale contrast domain and models supra-threshold effects like visual masking, while also accounting for contrast sensitivity and luminance (mal)adaptation.

To conclude, merits of high dynamic range imaging (HDRI) are currently widely recognized not only in photography, but also in computer graphics, computer vision, and other areas of digital imaging. Moreover, HDRI is becoming popular in interactive and real-time applications as well. Finally, data visualizations, digital cinema industry, computer games and other interactive applications gain new qualities thanks to HDRI.

Chapter 3

Color-to-Grayscale Image Conversions

All colors made me happy, even gray.
My eyes were such that literally they took photographs.
Vladimir Nabokov

Black-and-white photography has not lost any of its artistic appeal despite the wide availability of color imaging processes. Accordingly, the conversion from color to grayscale is an important piece of the computational photography puzzle. Furthermore, color images often have to be converted to grayscale for reproduction, or for subsequent processing. To that end, *color-to-grayscale conversions* basically perform a reduction of the three dimensional color data into one dimension, see Figure 3.1. The aim of color-to-grayscale conversions is usually to produce perceptually plausible grayscale results. Unfortunately, no analogous conversion is naturally present in the human visual system (HVS). However, one may measure perceptual differences between colors in subjective psychophysical experiments. Moreover, it is evident that some loss of information due to the conversion is inevitable. The other goal is therefore to reproduce maximum information from the original color image in the grayscale.

In conventional black-and-white photography, gray tones are determined by the spectral sensitivity of the emulsion, and can be modified in an active way by selecting different filters to enhance a specific part of the spectrum. This analog multispectral technique can now be emulated digitally, and this approach to the problem can be considered as a *global color-to-gray conversion* [Grun05]. Additionally, *local approaches* to color-to-grayscale conversion have been proposed [Gooc05, Rasc05]. These are conceptually similar to local HDR tone mapping methods (Section 2.1), because they aim at preserving color contrasts by introducing chrominance information locally into the luminance channel. Unfortunately, this may also lead to undesirable artifacts, and high computational complexity.

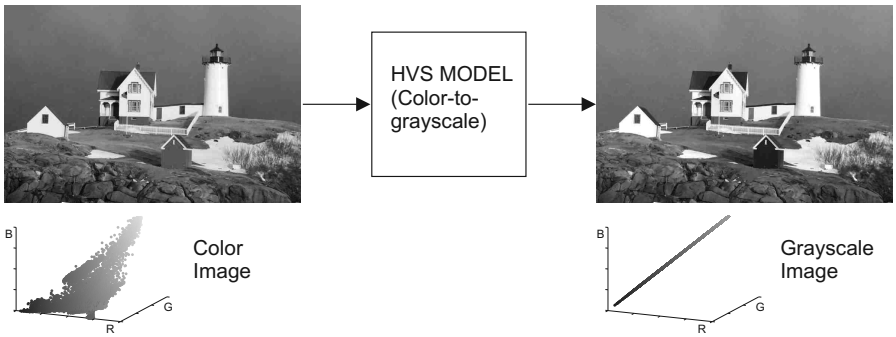


Figure 3.1: Color image to grayscale conversion often involves a model of human visual system (HVS).

3.1 Perception-based Color-to-Grayscale Conversion

Existing color-to-grayscale conversions [Gooc05, Rasc05] were eminently computationally intensive, which made them unsuitable for interactive conversions, especially with the ever increasing spatial resolution of digital images. To that end, a new *perceptually-plausible conversion* of color images to gray-scale, that operates in the gradient domain, was proposed [Neum07] (Appendix D). Two novel and efficient ways to construct a gradient field from a color image were developed. The first approach operates in the CIELab color space [Rein10], while the second uses the Coloroid color system [Nemc87], which is based on a strong experimental background. In this second approach, perceptually justifiable gray gradients equivalent to different color attributes are formulated and acquired by means of efficient experimental arrangements.

However, using one of these new approaches, one receives an inconsistent gradient field from a color image. As the inconsistent gradient field does not correspond to any real image, there is a problem of finding how to transform it to an output grayscale image. To solve the *gradient field inconsistency*, a new and efficient method was introduced. The algorithm converts inconsistent gradients into a consistent field which can be directly transformed into an image by a simple 2D integration. The complexity of this new method, called *gradient inconsistency correction*, is linear with respect to the number of pixels, making it suitable for high-resolution images. Experiments report that in comparison to existing approaches [Gooc05, Rasc05], the proposed method produces comparable results while being much faster to compute.

3.2 Perceptual Evaluation of Color-to-Grayscale Image Conversions

Many color-to-grayscale conversions have been proposed in the literature, however, their performance has not been objectively assessed yet. Accordingly, the strengths and weaknesses of these methods were unknown, and there was no standard testing dataset available. A thorough

evaluation of color-to-grayscale image conversions [Cadi08a] (Appendix E) was performed, and a testing dataset consisting of 24 color images was made publicly available. Two subjective experiments were conducted in which input images were converted to grayscale using 7 state-of-the-art conversions, and evaluated by 119 human subjects using paired comparison [Mant12]. This *new dataset*, which was made publicly available[1], is currently a widely recognized *de facto* standard for evaluating and testing of newly proposed color-to-grayscale conversions.

By looking at the color-to-grayscale evaluation results and at the evaluation of HDR tone mapping methods presented in Section 2.1, peculiar similarities become notable. If the aim is the natural reproduction of the original image, then simple global conversions perform efficiently on average. Advanced local techniques excel in certain cases, but may introduce unnatural artifacts, which reduce their robustness. The emotional or artistic dimension of color-to-grayscale conversion of photographs has not been explored yet, however, it is assumed that example-based style transfer techniques [Bae06, Aubr14] may be more efficient in such cases, in particular for amateur photographers. Professionals, on the other hand, often rely on simple manual conversions based on a weighted combination of color channels. The resulting grayscale image is, in this case, obtained by tweaking sliders which brighten or darken the respective tones in the photo. Accordingly, the main application of the methods presented above resides in fully automated color-to-grayscale conversions of images where it is critical to preserve color contrasts, e.g. for presentation or printing of business graphics and as a preprocessing step for further image editing. Finally, since color-to-grayscale conversion tightly couples the reproduction of brightness and contrast, it is usually combined with tonal modification methods, either automated (Section 2.1), or manual (Section 5.2).

[1] `http://cadik.posvete.cz/color_to_gray_evaluation/`

Chapter 4

Image and Video Quality Assessment

The goal of graphics is not to control light, but to control our perception of light.
Light is merely a carrier of the information we gather by perception.
Jack Tumblin, James A. Ferwerda

The goal of image and video quality assessment (IQA, VQA) is to computationally predict human perception of image and video quality. It is well known [Wang06, Wu05] that numerical distortion metrics, like root mean squared error (RMSE), are not adequate for the comparison of images, because they poorly predict the differences between the images as perceived by a human observer. To solve this problem properly, various perceptual *image and video quality metrics* (IQM, VQM) have been proposed [Wu05]. Image quality metrics traditionally comprise a computational human visual system (HVS) model to correctly predict image difference as a human would perceive it, be it a bottom-up [Mant11], or a top-down approach [Wang04]. Please refer to vision science textbooks [Palm02] for more in-depth information on human perception, and on HVS measurements related to masking, adaptation, contrast sensitivity, etc.

Image quality assessment is practical in various applications of computational photography. The main applications of IQA lie in the areas of image quality monitoring (e.g. in lossy image compression), benchmarking of imaging applications, and optimizing algorithms by tuning their parameter settings. Furthermore, image quality metrics have also been successfully applied to image database retrievals, evaluation of the perceptual impact of different computer graphics and vision algorithms, etc.

4.1 Dynamic Range Independent Video Quality Assessment

Full-reference image and video quality metrics are based on measuring the errors (signal differences) between a distorted image and the reference image, see Figure 4.1. The aim is to quantify the errors in a way that simulates human visual error sensitivity. A great variety of image quality metrics have been proposed in the literature [Wang06, Wu05]. Unfortunately, at

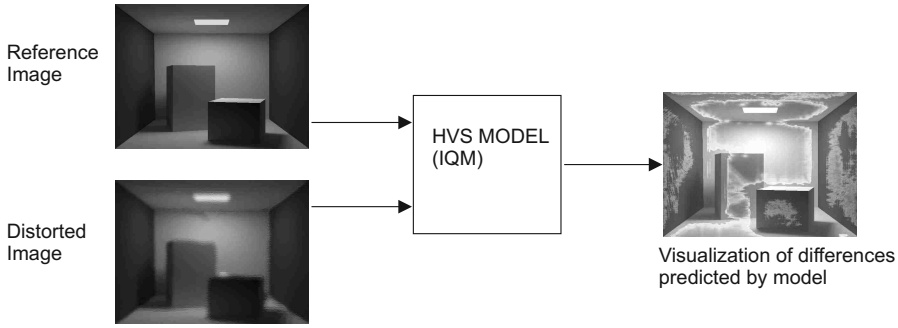


Figure 4.1: Full-reference image quality metrics (IQM) model human visual system (HVS) in a bottom-up or a top-down manner.

the time there was no available metric to compare HDR and LDR video sequences. Therefore, a *new dynamic range independent video quality metric* [Aydi10a] (Appendix F) was designed based on the computational bottom-up model of human perception. The implementation of this video metric is publicly available online[1] along with other IQMs, see Figure 4.2. According to the website statistics, this metric is regularly used by other researchers for evaluation of novel methods, which produce, enhance, or modify video sequences. To validate the metric and to foster future research in the field, a *new dynamic range independent dataset for evaluation of video quality metrics* was additionally published [Cadi11] (Appendix G).

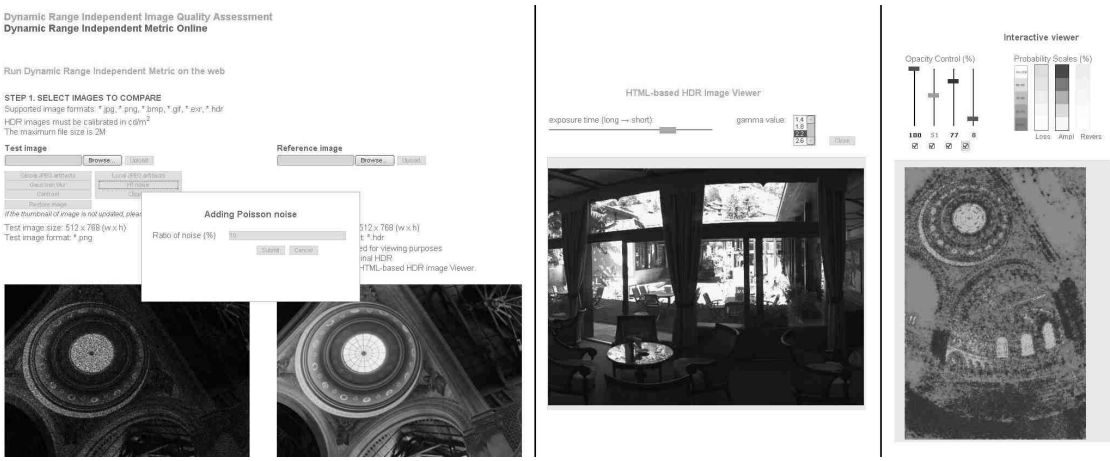


Figure 4.2: Image and video quality assessment online. Left: The user uploads the distorted and reference images or videos. Center: The HTML-based viewer enables viewing HDR images at various exposures and gamma values. Right: Visualization of the distortion map generated by the dynamic range independent metric.

[1] http://metrics.mpi-inf.mpg.de/

4.2 New Datasets and Evaluation of Image Quality Metrics

Even though knowledge about the human visual system (HVS) is continuously expanded, many unanswered questions and unverified hypotheses still remain. On that account, we are quite far from having an accurate bottom-up model of the HVS. Therefore, additionally to the bottom-up approaches shown above, top-down *data-driven approaches to image quality assessment* based on machine learning have been explored. Machine learning techniques have recently gained a lot of popularity and attention in many research areas. For such methods, it is of crucial importance to provide a sufficient amount of training data. Unfortunately, no usable dataset exhibiting localized distortion maps measured on human subjects is available thus far. Therefore, two experiments [Cadi12] were performed (Appendix H) where observers used a brush-painting interface to directly mark distorted image regions in the presence and absence of a high-quality reference image. The resulting *per-pixel image-quality datasets* enabled a thorough evaluation of existing full-reference IQMs. Furthermore, the dataset allowed to develop two new machine learning-based metrics, described hereafter.

4.3 Data-driven Full-Reference Metric for Synthetic Images

Although many image quality metrics have been developed in the past, they were often tuned for artifacts resulting from compression/transmission applications and have not been evaluated in the context of *synthetic computer generated image artifacts*. The unique datasets described above were utilized to develop a *Learning-based Predictor of Localized Distortions (LPLD)* [Cadi13] (Appendix J). LPLD is currently the best performing *full-reference metric* for synthetic images. The key element of the metric is a carefully designed set of features, which generalize over distortion types, image content, and superposition of multiple distortions in a single image. Additionally, two new datasets to validate this metric were created and made publicly available: a continuous range of basic distortions encapsulated in a few images, and the distortion saliency maps captured in the eye tracking experiment. The distortion maps are useful to benchmark existing and future IQMs and associated saliency maps could be used, for instance, in perceptual studies of human visual attention.

4.4 NoRM: No-Reference Image Quality Metric

So far, there was no available metric capable of predicting localized image distortions without knowing the original image, referred to as a *no-reference metric*, see Figure 4.3. To fill this gap in IQA, NoRM [Herz12] (Appendix I), a no-reference image quality metric for synthetic images was proposed. NoRM uses a supervised learning algorithm to predict a perceptual distortion map, which measures the probability of noticing the local distortions on the pixel-level. The proposed metric achieves prediction performance comparable to full-reference IQMs. The quality of the results of NoRM is owed to rendering-specific features extracted from the depth map

and the surface-material information.

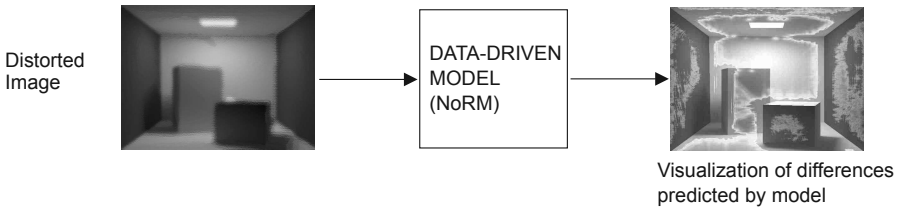


Figure 4.3: Data-driven no-reference image quality assessment metric (NoRM).

Despite many years of active research on image and video quality assessment, the developed metrics are often still far from being comparable to human observers. Existing universal metrics are not sufficiently robust to become widespread. However, to overcome this issue, one may develop specialized metrics tailored specifically to the particular problem. Recent examples of such metrics include the quality predictor for image completion [Kopf12], or similarity measure for illustration style [Garc14]. Furthermore, measuring vaguely defined quantities like interestingness of images [Gygl13] may be also feasible, perhaps thanks to the machine learning algorithms. Finally, the important area of multispectral image and video comparison [Le M14] remains currently almost unexplored.

Chapter 5

Advanced Image Editing

Every moment of searching is a moment of encounter.
Paulo Coelho

The majority of today's photographs are being altered after the capture: either manually, automatically, or both. Image editing methods are therefore the essential building blocks of the digital image processing chain. This section presents three different algorithms with broader applicability to image editing and computational photography. These algorithms have been tested on several use case scenarios. However, it is believed that they have numerous other applications.

More specifically, *visually significant edges* (Section 5.1) were shown to be beneficial in content-aware image resizing, tone mapping (Chapter 2), and HDR image stitching. The straightforward use of visually significant edges in image quality assessment (Chapter 4) remains to be explored.

The concept of *contrast prescription* (Section 5.2), on the other hand, facilitates the implementation of an interactive tool which may be utilized in photoediting applications. It enables convenient multiscale editing of image contrasts, and its wide applications are limited only by the creativity of the user (an artist, professional photographer, etc.). The use of contrast prescription in manual HDR image tone mapping (Chapter 2) is illustrated, and it is believed that it would be helpful in manual color-to-gray conversion (Chapter 3) as well.

Finally, an automatic *photo-to-3D model alignment* method (Section 5.3), which can register images into a 3D digital model, is presented. This enables a multitude of previously unseen or impossible applications such as adding haze or dehazing of photos, simulating changes in lighting, novel view synthesis, expanding the field of view, adding new objects into images or videos, and more.

5.1 Visually Significant Edges

Edge detection is a traditional problem in image processing and as such it is a building stone of many computational photography methods. Similarly, edge-aware image decompositions [Farb08, Fatt09] form the core of numerous applications, such as image abstraction, detail enhancement and HDR tone mapping. The result of the entire process is therefore critically dependent on the quality of edge computation. In *visually significant edges* [Aydi10b] (Appendix K), the way of determining the location and strength of edges is revised. Contrary to the widely used gradient magnitude-based edge strength model, the proposed algorithm accounts for visual significance of the edges, by modeling HVS mechanisms, such as luminance adaptation, spatial frequency sensitivity, and visual masking. The model was implemented using edge-aware image decomposition based on second generation wavelets. Benefits of the visually significant edges with respect to the gradient magnitude model in *image retargeting, HDR image stitching and tone mapping* have been demonstrated (Appendix K). In general, the visually significant edges provide qualitative improvements in applications utilizing edge strength at the cost of a modest computational burden due to the implemented HVS model.

5.2 Contrast Prescription for Multiscale Image Editing

This section presents a contribution to interactive image editing techniques, called *contrast prescription* [Pajk10b] (Appendix L). Contrast manipulation is a common process in digital photography. Recently developed multiscale image decompositions [Farb08, Fatt09] enable modifications of image contrasts at arbitrary scales. An important, but often ignored property of multiscale frameworks is the interaction between contrast at individual scales. Indeed, contrast modification in one band affects contrast in other bands, which is not intuitive for the user.

The concept of contrast prescription enables the user to lock the contrast in selected areas and bands, and make it immune to contrast manipulations in other bands. Additionally, an extension that allows the user to perform *countershading*, or halo editing was introduced. Countershading has been used by painters for centuries as it may enhance the *perceived contrast*. This approach is one of the few allowing control over such behavior. The hardware accelerated (GPU) implementation, combined with an intuitive user interface, provides real-time feedback to the user, which renders the proposed tool yet handier.

5.3 Automatic Photo-to-Terrain Alignment

Having a sufficiently accurate match between a photograph and a 3D model offers new possibilities for image enhancement. The goal of the proposed *automatic photo-to-terrain alignment* method [Babo11] (Appendix M) is to register outdoor pictures and movies into a Google-Earth-like 3D digital elevation model. Assuming the location of the photographer is known, the aim of the method is to accurately find the *orientation of the camera* used to capture the image. To this

extent, the orientation of the recording is calculated from an annotated 3D model by finding, in an efficient and robust way, the best match between significant edges (e.g., silhouettes) in both sources. While such an approach seems usually infeasible due to computational complexity, a careful mathematical reformulation allowed the solution to be practical.

Figure 5.1: Relighting of the photograph using the 3D model. Left: input image, middle and right: relighted results [Kopf08].

Such photo-to-terrain alignment could be used to transform photographs into a realistic virtual 3D experience. In particular, the system could be used to automatically highlight elements in the image, such as the travel path taken, names of mountains, or other landmarks. The work could further be used to augment the realism and level of detail of 3D applications by transferring information from images and movies directly onto the 3D model. Furthermore, the proposed photo-to-terrain alignment method produces a precise depth map of the queried photo, because the used digital elevation models are very accurate. Such a depth map, and the whole 3D model, can be utilized in many ways, as has been nicely illustrated by Kopf et al. for manually registered photographs [Kopf08]. Applications in photographs range from dehazing and relighting, to novel view synthesis, and overlaying with geographic information, see Figure 5.1. With the advent of Kinect-like depth sensors, such *model-based image enhancement and manipulation* approaches, besides their obvious strengths and benefits, will gain in importance.

Chapter 6

Conclusions and Future Work

Be the change that you wish to see in the world.
Mahatma Gandhi

This thesis has presented several contributions to computational photography (CP). It has covered a broad spectrum of CP methods, ranging from HDR image processing over the automatic visual quality assessment and color-to-grayscale conversions to model-based image and video enhancements. However, hand in hand with research progress, the area of computational photography is continuously expanding [Rask09, Szel10].

Work on presented publications revealed a number of possible directions for future research and development. Research in advanced visual attention models, in particular the top-down and task-driven ones, may have the potential to push the limits of current *image quality assessment*. Additionally, such models are likely to find many applications in related fields, including robotics and machine vision. The area of *color-to-grayscale* conversions requires robust temporal methods, as well as extensive experimentation and evaluations using video sequences. The technology of *high dynamic range imaging* is on the verge of being adopted by the market. However, the success of this technology depends to a large extent on the availability of HDR footage. Finally, the field of *advanced photo enhancement* awaits a boom of 3D model-assisted techniques [Khol14].

Research in the area of computational photography is currently very popular, and dozens of new methods are presented each year at major conferences on computer graphics, computer vision, and image processing. Presently, events specific to computational photography exist as well. In general, there are many unexplored research directions and accordingly, breakthrough ideas are continuously emerging [Shih13, Ito14, Laff14]. The general topic, on the other hand, is already quite well covered by textbooks [Rask09, Szel10], and is taught at leading technical universities all over the world. Furthermore, some traditional approaches to computational photography, e.g. panorama stitching, high dynamic range imaging, light fields or lumigraphs [Levo96, Gort96] have reached maturity and are being utilized by a number of people, and some of them are even

available on the market [Sphe, Lytr, Goog]. Not to mention basic algorithms like de-mosaicing or moiré suppression, which currently form an integral part of most of existing cameras.

It is expected that computational photography will become ubiquitous in the near future. Computational methods are, or will be present in hardware capture and display devices, internet services and applications, advertising, as well as many other aspects of everyday life. In addition, depictions that go beyond the capabilities of traditional imaging systems will be encountered increasingly often. Photography will be less and less limited by the properties of capture devices following examples of HDR technology (Section 2.1), lightfields [Lytr], and other achievements of CP. This progress will enable powerful digital editing, previously impossible modifications, as well as faithful reproductions of captured reality. However, despite a long history of photo fraud and manipulation [Fari08], people still consider photography a reliable tool of documentation. It will be interesting to observe how CP techniques will change the way people think of photographs. Accordingly, research in image forensics [Fari08] will also gain in importance.

Looking even further ahead, it is realistic to expect that augmented reality solutions (Section 5.3, [Meno14], [Goog], etc.) will become widespread, and mature enough to allow novel real-time applications. This will be supported by the progress of computer vision in image understanding, by computer graphics in acceleration of realistic image synthesis, and by advanced image processing techniques. The progress of vision science in understanding and modeling the human visual system and perception mechanisms is also essential, as well as technical advances in hardware capture devices and miniaturization.

Bibliography

[Adam05] A. Adams. *The Print*. Little, Brown, Bulfinch, 2005.

[Artu07] A. Artusi, B. Roch, Y. Chrysanthou, D. Michael, and A. Chalmers. Selective Tone Mapper. Tech. Rep. TR-07-05, University of Cyprus, 2007.

[Artu10] A. Artusi, B. Roch, A. Chalmers, and Y. Chrysanthou. US Patent, publication number GB2449272: Selective Tone Mapper. 2010. UK Patent Office.

[Aubr14] M. Aubry, S. Paris, S. W. Hasinoff, J. Kautz, and F. Durand. Fast Local Laplacian Filters: Theory and Applications. *ACM Transactions on Graphics*, Vol. 33, No. 5, pp. 167:1–167:14, Sep. 2014.

[Aydi10a] T. O. Aydın, M. Čadík, K. Myszkowski, and H.-P. Seidel. Video Quality Assessment for Computer Graphics Applications. *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia)*, Vol. 29, No. 6, pp. 161:1–161:12, Dec. 2010.

[Aydi10b] T. O. Aydın, M. Čadík, K. Myszkowski, and H.-P. Seidel. Visually significant edges. *ACM Transactions on Applied Perception*, Vol. 7, No. 4, pp. 27:1–27:15, 2010.

[Babo11] L. Baboud, M. Čadík, E. Eisemann, and H.-P. Seidel. Automatic Photo-to-terrain Alignment for the Annotation of Mountain Pictures. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR Orals)*, pp. 41–48, IEEE Computer Society, Washington, DC, USA, 2011.

[Bae06] S. Bae, S. Paris, and F. Durand. Two-scale Tone Management for Photographic Look. *ACM Transactions on Graphics*, Vol. 25, No. 3, pp. 637–645, July 2006.

[Boit14] R. Boitard, R. Cozot, D. Thoreau, and K. Bouatouch. Zonal brightness coherency for video tone mapping. *Signal Processing: Image Communication*, Vol. 29, No. 2, pp. 229–246, Feb. 2014.

[Cadi07] M. Čadík. Perception Motivated Hybrid Approach to Tone Mapping. In *Winter School of Computer Graphics (WSCG Full Papers)*, pp. 129– 136, Pilsen, Czech Republic, 2007.

[Cadi08a] M. Čadík. Perceptual Evaluation of Color-to-Grayscale Image Conversions. *Computer Graphics Forum*, Vol. 27, No. 7, pp. 1745–1754, 2008.

[Cadi08b] M. Čadík, M. Wimmer, L. Neumann, and A. Artusi. Evaluation of HDR Tone Mapping Methods Using Essential Perceptual Attributes. *Computers & Graphics*, Vol. 32, No. 3, pp. 330–349, 2008.

[Cadi11] M. Čadík, T. O. Aydın, K. Myszkowski, and H.-P. Seidel. On evaluation of video quality metrics: an HDR dataset for computer graphics applications. *Human Vision and Electronic Imaging XVI*, Vol. 7865, No. 1, 2011.

[Cadi12] M. Čadík, R. Herzog, R. Mantiuk, K. Myszkowski, and H.-P. Seidel. New Measurements Reveal Weaknesses of Image Quality Metrics in Evaluating Graphics Artifacts. *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia)*, Vol. 31, No. 6, pp. 147:1–147:10, 2012.

[Cadi13] M. Čadík, R. Herzog, R. Mantiuk, R. Mantiuk, K. Myszkowski, and H. Seidel. Learning to Predict Localized Distortions in Rendered Images. *Computer Graphics Forum*, Vol. 32, No. 7, pp. 401–410, 2013.

[Farb08] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski. Edge-Preserving Decompositions for Multi-Scale Tone and Detail Manipulation. *ACM Transactions on Graphics*, Vol. 27, No. 3, Aug. 2008.

[Fari08] H. Farid. Digital Image Forensics. *Scientific American*, Vol. 298, No. 6, pp. 66–71, 2008.

[Fatt09] R. Fattal. Edge-avoiding wavelets and their applications. *ACM Transactions on Graphics*, Vol. 28, No. 3, pp. 1–10, 2009.

[Garc14] E. Garces, A. Agarwala, D. Gutierrez, and A. Hertzmann. A Similarity Measure for Illustration Style. *ACM Transactions on Graphics*, Vol. 33, No. 4, pp. 93:1–93:9, July 2014.

[GIMP] GIMP GNU Image Manipulation Program. GNOME Foundation, GNU Project, `http://www.gimp.org/`. Accessed: Jul, 2014.

[Gooc05] A. A. Gooch, S. C. Olsen, J. Tumblin, and B. Gooch. Color2Gray: salience-preserving color removal. *ACM Transactions on Graphics*, Vol. 24, No. 3, pp. 634–639, 2005.

[Goog] Google Glass. Google, Inc., `http://www.google.com/glass/`. Accessed: Jul, 2014.

[Gort96] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The Lumigraph. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 43–54, ACM, New York, NY, USA, 1996.

[Grun05] M. Grundland and N. A. Dodgson. The Decolorize Algorithm for Contrast Enhancing, Color to Grayscale Conversion. Tech. Rep. UCAM-CL-TR-649, University of Cambridge, 2005.

[Gygl13] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool. The Interestingness of Images. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[Haye08] B. Hayes. Computational Photography. *American Scientist*, Vol. 96, No. 2, pp. 94–98, 2008.

[Herz12] R. Herzog, M. Čadík, T. O. Aydın, K. I. Kim, K. Myszkowski, and H.-P. Seidel. NoRM: No-Reference Image Quality Metric for Realistic Image Synthesis. *Computer Graphics Forum*, Vol. 31, No. 2, pp. 545–554, 2012.

[Iraw05] P. Irawan, J. A. Ferwerda, and S. R. Marschner. Perceptually Based Tone Mapping of High Dynamic Range Image Streams. In K. Bala and P. Dutre, Eds., *Rendering Techniques 2005 (Eurographics Symposium on Rendering)*, pp. 231–242, Eurographics Association, Konstanz, Germany, June 2005.

[Ito14] A. Ito, S. Tambe, K. Mitra, A. C. Sankaranarayanan, and A. Veeraraghavan. Compressive Epsilon Photography for Post-capture Control in Digital Imaging. *ACM Transactions on Graphics*, Vol. 33, No. 4, pp. 88:1–88:12, July 2014.

[Khol14] N. Kholgade, T. Simon, A. Efros, and Y. Sheikh. 3D Object Manipulation in a Single Photograph Using Stock 3D Models. *ACM Transactions on Graphics*, Vol. 33, No. 4, pp. 127:1–127:12, July 2014.

[Kopf08] J. Kopf, B. Neubert, B. Chen, M. F. Cohen, D. Cohen-Or, O. Deussen, M. Uyttendaele, and D. Lischinski. Deep Photo: Model-Based Photograph Enhancement and Viewing. *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia 2008)*, Vol. 27, No. 5, pp. 116:1–116:10, 2008.

[Kopf12] J. Kopf, W. Kienzle, S. Drucker, and S. B. Kang. Quality Prediction for Image Completion. *ACM Transactions on Graphics*, Vol. 31, No. 6, pp. 131:1–131:8, Nov. 2012.

[Laff14] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays. Transient Attributes for High-level Understanding and Editing of Outdoor Scenes. *ACM Transactions on Graphics*, Vol. 33, No. 4, pp. 149:1–149:11, July 2014.

[Le M14] S. Le Moan and P. Urban. Image-Difference Prediction: From Color to Spectral. *Image Processing, IEEE Transactions on*, Vol. 23, No. 5, pp. 2058–2068, May 2014.

[Levo96] M. Levoy and P. Hanrahan. Light Field Rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 31–42, ACM, New York, NY, USA, 1996.

[Lytr] Lytro Illum. Lytro, Inc., `https://www.lytro.com/`. Accessed: Jun, 2014.

[Mant11] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich. HDR-VDP-2: A Calibrated Visual Metric for Visibility and Quality Predictions in All Luminance Conditions. *ACM Transactions on Graphics*, Vol. 30, No. 4, pp. 40:1–40:14, July 2011.

[Mant12] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk. Comparison of Four Subjective Methods for Image Quality Assessment. *Computer Graphics Forum*, Vol. 31, No. 8, pp. 2478–2491, Dec. 2012.

[Meno14] A. Menozzi, B. Clipp, E. Wenger, J. Heinly, E. Dunn, H. Towles, J.-M. Frahm, and G. Welch. Development of vision-aided navigation for a wearable outdoor augmented reality system. In *Position, Location and Navigation Symposium - PLANS 2014, 2014 IEEE/ION*, pp. 460–472, May 2014.

[Mysz08] K. Myszkowski, R. Mantiuk, and G. Krawczyk. *High Dynamic Range Video*. Morgan & Claypool, 2008.

[Nemc87] A. Nemcsics. Color space of the Coloroid color system. *Color Research and Application*, Vol. 12, No. 3, pp. 135–146, 1987.

[Neum07] L. Neumann, M. Čadík, and A. Nemcsics. An Efficient Perception-Based Adaptive Color to Gray Transformation. In *Proceedings of Computational Aesthetics 2007*, pp. 73– 80, Eurographics Association, Banff, Canada, 2007.

[Pajk10a] D. Pająk, M. Čadík, T. O. Aydın, K. Myszkowski, and H.-P. Seidel. Visual maladaptation in contrast domain. *Human Vision and Electronic Imaging XV*, Vol. 7527, No. 1, p. 752710, 2010.

[Pajk10b] D. Pająk, M. Čadík, T. O. Aydın, M. Okabe, K. Myszkowski, and H.-P. Seidel. Contrast Prescription for Multiscale Image Editing. *The Visual Computer Journal*, Vol. 26, No. 6-8, pp. 739–748, June 2010.

[Palm02] S. E. Palmer. *Vision science – photons to phenomenology*. The MIT Press, Cambridge, 3rd Ed., 2002.

[Patt00] S. N. Pattanaik, J. Tumblin, H. Yee, and D. P. Greenberg. Time-dependent Visual Adaptation for Fast Realistic Image Display. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 47–54, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 2000.

[Phot] Photoshop Creative Cloud. Adobe Systems Incorporated, `http://www.adobe.com/en/products/photoshop.html`. Accessed: Aug, 2014.

[Rasc05] K. Rasche, R. Geist, and J. Westall. Re-coloring Images for Gamuts of Lower Dimension. *Computer Graphics Forum*, Vol. 24, No. 3, pp. 423–432, 2005.

[Rask09] R. Raskar and J. Tumblin. *Computational Photography: Mastering New Techniques for Lenses, Lighting, and Sensors*. A. K. Peters, Ltd., Natick, MA, USA, 2009.

[Rein10] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski. *High Dynamic Range Imaging, Second Edition: Acquisition, Display, and Image-Based Lighting*. Morgan Kaufmann, 2nd Ed., June 2010.

[Shih13] Y. Shih, S. Paris, F. Durand, and W. T. Freeman. Data-driven Hallucination of Different Times of Day from a Single Outdoor Photo. *ACM Transactions on Graphics*, Vol. 32, No. 6, pp. 200:1–200:11, Nov. 2013.

[Sphe] Spheron VR. Spheron-VR AG, `https://www.spheron.com/`. Accessed: Jul, 2014.

[Szel10] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag New York, Inc., New York, NY, USA, 1st Ed., 2010.

[Wang04] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, Vol. 13, No. 4, pp. 600–612, 2004.

[Wang06] Z. Wang and A. C. Bovik. Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing*, Vol. 2, No. 1, pp. 1–156, 2006.

[Wu05] H. R. Wu and K. R. Rao. *Digital Video Image Quality and Perceptual Coding (Signal Processing and Communications)*. CRC Press, Inc., Boca Raton, FL, USA, 2005.

Appendices – Paper Reprints

Appendix A

Evaluation of HDR Tone Mapping Methods Using Essential Perceptual Attributes

M. Čadík, M. Wimmer, L. Neumann, and A. Artusi. Evaluation of HDR Tone Mapping Methods Using Essential Perceptual Attributes. *Computers & Graphics*, Vol. 32, No. 3, pp. 330–349, 2008.
IF=1.029

Computers & Graphics 32 (2008) 330–349

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Technical Section

Evaluation of HDR tone mapping methods using essential perceptual attributes

Martin Čadík [a,*], Michael Wimmer [b], Laszlo Neumann [c], Alessandro Artusi [d]

[a] Department of Computer Science and Engineering, CTU in Prague, Karlovo nam. 13, 121 35 Prague, Czech Republic
[b] Institute of Computer Graphics and Algorithms, Vienna University of Technology, Austria
[c] Computer Vision and Robotics Group, University of Girona, and ICREA, Spain
[d] Warwick Digital Laboratory, University of Warwick, UK

ARTICLE INFO

Article history:
Received 18 June 2007
Received in revised form
7 April 2008
Accepted 10 April 2008

PACS:
07.05.Pj
07.05.Rm
07.68.+m

Keywords:
High dynamic range
Tone mapping
Image attributes
Visual perception
Psychophysics
Subjective testing
Evaluation of methods

ABSTRACT

The problem of reproducing high dynamic range images on media with restricted dynamic range has gained a lot of interest in the computer graphics community. There exist various approaches to this issue, which span several research areas including computer graphics, image processing, color vision, physiological aspects, etc. These approaches assume a thorough knowledge of both the objective and subjective attributes of an image. However, no comprehensive overview and analysis of such attributes has been published so far.

In this contribution, we present an overview about the effects of basic image attributes in high dynamic range tone mapping. Furthermore, we propose a scheme of relationships between these attributes, leading to the definition of an overall image quality measure. We present results of subjective psychophysical experiments that we have performed to prove the proposed relationship scheme. Moreover, we also present an evaluation of existing tone mapping methods (operators) with regard to these attributes. Finally, the execution of with reference and without a real reference perceptual experiments gave us the opportunity to relate the obtained subjective results.

Our effort is not just useful to get into the tone mapping field or when implementing a tone mapping method, but it also sets the stage for well-founded quality comparisons between tone mapping methods. By providing good definitions of the different attributes, user-driven or fully automatic comparisons are made possible.

1. Introduction

The dynamic range of visual stimuli in the real world is extremely large. A high dynamic range (HDR) image can be generated either synthetically or acquired from the real world, but the conventional media used to present these images can only display a limited range of luminous intensity. This problem, i.e., displaying high contrast images on output devices with limited contrast, is the task of HDR *imaging*, and it is approached by HDR tone mapping (TM). A number of different TM methods (operators) have been proposed in history [1,2]. However, also due to their sheer number, the advantages and disadvantages of these methods are not immanently clear, and therefore a thorough and systematic comparison is highly desirable.

The field of TM assumes extensive knowledge of findings from various scientific areas. In order to conduct a comparison of TM methods, it is necessary to settle upon a set of *image attributes* by which the images produced by the methods should be judged. These attributes are not independent, and their interrelationships and the influence on the overall image quality need to be carefully analyzed. This is useful not just for comparing existing HDR approaches, but for evaluating *future ones* as well. The human visual system (HVS) is extremely complex and, besides highly focused laboratory studies, there is a lack of comprehensive user experiments we could build on.

In this contribution, we give a comprehensive list of most of the important attributes involved in the evaluation of a TM method, and we show which relationships exist between the basic attributes by means of two different subjective testing methods. Namely, we investigate the perceived quality of the images produced by particular TM methods with and without the possibility of direct comparison to the original real-world scenes. The evaluation of the attributes and their relationships leads to the definition of an *overall image quality* (OIQ). This metric can be used to judge how well a given TM method is able to produce naturally looking images. Furthermore, we present the most

* Corresponding author. Tel.: +420 737 049 097; fax: +420 224 923 325.
E-mail address: cadikm@fel.cvut.cz (M. Čadík).

0097-8493/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cag.2008.04.003

comprehensive comparison to date in terms of the number of TM methods considered, including 14 different methods.

The article is organized as follows. In Section 2, we overview the previous work on comparison of TM methods and other related work. In Section 3, we introduce and describe the term "overall image quality". In Section 4, we give a survey of the most important image attributes for TM, and we describe how different methods reproduce these attributes. In Section 5 we propose a new scheme of relationships between the image attributes. In Section 6 we describe the two applied experimental methods based on human observations, and finally in Section 7, we show and discuss the results of these experiments. The survey of image attributes and the relationships (Sections 4, 5) is extended from [3] and incorporates our new findings.

2. Previous work

The history of evaluation of TM methods is short. The following works (the only ones, to our best knowledge) were published only in the last few years. This is due to the recent increase in published TM methods on the one hand, and due to the very high time, implementation, human and other demands involved in such an evaluation on the other hand. While this section surveys the previous work, we relate our results to these works in Section 7.5.

2.1. Experimental evaluations of TM methods

Drago et al. [4] performed a perceptual evaluation of six TM methods with regard to similarity and preference. In their study, observers were asked to rate a difference for all pairwise comparisons of a set of four HDR images tone mapped with six TM methods (24 images in total) shown on the screen. A multidimensional perceptual scaling of the subjective data from 11 observers revealed the two most salient stimulus space dimensions. The authors unfolded these dimensions as naturalness and detail and also identified the ideal preference point in the stimulus space. These findings were then used for a final ranking of the six TM methods.

In 2005, Yoshida et al. [5] compared seven TM methods using two real-world architectural interior scenes. The 14 observers were asked to rate basic image attributes (contrast, brightness, details) as well as the naturalness of the images. The results of this perceptual study exhibited differences between global and local TM methods. Global methods performed better than local methods in the reproduction of brightness and contrast; however, local methods exhibited better reproduction of details in bright regions of images.

Kuang et al. [6] tested eight TM algorithms using 10 HDR images. The authors implemented two paired comparison psychophysical experiments assessing the color and gray scale TM performance, respectively. In these tests, 30 observers were asked to choose the preferred image for each possible pair. The results showed the consistency of TM performance for gray scale and color images. In the continuation of this research, Kuang et al. [7] removed two TM methods and added two new images to the group of input stimuli. The authors examined the overall image preference (using paired comparison performed on an LCD desktop monitor) and preferences for six image attributes (using a rating scale)—highlight details, shadow details, overall contrast, sharpness, colorfulness, artifacts. The results show that shadow details, overall contrast, sharpness and colorfulness have high correlations with the overall preference. More recently and parallel to our work, Kuang et al. [8] used three indoor scenes and 19 subjects to evaluate seven TM algorithms. Using two paired comparisons, the authors evaluated image contrast, colorfulness and overall accuracy. The results showed that bilateral filtering [9] generated more accurate results than other algorithms. Results of the three experiments performed by Kuang and colleagues are summarized in [10].

Ashikhmin and Goyal [11], parallel to our work, demonstrated that using real environments is crucial in judging performance of TM methods. The authors compared five TM methods using four real-world indoor environments plus two additional HDR images. Fifteen subjects were involved in three ranking experiments: first two tests (preference and fidelity) were performed without ground truth while the third (fidelity) was conducted with reference (real scene). The results indicate that there is statistically no difference between preference and fidelity when there is no reference (i.e., equivalence of liking and naturalness criteria). However, the results show a difference in subject's responses for the fidelity test with reference and without reference.

2.2. Evaluations using HDR displays

Ledda et al. [12] ran an evaluation of six TM methods by comparing to the reference scenes displayed on an HDR display. This HDR display allowed authors to involve many (23) input scenes. Subjects were presented three images at once (the reference and two tone mapped images) and had to choose the image closest to the reference. Statistical methods were used to process subjective data and the six examined methods were evaluated with respect to the overall quality and to the reproduction of features and details.

In the field of HDR displays, Yoshida et al. [13] analyzed the reproduction of HDR images on displays of varying dynamic range. The authors ran two perceptual experiments to measure subjective preferences and the perception of fidelity of real scenes. Twenty-four participants, 25 HDR images and three real-world scenes were involved in the experiments. An outcome of this work is the analysis how users adjust parameters of a generic global TM method to achieve the best looking images and the images that are closest to the real-world scenes.

Akyüz et al. [14] investigated how LDR images are best displayed on current HDR monitors. In two subjective experiments, authors exhibited 10 HDR images to 22 and 16 subjects, respectively. The results show that HDR displays outperform LDR ones and that LDR data do not require sophisticated treatment to produce a HDR experience. More surprisingly, results show that tone mapped HDR images are statistically no better than the best single LDR exposure.

2.3. Other related studies

Some exciting contributions were published in the domain of image quality measurement of ordinary LDR images (see the book by Janssen [15] for an overview on this topic). Rogowitz et al. [16] conducted two psychophysical scaling experiments for the evaluation of image similarity. The subjective results were compared to two algorithmic image similarity metrics and analyzed using multidimensional scaling. The analysis showed that humans use many dimensions in their evaluations of image similarity, including overall color appearance, semantic information, etc.

We find related work also in the field of psychophysical color research and photography, e.g., Fedorovskaya et al. [17] varied chroma of four input images to determine its effect on perceived image quality, colorfulness and naturalness. Results indicate that the enhancement of colorfulness leads to higher perceptual quality of an image. Savakis et al. [18] performed an experiment

on image appeal in consumer photography. While image quality is generally an objective measure, image appeal is rather subjective. During the experiment, authors showed 30 groups of prints to 11 people. The task of each subject was to select such a picture from each group that would receive the most attention in a photo album. Moreover, subjects had to comment the positive and negative attributes they used for the selection of the picture. The results show that the most important attributes for image appeal fall into the groups of composition/subject and people/expression, leaving objective attributes less significant.

Jobson et al. [19] investigated contrast and lightness in visually optimized LDR images. The authors approach the lightness as the image mean and the contrast as the mean of regional standard deviations. Inspecting these measures, the authors experimentally show that visually optimized LDR images are clustered about a single mean value and have high standard deviations, i.e., both the lightness and contrast are improved with the latter being more affected.

In a forthcoming paper, Mantiuk and Seidel [20] show an application of their generic (black-box) TM operator to the analysis of TM methods. The authors fit the generic operator to 12 TM methods to visualize their characteristics using fitted parameters of the generic operator. Moreover, they apply the generic operator to HDR image compression. It is interesting to observe that global TM methods result in less distorted reconstruction than local ones, even though one would favor local methods to preserve more information.

2.4. Our approach

Differently from the mentioned approaches, we adopt both a direct rating (with reference) comparison of the tone mapped images to the real scenes, and a subjective ranking of tone mapped images without a real reference. This enables us to confront the results from these two subjective experiments. Moreover, we present a methodology for evaluating TM methods using generally known image attributes. With 14 methods in total, and three typical real-world HDR scenes, the subjective studies carried out to confirm this methodology also contain one of the most comprehensive comparison of TM methods. We have already presented [3] preliminary ideas of this project and we conducted an initial pilot study to examine the experimental setup. It was observed that the overall image quality is not determined by a single attribute, but rather a composition of them. Next, we assessed [21] the results concerning the indoor scenes. Encouraged by these findings, we conducted a full experiment (we extended the input stimuli group by two additional, different outdoor scenes), the results of which, including a thorough discussion, new statistical methodology, etc. are presented in this contribution.

3. Overall image quality

In this section, we motivate and describe a measure which is useful for determining the performance of a particular TM method.

The first question is whether it is possible at all to find an optimal or "exact" method to tone map an arbitrary HDR input image, based on human vision. Unfortunately, the answer seems to be negative. Take for example a beach scene, where the absolute illuminance is often above 50,000 lux. A captured photograph of that scene, viewed under normal room illumination (about 200 lux), can never reproduce the same amount of colorfulness, because this is a *psychophysiological* effect that depends on the absolute illuminance (vivid colors start to be perceived above 2000 lux). Therefore, a natural reproduction is only possible to a limited degree.

Another important question is the intent of the reproduction. The classical *perceptual* approach tries to simulate the human vision process and design the TM method accordingly. For example, a scene viewed at night would be represented blurred and nearly monochromatic due to scotopic vision. However, if it is important to understand some fine details or the structure of the visible lines in the result, i.e., the content of the image, the same scene would be represented with full detail, which would be called the *cognitive* approach. If the goal is only the pleasant appearance of the image, we speak about an *aesthetical* approach. Any given TM method will realize a mixture of these three approaches, with a different weighting given to each [22].

In this contribution, we concentrate on the perceptual approach, and aim to characterize the *overall image quality* (OIQ) resulting from a TM technique in a perceptual sense. In addition, we have chosen a number of important image attributes which are typically used to characterize tone mapped images, and study how well TM methods reproduce these attributes: brightness, contrast, color, detail and artifacts. The chosen attributes are mostly perceptual, but contain cognitive and aesthetics aspects as well. Beyond these attributes, which are related to color and spatial vision, there are some other important aspects and some "special effects" which can improve or modify the final appearance. Since some of the attributes are not mutually independent (as we will explain later), we propose a scheme of relationships between them (Fig. 6). The goal of this work is to investigate the influence these attributes have on overall image quality, based on a subjective study.

4. Image attributes

In this section, we briefly survey particular image attributes for TM, and we list some typical TM methods that attempt to reproduce them correctly. As this part has the character of a survey, an informed reader can skip directly to the experiments described in Section 6.

4.1. Brightness

Brightness is a quantity that measures the subjective sensation produced by the absolute amount of luminance [23]. More specifically, brightness is the attribute of a visual sensation according to which an area appears to emit more or less light [24]. The magnitude of brightness can be estimated for unrelated visual stimuli (since it is an absolute unit) as well as for related visual stimuli. *Lightness* is defined as the attribute of a visual sensation according to which the area in which the visual stimulus is presented appears to emit more or less light in proportion to that emitted by a similarly illuminated area perceived as a "white" stimulus [24]. Lightness has thus meaning only for related visual stimuli. As lightness is judged with reference to the brightness of the "white" stimulus, it may be considered a special form of brightness measure that could be referred to as relative brightness [24]. In this study, we concern ourselves with the quality of reproduction of an "overall" brightness of the inquired HDR scene.

Stevens and Stevens, see [25], proposed an expression for the apparent brightness, but although the expression gives a convenient relationship between luminance and brightness for simple targets, the overall brightness of an image is more complex. A method by Tumblin and Rushmeier [26] attempts to preserve the overall impression of brightness using a mapping function that is based on the model by Stevens and Stevens [25].

This mapping function matches the brightness of a real-world luminance to the brightness of a display luminance. Recently, Krawczyk et al. [27] proposed a method which aims for an accurate estimation of lightness in real-world scenes by means of the so-called anchoring theory of lightness perception. The method is based on an automatic decomposition of the HDR image into frameworks (consistent areas). Lightness of a framework is then estimated by the anchoring to the luminance level that is perceived as white, and finally, the global lightness is computed.

4.2. Contrast

Image contrast is defined in different ways, but it is usually related to variations in image luminance. There exist various basic formulae for computation of contrast, see the thesis by Winkler [28] for an overview. Matkovic et al. [29] proposed a complex computational global *contrast measure* called global contrast factor that uses contrasts at various resolution levels in order to compute overall contrast. In this study, we think about overall contrast in a similar way.

Ward's [30] initial TM method focuses on the preservation of *perceived contrast*. This method transforms input luminance to output luminance using a scaling factor. The computation of the factor is based on Blackwell's [31] psychophysical contrast sensitivity model. Because Ward's method scales image intensities by a constant, it does not change scene contrasts for display. Almost the same principle of contrast preservation is exploited also in other methods [32,33].

Advanced local TM methods (e.g., the method [34] or [35]) are based on a multi-resolution decomposition of the image and approximate contrast in a way similar to Peli [36], see Fig. 1. Mantiuk et al. [37] proposed a framework for perceptual contrast processing of HDR images. The authors define contrast as a difference between a pixel and one of its neighbors at a particular level of a Gaussian pyramid. This approach resembles the gradient-domain method by Fattal et al. [38].

4.3. Reproduction of colors

The sensation of color is an important aspect of the HVS, and a correct reproduction of colors can increase the apparent realism of an output image. One important feature of the HVS is the capacity to see the level of colors in a bright environment. This ability, measured as color sensitivity, is reduced in dark environments, as the light sensitive rods take over for the color-sensitive cone system, see Fig. 2. As the luminance level is raised, the cone system becomes active and colors begin to be seen. Furthermore, the HVS has the capability of *chromatic adaptation*. Humans are able to adjust to varying colors of illumination in order to approximately preserve the appearance of object colors. See Fairchild's book [25] for more information on color appearance modeling.

The TM method by Ferwerda et al. [32] captures changes in threshold color appearance by using separate threshold versus intensity (TVI) functions for rods and cones and interpolation for the mesopic luminance range. Ward et al. [33] used a very similar approach. Pattanaik et al. [39] proposed a comprehensive multi-scale model that accounts for changes both in threshold color discriminability and suprathreshold colorfulness. Using opponent color processing, the model is able to handle changes in chromatic and luminance-level adaptation as well. In their work, Reinhard and Devlin [40] adapted a computational model of photoreceptor behavior that incorporates a chromatic transform that allows the white point to be shifted.

4.4. Reproduction of details

The reproduction of details is an issue mainly in very dark and very bright areas, because truncation of values occurs most

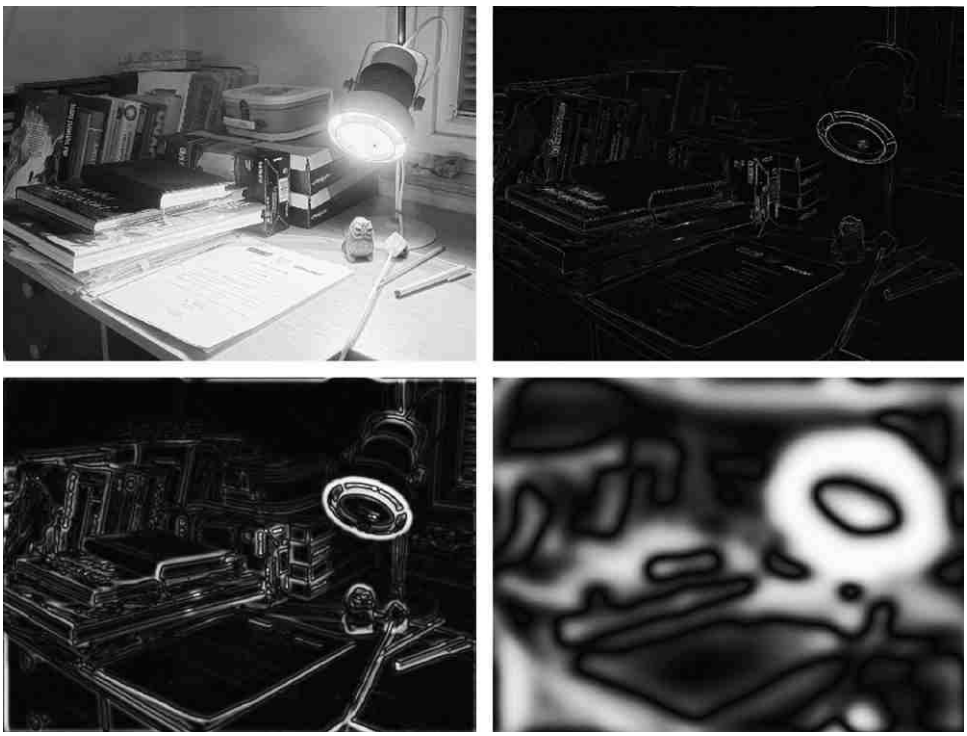


Fig. 1. Peli's local band-limited contrast on three different spatial resolutions (top-left: original image).

Fig. 2. Simulation of color sensitivity. Left: original image—no color sensitivity simulation. Right: simulation of the loss of color sensitivity in the dark.

Fig. 3. Reproduction of details in a very bright area. Left: global TM method exhibits the loss of details. Right: details preservation owing to mapping by a local method.

frequently in these areas as a result of the dynamic range limitations of the output device. The simplest methods (e.g., linear scaling or clamping) will usually reduce or destroy important details and textures (see Fig. 3). On the other hand, the effort to reproduce details well is a potential cause of *artifacts*.

Several TM methods focus especially on the reproduction of details. Tumblin and Turk's LCIS method [41] produces a high detail, low contrast image by compressing only the large features and adding back all small details. The idea of compressing just the large features and then adding subtle noncompressed details is also used in the methods based on the bilateral [9] and trilateral filter [42].

A different approach was presented by Ward [33]. Ward's method based on histogram adjustment aims to preserve *visibility*, where visibility is said to be preserved if we can see an object on the display if and only if we can see it in the real scene. Ward's method does not strive to reproduce all the details available, but exploits the limitations of human vision to reproduce just the *visible* details. Also, most local TM methods try to preserve detail along with contrast.

4.5. Artifacts

As a consequence of tone mapping, *artifacts* may appear in the output image. The artifacts degrade the *overall quality* of the output image. Some local TM methods [43,44] exhibit typical *halo artifacts*, see Fig. 4. These artifacts are caused by contrast reversals, which may happen for small bright features or sharp high contrast edges, where a bright feature causes strong attenuation of the neighboring pixels, surrounding the feature or high contrast edge with a noticeable dark band or halo.

Another possible artifact of TM methods stems from the superficial handling of colors. Many TM methods use very simple rules in handling of the colors, e.g., doing the HDR to LDR transformation just for the luminance component with consequential restoration of the color information. Apart from poor values for the color reproduction image attribute, this can also lead to visible *color artifacts* like oversaturation, see Fig. 4. Closely related to color artifacts are *quantization artifacts*, especially in dark regions, which stem from applying transformations (like gamma correction) to a low precision representation of color values.

4.6. Special attributes

The following image attributes show up just under special conditions and we do not consider them in our current experiments, in favor of the basic ones. Moreover, we avoided testing of glare and visual acuity simulation, because these effects are usually implemented in the same way as a postprocess after the TM step. However, we present these attributes here to complete the survey of image attributes for TM and it will be an interesting task to include them in future special evaluations.

Visual acuity is the ability of the HVS to resolve spatial detail. The visual acuity decreases in the dark, since cones are not responding to such low light levels. It is interesting that simulating this phenomenon, i.e., reducing the detail in an image, actually enhances the *perceptual quality* of the image.

Owing to the scattering of light in the human cornea, lens and retina, and due to diffraction in the cell structures on the outer radial areas of the lens, phenomena commonly referred to as *glare effects* [45] are seen around very bright objects, see Fig. 5.

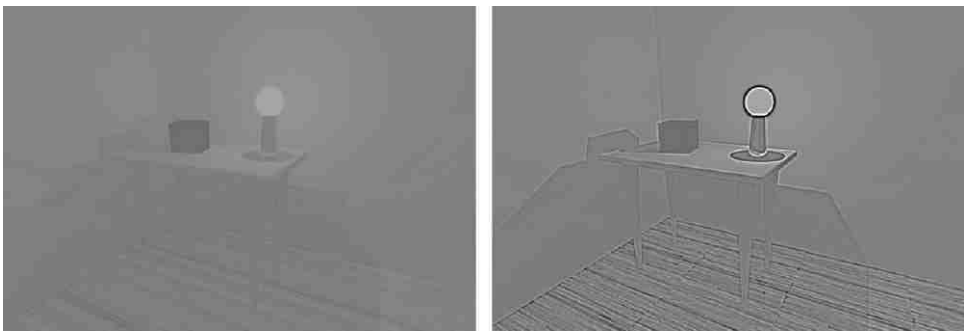


Fig. 4. Halo artifacts and oversaturation. Left: HDR image after successful tone mapping without halo artifacts. Right: the same image after tone mapping using the local method exhibiting a massive amount of halo artifacts. Both images exhibit oversaturation.

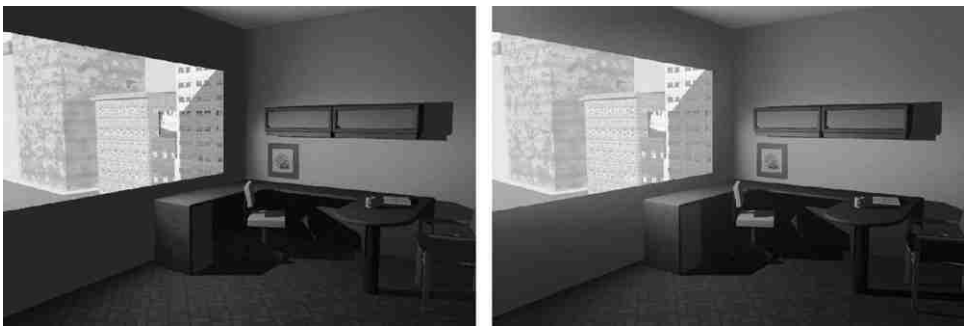


Fig. 5. Bloom (veiling luminance) simulation. Left: the original scene without bloom simulation. Right: the same scene with bloom simulation. *Source HDR image courtesy of Greg Ward.*

Since the dynamic range of traditional output devices is not sufficient to evoke such phenomena, we must simulate the human response artificially to improve the *perceptual quality* of the image.

5. Attribute relationships

In the previous section, we have surveyed the image attributes that are important for TM and influence the overall quality of the output image. These attributes are not independent, and we present a description of their interrelationships in this section.

We propose the scheme shown in Fig. 6 to illustrate the relationships between the attributes. The *overall image quality*, our measure, is determined by all the attributes. It depends strongly on the overall perceived *brightness*, i.e., highly illuminated scenes should be reproduced bright, while dim scenes should appear dark. Apparent *contrast* should also be reproduced well to make the result natural. The reproduction of *details* or rather the reproduction of *visibility* of objects is certainly essential to make the output image appear natural. Furthermore, since we are typically facing a limited display gamut, the reproduction of *color* is an important factor for perceptual quality as well. The simulation of *visual acuity* loss can significantly improve the perceptual quality of dim or night scenes, while the simulation of *glare* can enhance the perceptual quality of the dark scenes with strong light sources. There is no doubt that the presence of disturbing *artifacts* degrades perceptual quality. But there are also important interrelationships of the attributes:

The perception of *brightness* is affected greatly by the *contrast arrangement* (i.e., by the semantics of an image). Fairchild [25] described the effect of image contrast on the perceived brightness and concluded that the brightness typically increases with *contrast*. It has been shown that brightness increases as a function

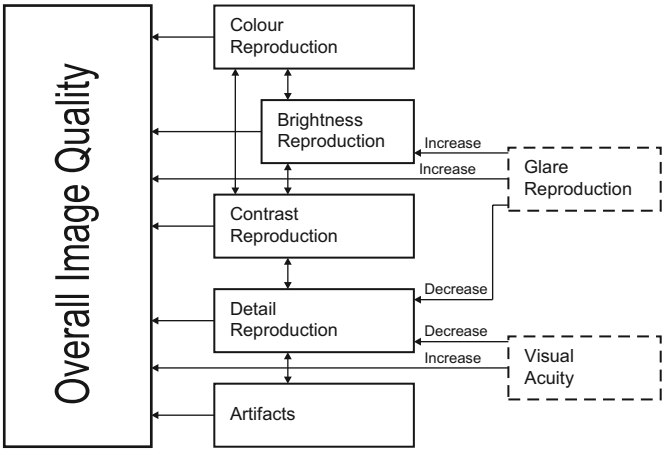


Fig. 6. The relationships between image attributes. The attributes we did not evaluate in subjective perceptual experiments are in dashed boxes.

of chroma (Helmholtz–Kohlrausch effect). Moreover, the simulation of color appearance at scotopic levels of illumination can substantially change the perceived brightness. Finally, the *simulation of glare* plays an important role for the brightness perception. The glare simulation increases the apparent brightness of light sources.

It was shown that *contrast* increases with the *luminance* (Stevens effect, see [25]). Since we can identify the contrast at different spatial resolutions, the perception of contrast is obviously affected by the reproduction of *details*. The experimental results of Calabria and Fairchild [46] confirmed that the perceived contrast depends also on image *lightness, chroma* and *sharpness*.

Fig. 7. Example of subjective perceptual experiments setups. Left: rating experiment with real references, Right: ranking experiment without references.

Colors are related to brightness, because the colorfulness increases with the luminance level (i.e., the Hunt effect [25]).

The reproduction of *details* is strongly affected by the simulation of the *visual acuity*. Since there are available data that represent the visual acuity (e.g., Shaler's curve [32]), these data place limits on the reproduction of fine details, and may also be utilized to verify the perceptual quality of detail reproduction. Furthermore, the *visibility preservation* diminishes the reproduced details using a threshold function (e.g., the TVI). The simulated *glare* can obscure otherwise reproducible details near strong light sources.

Using subjective testing results, Spencer et al. [45] verified that the *simulation of glare* can substantially increase the apparent *brightness* of light sources in digital images.

In the scheme of relationships (Fig. 6), we can identify attributes that represent *limitations* of the HVS: the simulation of glare, the simulation of visual acuity and (in part) the reproduction of color (in the sense of simulation of the scotopic vision). These attributes enhance the perceptual quality of the output image, but are not desirable when the goal is different, for example when we aim to reproduce as many details as possible.

6. Subjective perceptual studies

We have conducted two separate and technically different subjective perceptual studies: (1) a rating-based experiment with reference real-world scenes and (2) a ranking-based experiment with no references, see Fig. 7. These experiments were conducted to encourage the proposed idea of an *overall image quality* measure and to verify the correlations to and between the image attributes shown in Fig. 6. Moreover, the execution of two principally different studies gave us the opportunity to relate the obtained subjective results. Finally, we used the results of perceptual studies to evaluate the strengths and weaknesses of 14 TM methods.

Prior to the main experiments we have conducted a pilot study to examine the setup and to verify that subjects were able to rate "soft-copy" images against the real scenes (i.e., rating experiment verification). During this study we have also fine-tuned the parameters of several TM methods, and we have refined instructions given to subjects. Preliminary ideas of the project as well as the results of our pilot study have been presented in [3].

It is worth noting that apart from the evaluation of the 14 involved TM methods, the results concerning the relations of image attributes and overall perceptual quality of an image are totally independent of any particular TM method or of the values of its parameters (i.e., the 14 tone mapped images represent a collection of natural input visual stimuli in our subjective perceptual studies). We believe that the collection of images we used is much more natural than the usual artificial stimuli used in vision science for narrow perceptual studies, where images are very simple derivations of an original LDR image (thresholding, scaling, chroma variations or so).

Table 1
Numerical luminance values ($\log_{10}\,\mathrm{cd/m^2}$) for the experimental HDR images

	Min	Max	Mean	Dynamic range
Night scene	−2.33	2.77	−0.99	5.13
Indoor scene	−1.09	4.27	0.82	5.37
Outdoor scene	0.63	6.08	2.69	5.45

6.1. Subjective testing setup

We arranged three representative HDR real-world scenes for our experiments: a typical real-world indoor HDR scene, see Table 3, a typical HDR outdoor scene, see Table 4, and a night urban HDR scene, see Table 5. We acquired a series of 15 photos of each scene using a digital camera (Canon EOS300D, Sigma DC 18-200) with varying exposure (fixed aperture $f/11$, varying shutter speeds) from a locked-down tripod. The focal length was around 50 mm (crop factor equivalent) for all scenes—which corresponds to the normal FOV of an observer. The HDR radiance maps were recovered from the recorded series using the method of Debevec and Malik [47]. The dynamic ranges of the resulting HDR images of the indoor scene, outdoor scene and night urban scene were about $10^5{:}\,10^{-1}\,\mathrm{cd/m^2}$, $10^6{:}\,10^{1}\,\mathrm{cd/m^2}$ and $10^3{:}\,10^{-3}\,\mathrm{cd/m^2}$, respectively (numerical values as reported by the `pfsstat` utility[1] are summarized in Table 1), luminance histograms are shown in Fig. 8.

We transformed these input HDR images using 14 different TM methods, so that we obtained 14 LDR images[2] per scene for investigation. We attempted to include the largest possible amount of methods (see [1,2] for an overview) into the evaluation, and came up with the 14 techniques (see Table 2) to be included into our experiment (abbreviations are used through the entire

[1] Available at http://www.mpi-inf.mpg.de/resources/pfstools/

[2] All the tone mapped images as well as the original HDR images are available on the web pages of the project: http://www.cgg.cvut.cz/~cadikm/tmo

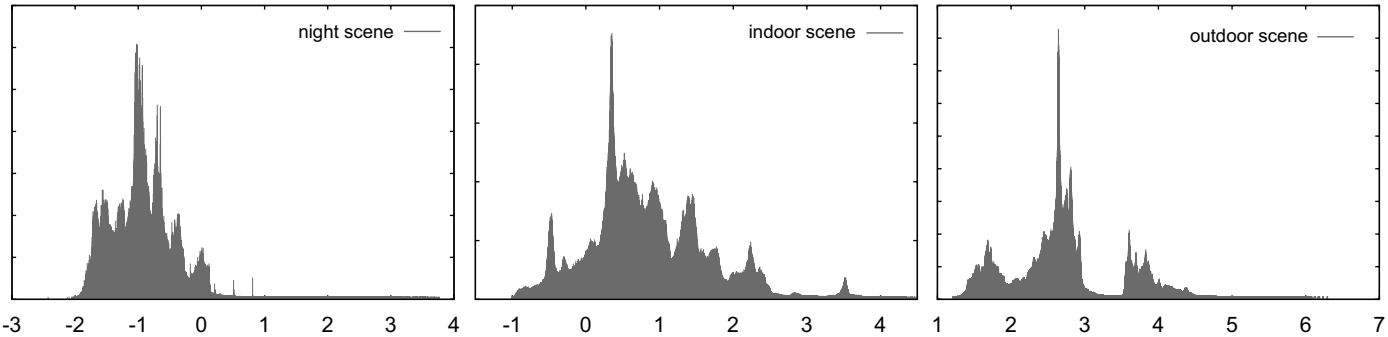


Fig. 8. Luminance histograms ($\log_{10}$) of the experimental HDR images, from left: night scene, indoor scene, outdoor scene.

Table 2
Abbreviations of evaluated tone mapping methods

Abbreviation	Method description	Publication	Global/Local
Ashikhmin02	A tone mapping algorithm for high contrast images	[35]	L
Chiu93	Spatially nonuniform scaling functions for high contrast images	[43]	L
Choudhury03	The trilateral filter for high contrast images and meshes	[42]	L
Drago03	Adaptive logarithmic mapping for displaying high contrast scenes	[48]	G
Durand02	Fast bilateral filtering for the display of HDR images	[9]	L
Fattal02	Gradient domain high dynamic range compression	[38]	L
LCIS99	Low curvature image simplifier	[41]	L
Pattanaik02	Adaptive gain control for HDR image display	[49]	L
Reinhard02	Photographic tone reproduction for digital images	[34]	L
Schlick94	Quantization techniques for visualization of HDR pictures	[44]	L
Tumblin99	Revised Tumblin–Rushmeier tone reproduction operator	[50]	G
Ward94	A contrast-based scalefactor for luminance display	[30]	G
Ward97	A visibility matching tone reproduction operator for HDR scenes	[33]	G
Linear Clip	Manual linear clipping		G

paper); for the resulting images see Tables 3–5. All the evaluated methods were implemented by the first author with some discussions and help from the original authors of these methods.

The sequence of 14 LDR TM images represented the input visual stimuli for each observer, all the testings were performed under controlled ambient luminance level. A total number of 20 subjects aged between 26 and 52 were involved in our experiments. All participating subjects had normal or corrected-to-normal vision and were nonexperts in the field of TM. In the two experimental studies, we collected in total $3_{(\text{scenes})} \cdot (10+10)_{(\text{subjects})} \cdot 6_{(\text{attributes})} \cdot 14_{(\text{methods})} = 5040$ values of observation scores.

In the first experiment, based on *rating* (see Fig. 7—left), we simultaneously presented an original (*real-world*) *HDR scene* and the appropriate TM images of this scene to human observers. In order to keep the illumination moderately constant, we performed all the testing procedures at the same time of the day as the HDR image was acquired, continually inspecting the illumination conditions using an exposure meter. The TM images were shown separately in random order on a calibrated monitor[3] to a group of 10 subjects. The task of each subject was to express the overall image quality, and the quality of reproduction of basic attributes—overall brightness, overall contrast, reproduction of details, overall reproduction of colors and the lack of disturbing image artifacts for a particular image by *ratings* (on the scale 1–10, where 10 represents the best result, while 1 is the worst) with respect to the actual scene. All subjects were verbally introduced to the experiment and they were instructed to "Rate the images on how close the particular image attribute matches in appearance to the real-world scene" (attribute reproduction results) and to "Rate the images on how close the overall match in appearance is to the real-world scene" (overall image quality results). To avoid any confusion, subjects were personally informed that we were interested in *quality of reproduction* (not the amount or quantity) of inquired image attributes (e.g., "Less detail in the image than in the ground truth is bad, more detail in the image than in the ground truth is bad as well, the closer to the ground truth the better the score should be.") and that they should judge only the particular attribute and avoid any influence of other attributes. Subjects sat at the place of the camera at common viewing distance from the display (approximately 60 cm) and they were able to directly observe both the real scene and the display. However, subjects were always instructed to take a few seconds to adapt to each. The procedure took approximately 45 min for one observer and one scene. We chose the rating scale method in this experiment to stimulate observers to do the direct comparison of the TM image to the real scene.

In the second experiment, based on *ranking* (see Fig. 7—right), we investigated what happens when subjects have no possibility of directly comparing to the ground truth (or are not affected by a previous experience with the real scene). A group of 10 observers (different ones than in the first experiment), who have *never seen the real HDR scenes* and had therefore virtually no idea about the attributes of original scenes, was selected. The task of each subject was to order (*rank*) *image printouts* resulting from the 14 methods according to the overall image quality, and the quality of reproduction of overall contrast, overall brightness, colors, details and image artifacts. Similarly to the first experiment, all subjects were verbally introduced to the experiment and they were

[3] FSC P19-2, 19″ LCD display, with maximum luminance of 280 cd/m^2. We used manufacturer's ICC profiles (D65) for both the monitor and the camera to perform the colorimetric characterization of the devices.

Table 3
Strengths and weaknesses of evaluated TM methods—*indoor scene*

Method	Image	Brightness	Contrast	Details	Colors	Overall quality	Method	Image	Brightness	Contrast	Colors	Details	Overall quality
Linear Clip		**10.6** *2.8* **6.3** *3.1*	**7.6** *3.9* **6.6** *3.9*	**7.6** *4.7* **3.6** *3.7*	**11.3** *3.6* **8.4** *3.7*	**8.9** *3.0* **5.7** *2.9*	LCIS99		**4.1** *1.5* **11.2** *2.1*	**6.2** *2.6* **10.8** *1.9*	**5.4** *3.9* **9.8** *3.3*	**3.4** *1.2* **9.3** *3.3*	**4.6** *1.3* **11.9** *2.4*
Ward94		**7.7** *3.0* **7.1** *2.7*	**8.1** *4.0* **9.7** *2.2*	**5.3** *1.9* **6.8** *2.5*	**9.6** *2.4* **9.8** *2.5*	**9.7** *3.7* **7.9** *3.4*	Pattanaik02		**11.1** *3.6* **8.2** *4.5*	**8.9** *3.6* **8.2** *3.9*	**12.4** *2.5* **9.4** *3.5*	**8.6** *3.3* **8.1** *3.5*	**6.8** *3.3* **7.6** *3.2*
Tumblin99		**11.1** *1.6* **9.0** *1.8*	**9.5** *3.3* **9.6** *3.5*	**7.5** *3.9* **6.9** *1.9*	**10.3** *1.9* **10.4** *1.9*	**10.8** *3.1* **8.0** *2.5*	Choudhury03		**5.2** *1.5* **10.2** *2.2*	**5.9** *2.3* **8.8** *2.5*	**7.0** *2.6* **9.6** *2.3*	**5.4** *1.4* **7.7** *2.3*	**2.2** *3.6* **10.4** *2.0*
Reinhard02		**10.8** *1.9* **11.9** *1.9*	**11.6** *2.7* **12.1** *2.6*	**10.4** *2.8* **11.8** *1.7*	**12.5** *1.4* **12.6** *1.8*	**12.2** *1.1* **11.9** *2.3*	Drago03		**10.9** *1.5* **7.5** *2.2*	**9.5** *1.8* **5.8** *1.9*	**6.9** *3.9* **5.1** *2.3*	**9.0** *3.0* **6.5** *2.3*	**8.9** *1.5* **7.2** *1.7*
Schlick94		**3.8** *2.4* **6.9** *4.2*	**7.1** *3.3* **8.7** *4.3*	**6.2** *3.8* **6.7** *3.9*	**5.6** *2.9* **9.1** *3.9*	**9.3** *3.1* **8.2** *4.6*	Ashikhmin02		**8.3** *2.5* **7.3** *2.9*	**8.0** *3.7* **6.5** *2.0*	**10.2** *2.6* **9.6** *2.6*	**8.3** *3.2* **5.0** *2.6*	**7.6** *3.3* **7.5** *1.3*
Ward97		**8.8** *2.5* **10.4** *2.3*	**9.8** *3.3* **9.7** *2.1*	**8.1** *3.2* **9.4** *1.9*	**10.3** *2.3* **10.0** *1.9*	**11.5** *1.8* **10.8** *2.2*	Fattal02		**3.2** *1.0* **3.7** *2.7*	**5.4** *3.6* **3.8** *2.6*	**7.4** *4.2* **8.2** *2.7*	**5.0** *1.8* **2.4** *2.7*	**5.8** *2.4* **3.4** *3.4*
Durand02		**8.4** *3.7* **2.9** *1.6*	**4.7** *4.4* **2.2** *0.8*	**6.9** *4.0* **5.0** *0.9*	**4.6** *2.9* **2.4** *0.9*	**3.5** *2.7* **2.75** *2.4*	Chiu93		**1.1** *0.3* **2.5** *1.9*	**2.7** *3.0* **2.7** *2.6*	**3.0** *2.5* **3.3** *1.6*	**1.1** *0.3* **3.5** *1.6*	**1.8** *1.3* **1.9** *0.9*

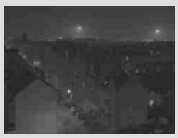
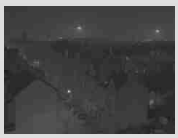
In bold: average ranking scores (1st line) and average rating scores (3rd line); in italics: standard deviations for ranking (2nd line) and for rating scores (4th line). The higher values represent the higher reproduction quality.

Table 4
Strengths and weaknesses of evaluated TM methods—*outdoor scene*

Method	Image	Brightness	Contrast	Details	Colors	Overall quality	Method	Image	Brightness	Contrast	Colors	Details	Overall quality
Linear Clip		**12.3** *2.2* **10.3** *3.1*	**13.2** *0.9* **10.7** *2.9*	**12.5** *1.9* **9.4** *3.6*	**13.4** *0.5* **10.1** *3.6*	**13.2** *0.9* **11.8** *2.5*	LCIS99		**9.4** *2.2* **7.8** *3.3*	**7.9** *1.6* **7.8** *3.2*	**8.8** *1.5* **8.7** *2.6*	**7.9** *2.0* **8.5** *2.6*	**7.8** *1.7* **7.7** *3.0*
Ward94		**4.3** *2.0* **7.4** *3.1*	**6.2** *2.6* **10.2** *2.3*	**4.1** *3.1* **7.2** *2.3*	**5.7** *1.4* **9.1** *2.3*	**8.2** *1.8* **9.7** *2.2*	Pattanaik02		**6.1** *4.1* **2.3** *0.8*	**4.1** *1.8* **2.1** *0.8*	**3.4** *2.2* **4.8** *3.4*	**2.1** *0.3* **3.1** *2.1*	**2.1** *0.3* **2.3** *0.7*
Tumblin99		**12.4** *2.3* **10.3** *2.5*	**12.7** *1.2* **8.4** *2.1*	**12.5** *1.9* **7.3** *4.1*	**13.6** *0.5* **9.8** *3.4*	**12.9** *1.0* **11.2** *1.5*	Choudhury03		**8.2** *1.7* **6.0** *2.4*	**5.8** *1.3* **6.1** *3.2*	**7.9** *1.4* **7.8** *3.1*	**7.3** *1.7* **7.1** *2.4*	**6.0** *1.5* **7.7** *3.0*
Reinhard02		**9.2** *2.0* **13.1** *1.1*	**10.6** *0.9* **12.0** *1.7*	**9.0** *1.6* **9.7** *4.6*	**10.1** *1.1* **11.7** *1.9*	**7.9** *3.2* **9.7** *2.9*	Drago03		**3.6** *2.1* **6.2** *2.3*	**4.1** *2.4* **6.8** *2.3*	**5.1** *2.2* **6.1** *1.7*	**5.4** *2.5* **5.0** *3.6*	**6.9** *2.7* **5.7** *1.5*
Schlick94		**9.4** *2.7* **10.6** *2.0*	**10.6** *1.6* **10.3** *2.6*	**10.7** *2.6* **9.7** *2.7*	**10.9** *0.5* **9.8** *2.2*	**11.5** *1.4* **11.1** *2.3*	Ashikhmin02		**8.7** *2.5* **6.6** *2.4*	**6.9** *1.8* **6.0** *2.1*	**7.7** *2.2* **7.9** *4.1*	**6.8** *1.5* **7.1** *2.3*	**4.9** *1.3* **5.2** *2.5*
Ward97		**10.6** *2.1* **9.7** *2.9*	**11.6** *2.3* **9.0** *3.7*	**12.3** *1.1* **8.9** *3.4*	**11.4** *0.8* **8.7** *2.5*	**11.3** *1.8* **8.9** *1.5*	Fattal02		**2.8** *1.3* **4.4** *4.1*	**1.8** *1.1* **2.9** *0.9*	**3.1** *1.4* **5.5** *4.7*	**3.7** *1.9* **2.0** *0.7*	**3.5** *1.2* **2.3** *0.9*
Durand02		**3.6** *2.0* **7.8** *3.4*	**6.0** *2.9* **10.4** *2.5*	**4.8** *2.6* **7.0** *2.0*	**5.6** *1.3* **9.0** *2.4*	**7.7** *1.6* **10.1** *2.7*	Chiu93		**4.4** *3.8* **2.9** *2.2*	**3.5** *3.0* **2.6** *2.6*	**3.1** *2.3* **5.3** *5.1*	**1.1** *0.3* **4.4** *3.8*	**0.3** *1.1* **1.8** *0.5*

In bold: average ranking scores (1st line) and average rating scores (3rd line); in italics: standard deviations for ranking (2nd line) and for rating scores (4th line). The higher values represent the higher reproduction quality.

Table 5
Strengths and weaknesses of evaluated TM methods—*night scene*

Method	Image	Brightness	Contrast	Details	Colors	Overall quality	Method	Image	Brightness	Contrast	Colors	Details	Overall quality
Linear Clip		**11.3** *3.6* **10.0** *3.0*	**12.8** *1.2* **9.8** *2.1*	**13.2** *1.3* **8.8** *3.3*	**12.2** *2.7* **10.4** *2.9*	**12.9** *1.0* **10.2** *3.1*	LCIS99		**6.5** *2.3* **7.6** *2.2*	**6.2** *1.5* **6.9** *2.9*	**5.8** *0.9* **7.5** *2.1*	**6.1** *2.7* **6.5** *2.0*	**5.7** *0.5* **7.7** *2.6*
Ward94		**10.6** *3.0* **10.2** *3.2*	**12.1** *1.8* **11.1** *3.2*	**12.1** *1.5* **9.6** *3.9*	**11.9** *2.2* **11.5** *3.0*	**12.5** *1.4* **11.9** *1.7*	Pattanaik02		**9.1** *2.0* **10.4** *3.9*	**10.0** *1.0* **12.2** *1.9*	**10.5** *1.8* **12.7** *1.3*	**9.6** *2.5* **11.7** *2.0*	**9.2** *1.6* **12.5** *1.8*
Tumblin99		**7.4** *2.3* **8.8** *2.7*	**7.6** *1.7* **8.6** *3.1*	**8.1** *1.0* **6.8** *2.9*	**8.4** *1.5* **8.4** *3.4*	**8.8** *1.3* **9.4** *2.7*	Choudhury03		**7.1** *2.6* **6.7** *1.9*	**6.6** *2.3* **5.6** *3.4*	**5.3** *0.9* **7.3** *2.9*	**5.9** *2.6* **6.5** *2.8*	**5.1** *1.0* **6.9** *2.5*
Reinhard02		**9.1** *3.6* **6.8** *4.2*	**9.0** *2.5* **8.5** *2.4*	**9.4** *1.3* **9.8** *3.7*	**9.7** *1.4* **10.0** *2.3*	**9.3** *1.3* **8.7** *2.6*	Drago03		**4.9** *4.2* **6.9** *3.9*	**4.9** *3.9* **4.9** *2.5*	**3.3** *0.6* **5.2** *3.0*	**3.8** *2.6* **3.9** *3.2*	**3.4** *1.2* **4.4** *1.9*
Schlick94		**8.8** *2.6* **7.5** *3.1*	**9.3** *2.1* **10.9** *1.8*	**9.7** *1.3* **9.5** *2.5*	**9.5** *1.8* **9.3** *1.9*	**10.4** *1.7* **8.8** *1.9*	Ashikhmin02		**5.1** *3.4* **8.2** *3.2*	**3.6** *1.1* **4.8** *2.2*	**3.6** *1.0* **8.5** *4.4*	**4.5** *2.7* **5.2** *2.6*	**3.3** *1.0* **4.8** *2.3*
Ward97		**8.7** *2.8* **3.7** *2.8*	**7.0** *1.8* **6.3** *2.0*	**7.8** *2.0* **6.5** *2.0*	**8.1** *1.4* **6.9** *2.0*	**7.7** *1.3* **5.9** *1.9*	Fattal02		**4.0** *2.5* **5.7** *4.8*	**2.4** *0.5* **3.7** *3.2*	**2.4** *0.7* **3.0** *3.0*	**2.6** *0.5* **2.6** *1.1*	**2.7** *0.6* **2.1** *0.4*
Durand02		**11.4** *2.5* **8.9** *3.7*	**12.5** *1.4* **10.9** *2.5*	**12.8** *0.4* **8.9** *2.9*	**11.7** *2.2* **10.9** *2.2*	**13.0** *0.6* **11.0** *2.4*	Chiu93		**1.0** *0.0* **3.9** *5.0*	**1.0** *0.0* **1.1** *0.2*	**1.0** *0.0* **1.2** *0.3*	**1.0** *0.0* **1.3** *0.6*	**1.0** *0.0* **1.1** *0.2*

In bold: average ranking scores (1st line) and average rating scores (3rd line); in italics: standard deviations for ranking (2nd line) and for rating scores (4th line). The higher values represent the higher reproduction quality.

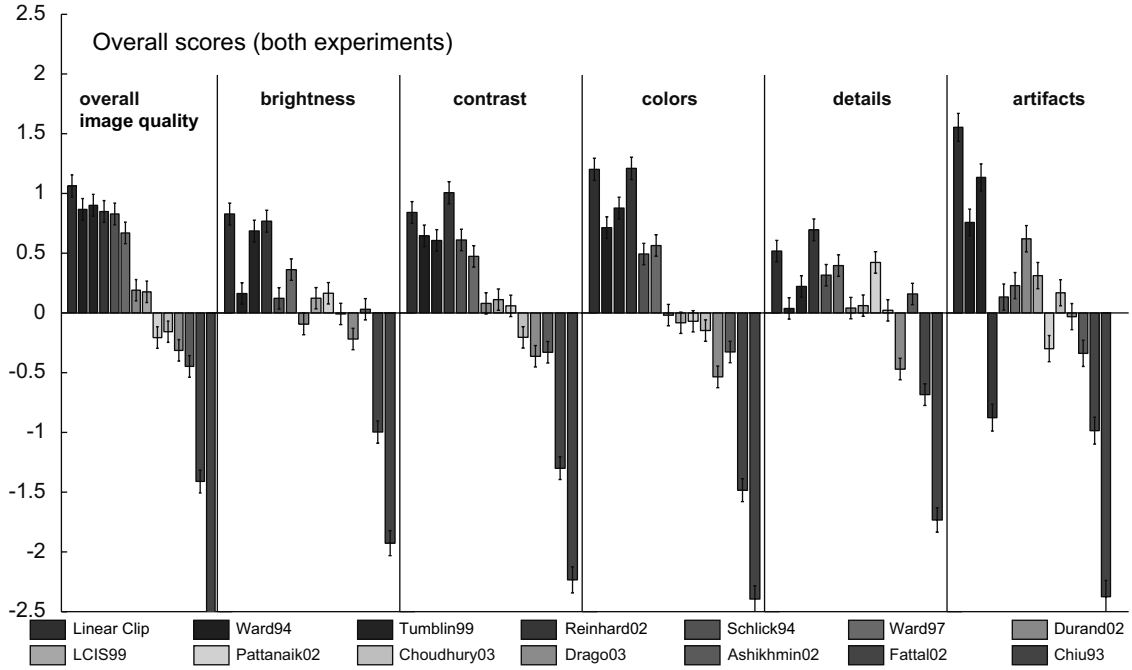


Fig. 9. Overall accuracy scores for all examined TM methods. Left to right: overall perceptual quality, reproduction of brightness, reproduction of contrast, reproduction of details, reproduction of colors, lack of disturbing artifacts. In each chart the higher value represents the higher reproduction quality.

instructed to "Rank the printouts on how close the particular image attribute matches in appearance to a *hypothetical* real-world scene," the idea being that when a human views an image, she always forms a mental model of the original scene. Thus, the description of image attributes was the same as in the first experiment, but observers were instructed to "Imagine how the original real-world scene would look like" and rank the printouts accordingly. The procedure took approximately 35 min for one observer and one series of input images. The investigated printouts were high quality color image printouts on a glossy paper of the same 14 tone mapped images as in the first experiment.[4] Printouts were observed in an office under standard illumination of approximately 550 lux.

7. Results and discussion

In order to make the results of the two conducted experiments comparable, we converted the *rating* observation scores to the ranking scale by computing the ranks of observations for each person and attribute with adjustment for ties (if any values were tied, we computed their average rank) prior to the following evaluations. For example, a rating observation vector $\mathbb{X}$ is converted to the rank vector $\mathbb{X}'$ as follows:

$$\mathbb{X} = (3\ 7\ 2\ 6\ 2\ 1\ 5\ 6\ 9\ 5\ 6\ 8\ 8\ 4)$$
$$\mathbb{X}' = (4\ 11\ 2.5\ 9\ 2.5\ 1\ 6.5\ 9\ 14\ 6.5\ 9\ 12.5\ 12.5\ 5)$$

We analyzed the data using non-parametric statistical tests.[5] Moreover, we also converted these rank order data using the Thurstonian model (condition D) [51,52] to interval scales. Tables 3–5 show the numerical results separately for each scene, while interval scales are shown along with standard errors in Fig. 9 (overall average results), in Fig. 11 (average values for each experiment), and Fig. 13 (overall image quality ratings for each input scene for each experiment). We describe and discuss the obtained results in the following text: Sections 7.1 and 7.2 statistically prove that neither the experimental setup nor the choice of scenes has a systematic influence on the results. In Section 7.3 we discuss the results of examined TM methods. In Section 7.5 we quantify the relationship between image attributes proposed in Section 5. Finally, in Section 7.5 we compare our results to results obtained in previous work.

7.1. Effects of input scenes and methods

First, we have to inquire if the *input scene* has a significant systematic effect on the evaluation of the methods and image attributes. We use Friedman's nonparametric two-way analysis of variance (ANOVA) test [53] for each image attribute independently for ranking and rating data sets. We state the null hypothesis H_0 as follows: there is no significant difference between observation values for the input scenes.

We summarize the results for all image attributes in Table 6. If the value of Friedman's statistics Q is higher than the tabulated critical value Q_{crit}, we reject the null hypothesis H_0. For all the cases we use a significance level of $p<0.05$. As we can observe in Table 6, we cannot reject the null hypothesis for any of the attributes for both experimental setups. This means we were not able to find a statistically significant difference between the three input scenes and we can thus proceed with the evaluation independently of the input scenes.

Next, we have to verify that there are significant differences between the *TM methods* and the evaluation of TM methods thus makes sense. We use Friedman's analysis independently for ranking and rating, with the null hypothesis H_0: there is no significant difference between observation values for 14 evaluated methods.

The results are summarized in Table 7. Since all obtained Q values are much higher than Q_{crit}, we reject the null hypothesis for

[4] A HP Color Laserjet 3500 was used, with the manufacturer's ICC profile to perform colorimetric characterization, in order to achieve a reasonably comparable color representation as in the first experiment.

[5] Since we have non-normally distributed observation values (rank orders), we use nonparametric tests throughout this paper.

Table 6
Results of two separate Friedman's tests for the effect of input scenes

	Q_{rating}	$Q_{ranking}$
Overall quality	2.7984	4.5823
Brightness	2.2857	2.7648
Contrast	1.2857	0.3984
Details	0.1429	3.8353
Colors	1.2857	3.6545
Artifacts	0.1231	1.3740
Critical value $Q_{crit} = 5.99$		

Table 7
Results of two separate Friedman's tests for the effect of input methods

	Q_{rating}	$Q_{ranking}$
Overall quality	85.093	110.98
Brightness	72.772	83.494
Contrast	87.782	92.531
Details	56.826	89.617
Colors	91.939	111.91
Artifacts	75.833	92.768
Critical value $Q_{crit} = 19.16$		

all attributes. This means we found significant differences between the method scores for all attributes and both experiments and we can proceed with the evaluation of TM methods.

7.2. Effect of the experimental setup

The next question is if there is a statistically significant difference between the data obtained from the two different *experimental setups* (two conducted psychophysical experiments). Recall that in the rating experiment, observers were able to directly rate the quality of image attributes against the real reference (real HDR scene), while in the ranking experiment they had to rank the images according to the quality of image attributes without knowledge of the original scene, see Fig. 7. The second experiment, even though without reference, was not a simple preference experiment, since observers were instructed to rank images according their mental model of the original real-world scene. We chose two different evaluation methods because unlike in the second experiment, in the first experiment we did not want to show all the 14 images simultaneously with the reference scene. We rather wanted to stimulate the observer to rate a single image against the real reference, thus slightly eliminating the ranking of tested images (this is, however, never fully possible). The rating scale was chosen so that the scores were in the interval $[1, 10]$.

To examine the differences between the rating and ranking experiments[6] for each attribute, we used the Kruskal–Wallis test [53] (nonparametric version of one-way ANOVA). The critical value for the test ($\nu = 14 \times 10 \times 3 - 1 = 419$ degrees of freedom) is $\chi^2_{crit} = 467.73$. All the obtained results of the test were much smaller than the critical value, therefore we did not detect any significant difference between experiments for any attributes using the nonparametric ANOVA.

Since using the Kruskal–Wallis test we did not find any statistically significant differences between the rating and ranking experiments, we also applied another more rigorous test, the profile analysis [54,55], to the observed data. Profile analysis is a nonparametric test used to verify that changes in a particular stochastic variable have the same tendency for several different objects (rating and ranking experiments in our case). We state the null hypothesis H_0 as follows: the mean values of observation vectors $\mathbb{X}_{rat_i}$ and $\mathbb{X}_{ran_i}$, where $\mathbb{X}_{rat_i}$ and $\mathbb{X}_{ran_i}$ is a vector of observed values from the rating and ranking experiment, respectively, differ just in shift (we say they have parallel profiles). According to the profile analysis process, we compute the test quantity V^*_t for each variable t and we reject H_0 if V^*_t is higher than the computed critical value V^*_{crit}.

First, we calculated H_0 for the ranking and rating results for the profiles over the scenes for each image attribute separately. The observation vectors were then: $\mathbb{A}_{rat_i} = (A_{INDOORrat_i}, A_{OUTDOORrat_i}, A_{NIGHTrat_i})$ and $\mathbb{A}_{ran_i} = (A_{INDOORran_i}, A_{OUTDOORran_i}, A_{NIGHTran_i})$ where $\mathbb{A}$ denotes a particular image attribute, and $A_{DESK*_i}, A_{WINDOW*_i}$, and A_{NIGHT*_i} are the observation values for the Desk, Window and Night scene respectively. The obtained profile analysis results are summarized in Table 8. These results show that we can not reject H_0 for any attribute, this means we did not find a significant difference in profiles for each input scene for the rating and ranking experiments.

Next, we averaged the scores for the input scenes for each attribute for each experimental setup separately and we performed another profile analysis over the following vectors: $\mathbb{X}_{rat_i} = (OIQ_{rat_i}, Bri_{rat_i}, Con_{rat_i}, Det_{rat_i}, Col_{rat_i}, Art_{rat_i})$ and $\mathbb{X}_{ran_i} = (OIQ_{ran_i}, Bri_{ran_i}, Con_{ran_i}, Det_{ran_i}, Col_{ran_i}, Art_{ran_i})$, where OIQ_{*_i}, Bri_{*_i}, etc., are averages over input scenes for image attributes overall image quality, brightness, etc., for rating and ranking experiments. The critical value is in this case $V^*_{crit} = 2.6383$ and the resulting values are $V^*_{OIQ} = -0.3489$, $V^*_{Bri} = -0.4791$, $V^*_{Con} = 0.0565$, $V^*_{Det} = -0.1409$, $V^*_{Col} = -0.0404$, $V^*_{Art} = 0.1727$. Since the V^*_{crit} is higher than the resulting V^* for all image attributes, profile analysis did not find a significant difference in the rating and ranking observation data.

Finally, to account for all the factors (i.e., "subject (observer)", "TM method", "input scene" and "experimental setup") together

Table 8
Results of profile analysis

	V^*_{INDOOR}	$V^*_{OUTDOOR}$	V^*_{NIGHT}
Overall quality	−0.2016	0.0000	0.0000
Brightness	0.7928	0.0000	1.0296
Contrast	0.8021	0.3077	0.1156
Details	−0.1077	−0.1287	0.4361
Colors	0.4008	−0.7951	−0.1773
Artifacts	0.0000	0.2068	0.0000
Critical value $V^*_{crit} = 2.394$			

Table 9
Results of nonparametric MANOVA test

Source of variation	*SS*	*df*	*MS*	*F*	*p*
Experimental setup	−0	1	−0	−0	≈ 1
Input scene	0	2	0	0	≈ 1
TM method	5936.1	13	456.62	49.96	≈ 0
Subject (observer)	0	9	0	0	≈ 1
Residual	7439.4	814	9.13		
Total	13376	839			

SS denotes sum of squares, *df* means degrees of freedom, *MS* denotes mean square, *F* is *F*-value, and *p* is *p*-value for the null hypothesis.

[6] Recall that the rating is converted to ranking by computing the ranks of observations for each person and attribute with adjustment for ties.

in one statistical test, we utilized the recently published permutational multi-factorial MANOVA [56]. This test is a nonparametric analogy of the parametric multi-factorial multivariate ANOVA [57]. Results of permutational MANOVA (summarized in Table 9) show that the factors "subject", "input scene" and "experimental setup" are statistically not significant, i.e., scenes, subjects and types of experiment do not have a significant effect on the resulting scores. The only significant *main effect* is with the factor "TM method", which means that there are significant differences in responses of subjects depending on the type of the TM method. This correlates with the results reported above, and again justifies our experimental setup. Moreover, we also inquired *interaction effects* and found a significant effect of "input scene"×"TM method" ($F = 11.23$, $p < 0.001$) which means that the scores depend on the combination of scene and input method, i.e., there probably exist methods whose performance differs for particular input scenes.

In this section, we made a lot of effort to find a statistically significant difference between the two experiments, but we have not found one. This is a very interesting and important result, because it suggests that for a *perceptual* comparison of TM methods it is sufficient to use ranking without a reference as experimental setup. This type of psychophysical testing is much cheaper in terms of money and time than the setup with original scene and ratings.

7.3. Evaluation of HDR TM methods

We should stress here again that all our evaluations are targeted at the perceptual dimension of TM, i.e., the holy grail is to reproduce the visual sensation of the real HDR scene as closely as possible (as opposed to for example information preservation). Moreover, since all the evaluated methods were implemented personally by the first author of the paper, the results in this section represent also the "achievability" of the results. We do not claim that better results for a particular method could not be achieved after a thorough parameter tuning. We have tested three different HDR scenes with a variety of characteristics, but other input scenes may potentially lead to different results. We should also stress that our evaluation does not reflect computation time, implementation difficulties and other factors, that are also significant in practical applications of TM methods.

The observed values represent the *quality* of reproduction of a particular image attribute, and not its *amount*. For example the average observation values for the reproduction of details show the quality of reproduction of details, not the amount of details. Subjects were instructed to rank/rate the images accordingly, therefore too many or too few details are both rated worse than the right amount of details.

7.3.1. Overall results

The overall results (see interval scores shown in Fig. 9) suggest that the *best overall quality* is generally observed in images produced by global TM methods (TM curves). Interestingly, the average best score is achieved by the simplest possible approach, the manual *linear clipping* of luminance values! However, this is not such a surprising result, because also our previous pilot studies [3] have shown the superiority of global approaches in the perceptual dimension of TM. A possible explanation of this is also suggested by our analysis (see Section 7.4): the proper reproduction of *overall* image attributes (overall contrast, overall brightness, colors) is essential for the natural perception of the resulting image, more so than *local* attributes. The HVS is evidently highly sensitive to any disruptive factors in the overall image attributes, far more than to the absence of some image details. Recall that the group of six best-rated TM methods contains just one local approach—the method Reinhard02 [34], but an essential part of that method is basically a global TM method with advanced parameter estimation.

The worst rated methods were Fattal02—the gradient-based approach, which we believe is a good method, but not so for perceptual applications, and an early local approach Chiu93. At the bounds of the quality interval, the best and the worst methods exhibit also the lowest variance, while the middle zone with often uncertain judgments has higher variances. The observers have typically the same opinion about the best/worst question, but difficulties with the evaluation of some similar cases.

The plot of means of the overall image quality attribute (obtained by a non-parametric MANOVA test [56]) with 95% confidence intervals shows the categorization of TM methods more clearly (see Fig. 10 (left)). As we may observe, there are no statistically significant differences in the overall image quality for

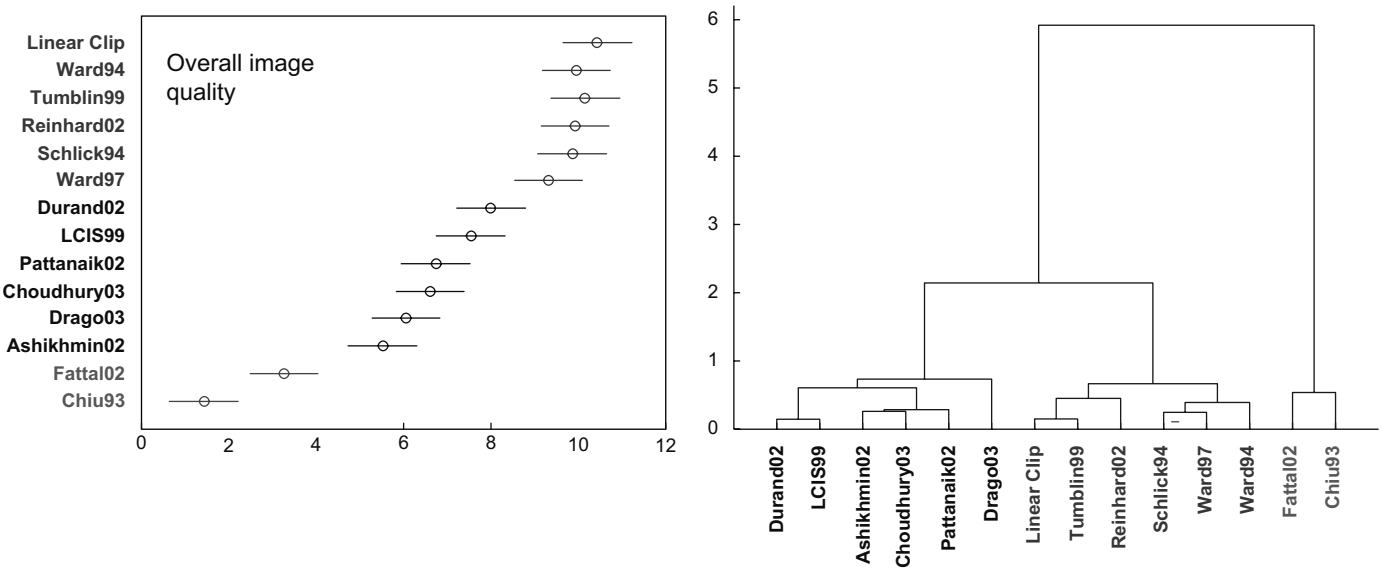


Fig. 10. Left: average overall image quality with confidence intervals. Circles show OIQ means with 95% confidence intervals (horizontal axis)—the higher value the better quality. Right: average Mahalanobis distances of overall image quality for all methods.

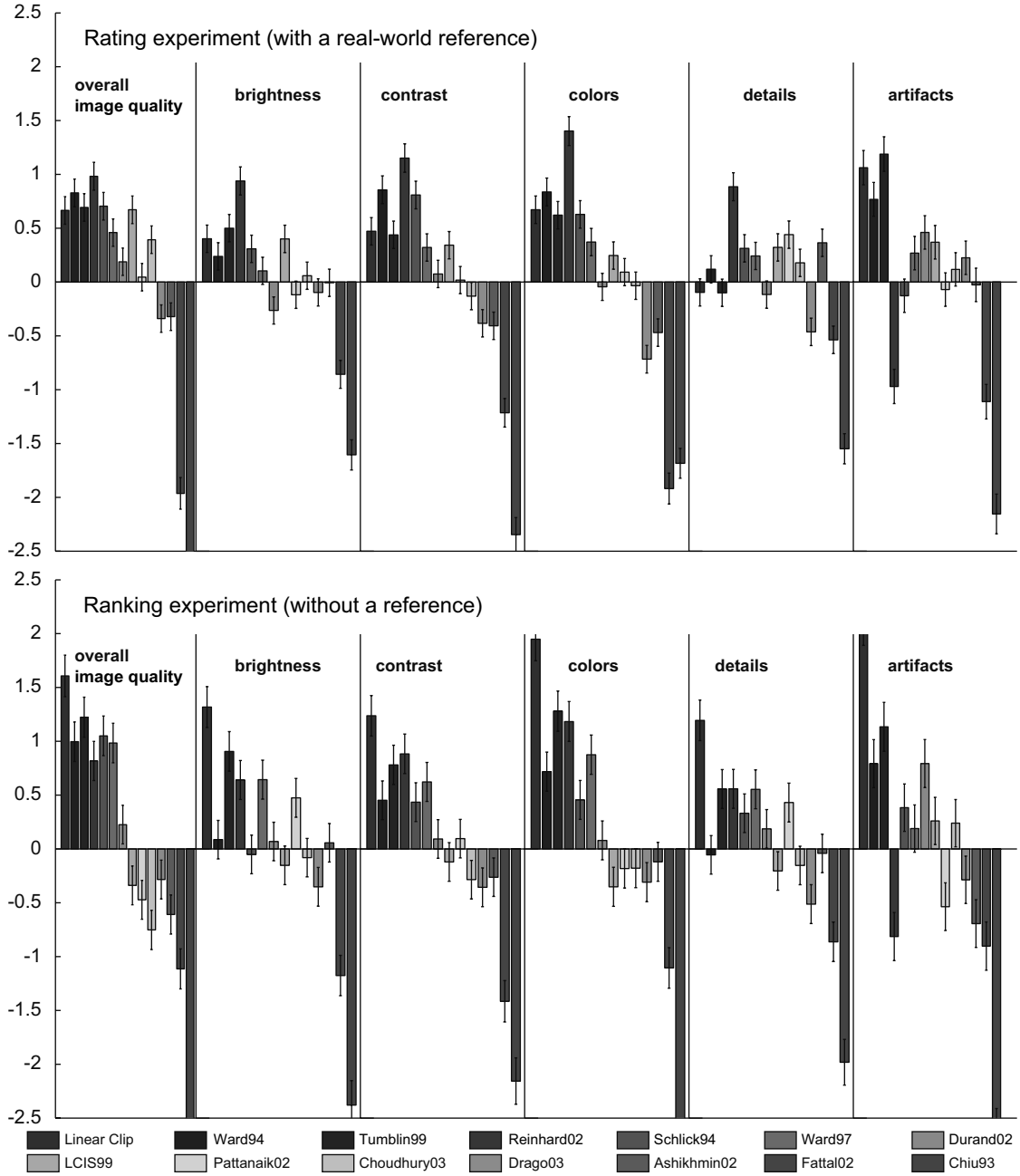


Fig. 11. Accuracy scores for rating (with a reference) experiment (top) and for ranking (without a reference) experiment (bottom) for all examined TM methods. Left to right: overall perceptual quality, reproduction of brightness, reproduction of contrast, reproduction of details, reproduction of colors, lack of disturbing artifacts. In each chart the higher value represents the higher reproduction quality.

the first six methods, which are largely the global TM methods (visualized in blue). The second group (black color) comprises in fact deeply local TM approaches that operate averagely in the perceptual dimension of HDR tone reproduction. Finally, in the third group (red color) are perceptually not satisfactory methods. In Fig. 10 (right) we show the dendrogram of distances of overall image quality between the enquired methods. This graph also shows the described clustering of the methods into three groups.

The evaluation of artifacts (the higher value the better quality, i.e., the less amount of artifacts) shows another interesting result. The approach by Reinhard et al. shows high variance in this attribute, because it produced two relatively good images, but one with very disturbing artifacts, see Table 4. Due to the nature of Reinhard's method, the artifacts could not be completely avoided.

7.3.2. Comparison of the two experiments

In Fig. 11 we show average results for the two performed experiments separately. These results indicate how well the methods performed in rating (with reference) and ranking (without a reference) experiments. Similarly to overall results, methods Chiu93 [43] and Fattal02 [38] performed constantly worst in both experiments. In the rating experiment, Reinard02 [34] exhibits the best scores in all attributes but the artifacts, where it is the third worst rated (alike in the ranking experiment). In the ranking experiment, the linear clipping exhibits constantly the best scores in all attributes.

Generally, the results exhibit similar trends for all the enquired attributes as suggested by statistical analysis in previous sections. The relations of two experiments for each image attribute are

visualized in Fig. 12 along with linear fit and coefficients of determination R^2 (R^2 is a measure of the global fit of the model; $R^2 = 1$ would indicate that the fitted model explained all variability, while $R^2 = 0$ indicates no linear relationship between the results of our two experiments.) The highest agreement between two experiments is for overall contrast, overall image quality, and for the lack of artifacts attribute. The lowest agreement is exhibited by the detail attribute and we deal with this result in the next section.

7.3.3. Comparison of the results for input scenes

Statistical analysis as reported in Section 7.2 suggests that even though our input scenes do not have a systematic effect on obtained results, there probably exist methods whose performance differs for particular scenes. To examine the effect of the input scenes on the results further, we show the overall image quality scores separately for each scene, see Fig. 13. We notice rather similar trends in results for the two outdoor scenes (outdoor and night scene), while the indoor scene exhibits a slightly different pattern. Since there is a book with tiny writing which dominates the indoor scene, perhaps, there is a higher stress on reproduction of details in this case.

Notice that methods visualized in shades of blue color perform very well for at least two scenes. Chiu93 and Fattal02 on the other hand perform constantly poorly over all scenes in both tests. Pattanaik02 shows interesting consistent behavior—it performs very well for the night scene, averagely for the indoor scene, and poorly for outdoor scene. In case of the indoor scene, LCIS99 and Choudhury03 show the highest discrepancy between rating and ranking experiments. In this case, subjects in the rating experiment perhaps put more stress to the detail attribute to the detriment of other attributes while subjects in the ranking experiment not that much. This is in accordance with results reported in Section 7.3.2, where the detail attribute exhibited the lowest agreement.

7.4. Overall image quality and relationships of attributes

Beyond the discussed results, we analyzed the *dependencies of overall image quality* on the quality of reproduction of the five evaluated perceptual image attributes. Our investigations are formulated by means of the experimental results in five-dimensional functions, namely as the dependence of the overall image quality on the brightness, the contrast, the color, the detail reproduction and the artifacts attributes.

We used different methods to fit functions to the attribute observation scores receiving the best approximation to the independently observed overall image quality. Using the simplest approach, *multivariate linear regression*, we obtained the following result:

$$OIQ = 0.07 \cdot Bri + 0.37 \cdot Con + 0.06 \cdot Det + 0.36 \cdot Col + 0.21 \cdot Art, \quad (1)$$

where OIQ is an overall image quality function, Bri, Con, Det, and Col, represent the quality of reproduction of brightness, contrast, details and colors, respectively, all in the interval of $[0, 1]$ (0 meaning the worst reproduction). Art denotes the artifacts attribute in the interval of $[0, 1]$ (1 meaning no artifacts). To state how well the model explains the data, we computed the coefficient of determination: $R^2 = 0.76$. The high value of R^2 shows in our case that the linear regression approach is reasonable (a satisfactory value of R^2 for psychophysical experiments is over 0.7). In the second step, we determined which of the attributes actually contributed to the model. For this, we used the p-values of each attribute:

$$p_{\mathrm{Bri}} = 0.8624, \quad p_{\mathrm{Con}} < 0.0001, \quad p_{\mathrm{Det}} = 0.0390, \quad p_{\mathrm{Col}} < 0.0001, \quad p_{\mathrm{Art}} < 0.0001.$$

The only p-value that is higher than the threshold 0.05 is the brightness attribute, which means that the reproduction of brightness does not significantly influence the model. Further-

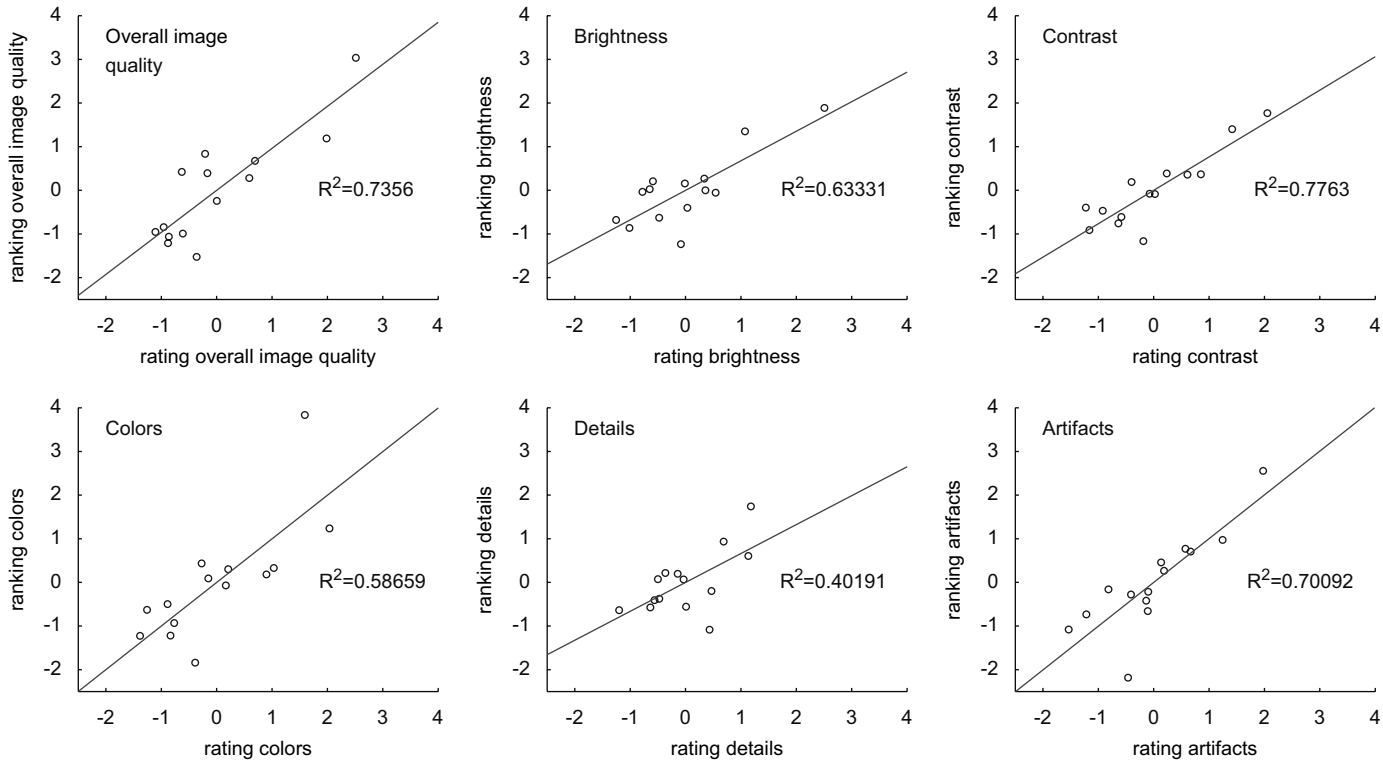


Fig. 12. Relations of the ranking experiment (vertical axes) and rating experiment (horizontal axes) interval scale results for all image attributes.

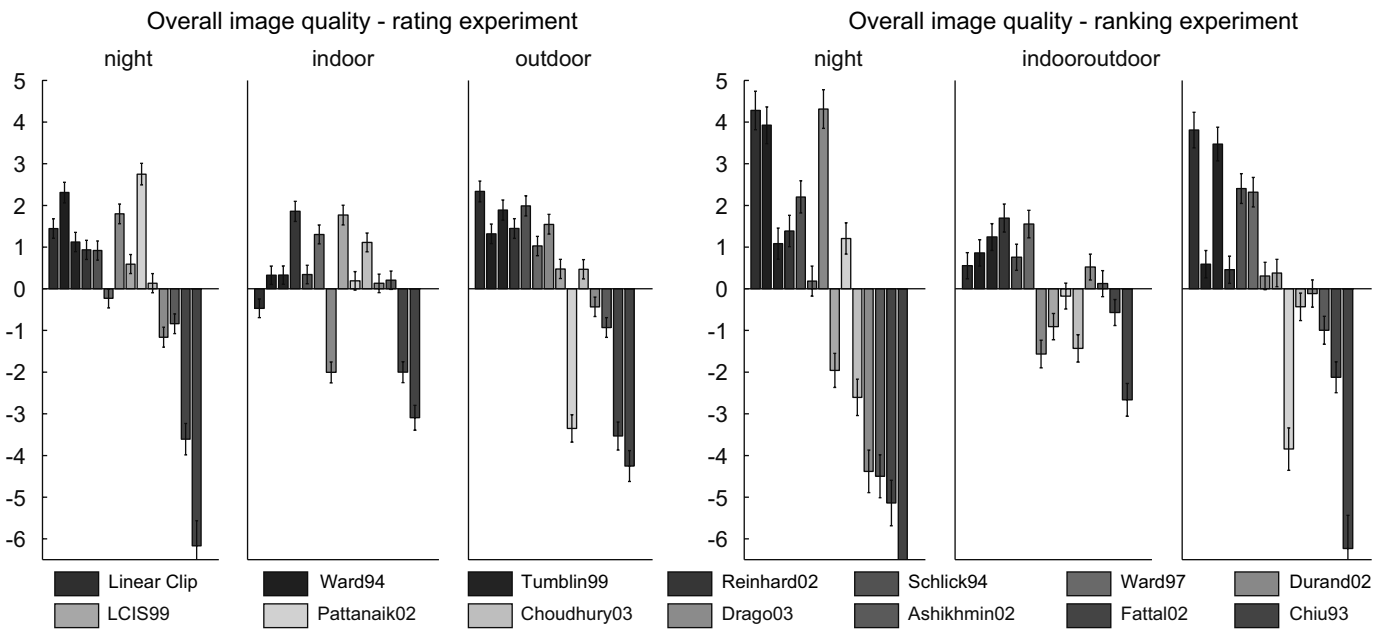


Fig. 13. Overall image quality scores for each input scene. Rating (with a reference) experiment results (left) and ranking (without a reference) experiment results (right) for all examined TM methods. Left to right: night scene, indoor scene, outdoor scene.

Table 10
Spearman correlations between the qualities of reproduction of image attributes

	OIQ	*Bri*	*Con*	*Det*	*Col*
Brightness (Bri)	0.58				
Contrast (Con)	0.80	0.64			
Details (Det)	0.66	0.60	0.66		
Colors (Col)	0.80	0.59	0.77	0.67	
Artifacts (Art)	0.65	0.43	0.55	0.55	0.56

more, we can observe in Eq. (1) that the *overall contrast* has the biggest weight factor and the detail reproduction the smallest one. This result may look surprising, as one would expect details to be more important. However, the global appearance of an image seems to depend much more on the quality of reproduction of other image attributes (contrast, color) and this confirms the good results of global TM methods as described in Section 7.3.

The low factor of *brightness reproduction* deserves special attention—it means that the brightness factor does not contribute to the proposed linear model. This could be caused by the fact that there is not a significant difference in reproduction of this attribute between the methods. However, we have found a significant difference in brightness already, see Section 7.1. To have another guideline, we computed the Spearman correlation coefficients between attributes, see the Table 10. These results show that there is a significant correlation between the brightness quality and the overall image quality. In the same time (not being in contradiction), Eq. (1) suggests that the impact of brightness quality spreads into the other attributes, it reveals itself only indirectly. This effect is perhaps the best example that the basic attributes are very coherent or inseparable. Incidentally, Eq. (1) shows *which attributes we should test* if we want to compare TM methods. There is no significant reason to evaluate the brightness since its effect is included in other attributes. The detail quality attribute shows a similar Spearman correlation coefficient and weight factor in formula (1) as the brightness. However, because of its very small p-value, it contributes directly to the overall image quality, in contrast to the brightness.

Finally, we used multiple linear regression to examine the image attribute relations (Fig. 6), with the following results:

$$Bri = 0.35 \cdot Con + 0.26 \cdot Det + 0.13 \cdot Col + 0.0004 \cdot Art$$

$$R^2 = 0.69, \quad p_{Con} < 0.0001, \quad p_{Det} < 0.0001, \quad p_{\mathrm{Col}} < 0.0001, \quad p_{\mathrm{Art}} = 0.99.$$

Since the p-value of artifacts is over the 0.05 threshold, this result implies that image artifacts do not contribute significantly to the perception of brightness quality.

$$Con = 0.22 \cdot Bri + 0.14 \cdot Det + 0.49 \cdot Col + 0.12 \cdot Art$$

$$R^2 = 0.67, \quad p_{\mathrm{Bri}} < 0.0001, \quad p_{\mathrm{Det}} < 0.0001, \quad p_{\mathrm{Col}} = 0.001, \quad p_{\mathrm{Art}} = 0.001,$$

$$Det = 0.25 \cdot Bri + 0.19 \cdot Con + 0.30 \cdot Col + 0.23 \cdot Art$$

$$R^2 = 0.56, \quad p_{\mathrm{Bri}} < 0.0001, \quad p_{\mathrm{Con}} < 0.0001, \quad p_{\mathrm{Col}} < 0.0001, \quad p_{\mathrm{Art}} < 0.0001,$$

$$Col = 0.10 \cdot Bri + 0.50 \cdot Con + 0.23 \cdot Det + 0.12 \cdot Art$$

$$R^2 = 0.66, \quad p_{\mathrm{Bri}} < 0.0001, \quad p_{\mathrm{Con}} < 0.0001, \quad p_{\mathrm{Det}} < 0.0001, \quad p_{\mathrm{Art}} < 0.0001,$$

$$Art = 0.08 \cdot Bri + 0.23 \cdot Con + 0.34 \cdot Det + 0.27 \cdot Col$$

$$R^2 = 0.39, \quad p_{\mathrm{Bri}} = 0.99, \quad p_{\mathrm{Con}} < 0.0001, \quad p_{\mathrm{Det}} < 0.0001, \quad p_{\mathrm{Col}} < 0.0001.$$

Due to rather small values of the coefficient of determination R^2 we cannot make a deeper observation from the above equations. However, they show evidence of the relations between the attributes and their approximate weight factors. Moreover, it is evident that the basic attributes are very hard to separate. As we predicted in Section 5, there are cross effects, or more complex basic factors, which are not directly observable. However, for the amount of observation data we have, the linear regression

approach is very reasonable and satisfactory, since we would need extremely large psychophysical experiments (with hundreds of subjects) for nonlinear fits with cross effects of image attributes.

7.5. Comparison to other studies

In this section, we discuss and relate our results to other studies. A complete direct comparison is not possible, because we have evaluated more methods than the previous studies, and the aims of particular studies were slightly different. We should emphasize that our study was targeted at the natural reproduction of real scenes. Since our experimental input data are bound to natural scenes, the global TM methods (and local methods with a proper global part) were generally ranked better than the "detail-hunting" and non-human vision-aware approaches.[7] Our results show that the quality of reproduction of overall brightness, overall contrast and colors is much more important than the reproduction of details when naturalness is ranked in real scenes.

Still, the *good performance of global methods* is perhaps the most surprising result of our study. However, this is in good accord with a recent psychophysical evaluation performed by Akyüz et al. [14], who show that outputs of sophisticated TM methods are statistically no better than the best single LDR exposure. Results of Yoshida et al. [5] also show distinctions between global and local methods, more specifically global methods performed better in the reproduction of brightness and contrast, while local methods exhibited better reproduction of details in bright regions of images. Even though Yoshida et al. claim that local methods perform better, we do not interpret their results so for the perceptual dimension, since (as one may see) in their results for naturalness (i.e., overall image quality) the first and the second best-rated (out of seven) methods are global TM curves (Ward97 and Drago03). In the results of Ledda et al. [12], two investigated global methods performed averagely, in favor of the iCAM [58] and Reinhard02 methods, but note that these methods are very strong in their global parts.[8] Looking at the results in the naturalness dimension reported by Drago et al. [4], we do not see the distinction between global and local methods, since Tumblin99 performs the best, but Ward97 is interestingly rated the worst. However, we should recall that observers did not have any reference in this experiment. Contrary to our results, Kuang et al. [10] report that local methods outperform global methods. However, basically the only global method that appears in their experiments is Ward97 with quite compelling results. To sum up: our results imply and we strongly believe that for a good performance in a perception targeted TM task, the TM method needs to have a significant global TM part. Then, the result may be sometimes enhanced using a local part that does not vanish in the global trend, e.g., [59].

The question of *correlation between the accuracy and preference experiments* is also very interesting. Ashikhmin and Goyal [11] demonstrate that using real environments is crucial in judging performance of TM methods and clearly show that there is a difference in subject's responses for a fidelity test with reference and without reference. Contrary to that, Kuang et al. [10] report a very strong correlation between the accuracy and preference experiments and state that one can use preference experiments in place of accuracy experiments with a real-world reference. Our results are perhaps closer to Kuang et al., since we did not detect statistically significant differences between the two performed experiments. However, our results do not exhibit as strong a correlation as that of Kuang for overall image quality, and specifically not for overall brightness and reproduction of the detail attributes.

Comparing *particular method performances* is quite tricky, since the results of TM methods may depend on implementation and used parameters. Our results are in good agreement with the evaluation performed by Drago et al. [4], where the Reinhard02 method was ranked the best and the Schlick94 method was also ranked quite well. The difference is in Ward97 (histogram-based approach), where authors deliberately omitted the human-based ceiling function (we did not) and therefore the method favors the reproduction of details at the expense of naturalness. The consequences of Kuang et al. [6,8] are also similar to ours: Fattal02 was considered not very natural while Reinhard02 (photographic mapping) was nearly the best ranked; we did not test iCAM. The only difference is with Durand02 (bilateral filtering method), which was ranked the best in Kuang's study (in our overall ranking Durand02 performed averagely). We believe this is caused by the implementation of the bilateral filter, since Kuang et al. use their specific modification of the original algorithm. In accordance with the original method description [9], we have compressed the base layer using a scale factor in the log domain. More plausible global compression would result in a positively better outcome, but we aimed to compare purely the original approaches. This supposition is also supported by the conclusions of Ledda et al. [12], where the bilateral filtering approach performed the worst while other overlapping methods show perfect agreement as well (in the overall similarity test). Similarly to our results, in Yoshida et al. [5], the best-natural rated method was Ward97, which is in accord with our results. The other results could not be compared easily, since Yoshida et al. tested the values (amount) of attributes while we inquired the reproduction quality of attributes.

8. Conclusions

In this article, we presented an overview of image attributes for TM that should facilitate access to the existing TM literature. Since the attributes are intimately related, we have proposed a scheme of relationships between them. Moreover, we have proposed a measure for the *overall image quality*, which can be expressed as a combination of these attributes based on psychophysical experiments. We have verified the proposed ideas by means of two different psychophysical experiments.

The presented overview of image attributes is helpful for getting into the TM field, or when implementing or developing a new TM method. On the other hand, the identification of the relationships between the attributes is very useful for the subjective comparison of TM methods. For example, we have found that overall brightness need not really be observed when the other attributes are available. It also simplifies the comparison process by reducing the actual number of attributes that can be used to evaluate a TM method. Finally, it represents the initial effort to design a truthful, objective comparison metric for HDR images.

Using the results of two different experimental studies, with three typical real-world HDR scenes and 14 different TM methods evaluated, this contribution presents one of the most comprehensive evaluations of TM methods yet. Although many interesting results in the field of local TM methods have been published, our results imply that the global part of a TM method is most

[7] Our results show that *statistically*, global techniques frequently outperform local TM approaches, even though local methods are generally claimed to perform better. Evidently this does not hold for all scenes, as can also be seen in our results. However, this is also a trend which matches our subjective personal experience.

[8] iCAM is generally a local method, but the adaptation values (for both luminance and colors) are calculated using a heavily blurred source image (very wide Gaussian), so that the method has a very strong global part and the method behaves to a big extent close to a global one.

essential to obtain good perceptual results for typical real-world scenes.

An interesting and important result of the two different testing methodologies used (rating with reference and ranking without reference) is that almost all of the studied image quality attributes can be evaluated without comparison to a real HDR reference.

The question remains how to numerically assess the quality of reproduction of particular image attributes. Although some approaches were proposed in literature [15,29], this area deserves further investigation and perceptual verification. In the future, we will conduct consequential tests targeted on individual image attributes to be able to computationally assess the overall quality of TM methods.

Acknowledgments

This work has been partially supported by the Ministry of Education, Youth and Sports of the Czech Republic under the research programs MSM 6840770014 and LC-06008, by the Kontakt OE/CZ Grant no. 2004/20, by the Research Promotion Foundation (RPF) of Cyprus IPE project PLHRO/1104/21, and by the Austrian Science Fund under contract no. P17261-N04. Part of this work was carried out during the tenure of an ERCIM "Allain Bensoussan" Fellowship Programme.

Thanks to all subjects at the CTU in Prague, the Intercollege Cyprus, and the MTA SZTAKI Budapest (thanks to Prof. Dmitry Chetverikov for helping in carrying out the experiments) that participated in the perceptual tests. Erik Reinhard provided a copy of their paper prior to its publication. Special gratitude to Jiří Bittner for his help in preparing this article and to the anonymous reviewers for their valuable comments.

References

[1] Devlin K, Chalmers A, Wilkie A, Purgathofer W. Star: tone reproduction and physically based spectral rendering. In: Fellner D, Scopignio R, editors. State of the art reports, eurographics 2002. The Eurographics Association; 2002. p. 101–23.
[2] Reinhard E, Ward LG, Pattanaik S, Debevec P. High dynamic range imaging: acquisition, display, and image-based lighting. Los Altos, CA: Morgan Kaufmann; 2005.
[3] Čadík M, Slavík P. The naturalness of reproduced high dynamic range images. In: Proceedings of the ninth international conference on information visualisation, Los Alamitos: IEEE Computer Society; 2005. p. 920–5.
[4] Drago F, Martens WL, Myszkowski K, Seidel, H-P. Perceptual evaluation of tone mapping operators. In: Proceedings of the SIGGRAPH 2003 conference on sketches & applications, GRAPH '03, New York, NY, USA: ACM Press; 2003. p. 1.
[5] Yoshida A, Blanz V, Myszkowski K, Seidel, H-P. Perceptual evaluation of tone mapping operators with real-world scenes. Human Vision & Electronic Imaging X, San Jose, CA, USA: SPIE;2005. p. 192–203.
[6] Kuang J, Yamaguchi H, Johnson GM, Fairchild MD. Testing HDR image rendering algorithms. In: Color imaging conference. 2004. p. 315–20.
[7] Kuang, J, Johnson, GM, Fairchild MD. Image preference scaling for HDR image rendering. In: Thirteenth color imaging conference. Scottsdale, Arizona. 2005. p. 8–13.
[8] Kuang J, Liu C, Johnson, GM, Fairchild MD. Evaluation of HDR image rendering algorithms using real-world scenes. In: International congress of imaging science, ICIS06. 2006.
[9] Durand F, Dorsey J. Fast bilateral filtering for the display of high-dynamic-range images. In: Proceedings of the 29th annual conference on computer graphics and interactive techniques. New York, NY, USA: ACM Press; 2002. p. 257–66.
[10] Kuang J, Yamaguchi H, Liu C, Johnson GM, Fairchild MD. Evaluating HDR rendering algorithms. ACM Transactions on Applied Perception 2007;4(2):9.
[11] Ashikhmin M, Goyal J. A reality check for tone-mapping operators. ACM Transactions on Applied Perception 2006;3(4):399–411.
[12] Ledda P, Chalmers A, Troscianko T, Seetzen H. Evaluation of tone mapping operators using a high dynamic range display. In: Proceedings of the 32nd annual conference on computer graphics and interactive techniques, SIGGRAPH '05. ACM Press; 2005, p. 640–8.
[13] Yoshida A, Mantiuk R, Myszkowski K, Seidel H-P. Analysis of reproducing real-world appearance on displays of varying dynamic range. Computer Graphics Forum 2006;25(3):415–26.
[14] Akyüz AO, Fleming R, Riecke BE, Reinhard E, Bülthoff HH. Do HDR displays support LDR content? A psychophysical evaluation. ACM Transactions on Graphics 2007;26(3).
[15] Janssen R, Computational image quality. Society of Photo-Optical Instrumentation Engineers (SPIE), 2001.
[16] Rogowitz BE, Frese T, Smith, JR, Bouman, CA, Kalin, EB. Perceptual image similarity experiments. In: Rogowitz BE, Pappas TN, editors. Proceedings of the SPIE, human vision and electronic imaging III, vol. 3299. 1998, p. 576–90.
[17] Fedorovskaya E, de Ridder H, Blommaert F. Chroma variations and perceived quality of color images of natural scenes. Color Research & Application 1997;22(2):96–110.
[18] Savakis A, Etz S, Loui A, et al. Evaluation of image appeal in consumer photography. Proceedings of the SPIE 2000;3959:111–20.
[19] Jobson D, Rahman Z, Woodell G. The statistics of visual representation. In: Rahman Z, Schowengerdt RA, Reichenbach SE, editors. Visual information processing XI, 2002, p. 25–35.
[20] Mantiuk R, Seidel HP. Modeling a generic tone-mapping operator. Computer Graphics Forum 2008;27(2), in press.
[21] Čadík M, Wimmer M, Neumann L, Artusi A. Image attributes and quality for evaluation of tone mapping operators. In: Proceedings of pacific graphics 2006. Taipei, Taiwan: National Taiwan University Press; 2006. p. 35–44.
[22] Neumann L, Neumann A. Gradient domain imaging. First EG workshop on computational aesthetics in graphics, imaging and visualization, 2005.
[23] Adelson EH. Lightness perception and lightness illusions. In: Gazzaniga M, editor. The cognitive neurosciences. Cambridge, MA: MIT Press; 1999. p. 339–51.
[24] Wyszecki G, Stiles WS. Color science, concepts and methods: quantitative data and formulae. 2nd ed. New York: Wiley; 1982 ISBN 0-471-02106-7.
[25] Fairchild MD. Color appearance models. 2nd ed. Chichester, UK: Wiley-IS&T; 2005.
[26] Tumblin J, Rushmeier H. Tone reproduction for realistic images. IEEE Computers Graphics and Applications 1993;13(6):42–8.
[27] Krawczyk G, Myszkowski K, Seidel, H-P. Computational model of lightness perception in high dynamic range imaging. In: Rogowitz BE, Pappas TN, Daly SJ, editors. Human vision and electronic imaging XI, IS&T/SPIE's 18th annual symposium on electronic imaging (2006). 2006. p. 1–12.
[28] Winkler S. Vision models and quality metrics for image processing applications, PhD thesis, EPFL, December 2000.
[29] Matkovic K, Neumann L, Neumann A, Psik T, Purgathofer W. Global contrast factor—a new approach to image contrast. In: Neumann L, Sbert M, Gooch B, Purgathofer W, editors. Computational aesthetics in graphics, visualization and imaging 2005. Eurographics Association; 2005. p. 159–68.
[30] Ward LG. A contrast-based scalefactor for luminance display. Graphics Gems 1994;IV:415–21.
[31] CIE, An analytical model for describing the influence of lighting parameters upon visual performance. vol. 1: technical foundations. CIE 19/2.1. International organization for standardization, 1981.
[32] Ferwerda JA, Pattanaik SN, Shirley P, Greenberg DP. A model of visual adaptation for realistic image synthesis. Computer Graphics 1996;30:249–58 (Annual Conference Series).
[33] Ward LG, Rushmeier H, Piatko C. A visibility matching tone reproduction operator for high dynamic range scenes. IEEE Transactions on Visualization and Computer Graphics 1997;3(4):291–306.
[34] Reinhard E, Stark M, Shirley P, Ferwerda J. Photographic tone reproduction for digital images. In: Proceedings of the 29th annual conference on computer graphics and interactive techniques. ACM Press; 2002. p. 267–76.
[35] Ashikhmin M. A tone mapping algorithm for high contrast images. In: 13th eurographics workshop on rendering. Eurographics Association; 2002. p. 145–56.
[36] Peli E. Contrast in complex images. Journal of the Optical Society of America A 1990;7(10):2032–40.
[37] Mantiuk R, Myszkowski K, Seidel H-P. A perceptual framework for contrast processing of high dynamic range images. In: Proceedings of the 2nd symposium on applied perception in graphics and visualization, APGV '05, New York, NY, USA: ACM Press; 2005. p. 87–94.
[38] Fattal R, Lischinski D, Werman M. Gradient domain high dynamic range compression. In: Proceedings of the 29th annual conference on computer graphics and interactive techniques. New York, NY, USA: ACM Press; 2002. p. 249–56.
[39] Pattanaik SN, Ferwerda JA, Fairchild MD, Greenberg DP. A multiscale model of adaptation and spatial vision for realistic image display. In: Proceedings of the 25th annual conference on computer graphics and interactive techniques. New York, NY, USA: ACM Press; 1998. p. 287–98.
[40] Reinhard E, Devlin K. Dynamic range reduction inspired by photoreceptor physiology. IEEE Transactions on Visualization and Computer Graphics 2005;11(1):13–24.
[41] Tumblin J, Turk G. Low curvature image simplifiers (LCIS). In: SIGGRAPH 99 conference proceedings. Annual conference series. Reading, MA: Addison Wesley; 1999. pp. 83–90.
[42] Choudhury P, Tumblin J. The trilateral filter for high contrast images and meshes. In: Proceedings of the 14th eurographics workshop on rendering, eurographics association, EGRW '03, 2003. p. 186–96.
[43] Chiu K, Herf M, Shirley P, Swamy S, Wang C, Zimmerman K. Spatially nonuniform scaling functions for high contrast images. In: Proceedings of graphics interface '93. 1993. p. 245– 53.

[44] Schlick C. An adaptive sampling technique for multidimensional ray tracing. In: Brunet P, Jansen FW, editors. Photorealistic rendering in computer graphics. Berlin: Springer; 1994. p. 21–9.
[45] Spencer G, Shirley P, Zimmerman K, Greenberg DP. Physically-based glare effects for digital images. In: Proceedings of the 22nd annual conference on computer graphics and interactive techniques. ACM Press; 1995. p. 325–34.
[46] Calabria AJ, Fairchild MD. Perceived image contrast and observer preference I: the effects of lightness, chroma, and sharpness manipulations on contrast perception. Journal of Imaging Science & Technology 2003;47:479–93.
[47] Debevec PE, Malik J. Recovering high dynamic range radiance maps from photographs. In: Whitted T, editor. SIGGRAPH 97 conference proceedings. Annual conference series, vol. 31 ACM SIGGRAPH, Reading, MA: Addison Wesley; 1997. p. 369–78. ISBN 0-89791-896-7.
[48] Drago F, Myszkowski K, Annen T, Chiba N. Adaptive logarithmic mapping for displaying high contrast scenes. Computer Graphics Forum 2003;22(3).
[49] Pattanaik S, Yee H. Adaptive gain control for high dynamic range image display. In: Proceedings of 18th spring conference on computer graphics, ACM Press; SCCG '02: 2002. p. 83–7.
[50] Tumblin J, Hodgins JK, Guenter BK. Two methods for display of high contrast images. ACM Transactions on Graphics 1999;18(1):56–94.
[51] Thurstone LL. A law of comparative judgement. Psychological Review 1927;34:278–86.
[52] Torgerson WS. Theory and methods of scaling. New York, NY, USA: Wiley; 1958.
[53] Siegel S, Castellan NJ. Nonparametric statistics for the behavioral sciences. 2nd ed. London: McGraw-Hill; 1988.
[54] Lehmacher W, Wall KD. A new nonparametric approach to the comparison of *k* independent samples of response curves. Biometrical Journal 1978;20:261–73.
[55] Rencher AC. Methods of multivariate analysis. 2nd Ed. Wiley series in probability and statistics, 2002.
[56] Anderson MJ, ter Braak CJF. Permutation tests for multi-factorial analysis of variance. Journal of Statistical Computation and Simulation 2003;73:85–113.
[57] Tabachnick BG, Fidell LS. Using multivariate statistics. 5th ed. Pearson Education, Inc.; 2007.
[58] Fairchild MD, Johnson GM, Kuang J, Yamaguchi H. Image appearance modeling and high-dynamic-range image rendering. In: Proceedings of the 1st symposium on applied perception in graphics and visualization, APGV '04, New York, NY, USA: ACM; 2004. p. 171.
[59] Čadík M. Perception motivated hybrid approach to tone mapping. In: Proceedings of WSCG (Full Papers). 2007. p. 129–36.

Computers & Graphics 32 (2008) 716–719

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Erratum

Erratum to "Evaluation of HDR tone mapping methods using essential perceptual attributes" [Comput. Graph. 32(3) (2008) 330–349]

Martin Čadík *

Department of Computer Science and Engineering, Czech Technical University in Prague, Karlovo nam. 13, 121 35 Prague, Czech Republic

There were errors in the published version of Tables 3–5 in this article. The publisher apologises for any problem caused as a result. The correct version of the tables appear here.

DOI of original article: 10.1016/j.cag.2008.04.003
* Tel.: +420 737 049 097; fax: +420 224 923 325.
E-mail address: cadikm@fel.cvut.cz

doi:10.1016/j.cag.2008.07.007

Table 3
Strengths and weaknesses of evaluated TM methods—indoor scene

Method	Image	Brightness	Contrast	Details	Colors	Overall Quality
Linear Clip		**10.6** *2.8*	**7.6** *3.9*	**7.6** *4.7*	**11.3** *3.6*	**8.9** *3.0*
		6.3 *3.1*	**6.6** *3.9*	**3.6** *3.7*	**8.4** *3.7*	**5.7** *2.9*
Ward94		**7.7** *3.0*	**8.1** *4.0*	**5.3** *1.9*	**9.6** *2.4*	**9.7** *3.7*
		7.1 *2.7*	**9.7** *2.2*	**6.8** *2.5*	**9.8** *2.5*	**7.9** *3.4*
Tumblin99		**11.1** *1.6*	**9.5** *3.3*	**7.5** *3.9*	**10.3** *1.9*	**10.8** *3.1*
		9.0 *1.8*	**9.6** *3.5*	**6.9** *1.9*	**10.4** *1.9*	**8.0** *2.5*
Reinhard02		**10.8** *1.9*	**11.6** *2.7*	**10.4** *2.8*	**12.5** *1.4*	**12.2** *1.1*
		11.9 *1.9*	**12.1** *2.6*	**11.8** *1.7*	**12.6** *1.8*	**11.9** *2.3*
Schlick94		**3.8** *2.4*	**7.1** *3.3*	**6.2** *3.8*	**5.6** *2.9*	**9.3** *3.1*
		6.9 *4.2*	**8.7** *4.3*	**6.7** *3.9*	**9.1** *3.9*	**8.2** *4.6*
Ward97		**8.8** *2.5*	**9.8** *3.3*	**8.1** *3.2*	**10.3** *2.3*	**11.5** *1.8*
		10.4 *2.3*	**9.7** *2.1*	**9.4** *1.9*	**10.0** *1.9*	**10.8** *2.2*
Durand02		**8.4** *3.7*	**4.7** *4.4*	**6.9** *4.0*	**4.6** *2.9*	**3.5** *2.7*
		2.9 *1.6*	**2.2** *0.8*	**5.0** *0.9*	**2.4** *0.9*	**2.75** *2.4*

Method	Image	Brightness	Contrast	Colors	Details	Overall Quality
LCIS99		**4.1** *1.5*	**6.2** *2.6*	**5.4** *3.9*	**3.4** *1.2*	**4.6** *1.3*
		11.2 *2.1*	**10.8** *1.9*	**9.8** *3.3*	**9.3** *3.3*	**11.9** *2.4*
Pattanaik02		**11.1** *3.6*	**8.9** *3.6*	**12.4** *2.5*	**8.6** *3.3*	**6.8** *3.3*
		8.2 *4.5*	**8.2** *3.9*	**9.4** *3.5*	**8.1** *3.5*	**7.6** *3.2*
Choudhury03		**5.2** *1.5*	**5.9** *2.3*	**7.0** *2.6*	**5.4** *1.4*	**2.2** *3.6*
		10.2 *2.2*	**8.8** *2.5*	**9.6** *2.3*	**7.7** *2.3*	**10.4** *2.0*
Drago03		**10.9** *1.5*	**9.5** *1.8*	**6.9** *3.9*	**9.0** *3.0*	**8.9** *1.5*
		7.5 *2.2*	**5.8** *1.9*	**5.1** *2.3*	**6.5** *2.3*	**7.2** *1.7*
Ashikhmin02		**8.3** *2.5*	**8.0** *3.7*	**10.2** *2.6*	**8.3** *3.2*	**7.6** *3.3*
		7.3 *2.9*	**6.5** *2.0*	**9.6** *2.6*	**5.0** *2.6*	**7.5** *1.3*
Fattal02		**3.2** *1.0*	**5.4** *3.6*	**7.4** *4.2*	**5.0** *1.8*	**5.8** *2.4*
		3.7 *2.7*	**3.8** *2.6*	**8.2** *2.7*	**2.4** *2.7*	**3.4** *3.4*
Chiu93		**1.1** *0.3*	**2.7** *3.0*	**3.0** *2.5*	**1.1** *0.3*	**1.8** *1.3*
		2.5 *1.9*	**2.7** *2.6*	**3.3** *1.6*	**3.5** *1.6*	**1.9** *0.9*

In bold: average ranking scores (1st line) and average rating scores (3rd line); in italics: standard deviations for ranking (2nd line) and for rating scores (4th line). The higher values represent the higher reproduction quality.

Table 4
Strengths and weaknesses of evaluated TM methods—outdoor scene

Method	Image	Brightness	Contrast	Details	Colors	Overall Quality
Linear Clip		**12.3**	**13.2**	**12.5**	**13.4**	**13.2**
		2.2	*0.9*	*1.9*	*0.5*	*0.9*
		10.3	**10.7**	**9.4**	**10.1**	**11.8**
		3.1	*2.9*	*3.6*	*3.6*	*2.5*
Ward94		**4.3**	**6.2**	**4.1**	**5.7**	**8.2**
		2.0	*2.6*	*3.1*	*1.4*	*1.8*
		7.4	**10.2**	**7.2**	**9.1**	**9.7**
		3.1	*2.3*	*2.3*	*2.3*	*2.2*
Tumblin99		**12.4**	**12.7**	**12.5**	**13.6**	**12.9**
		2.3	*1.2*	*1.9*	*0.5*	*1.0*
		10.3	**8.4**	**7.3**	**9.8**	**11.2**
		2.5	*2.1*	*4.1*	*3.4*	*1.5*
Reinhard02		**9.2**	**10.6**	**9.0**	**10.1**	**7.9**
		2.0	*0.9*	*1.6*	*1.1*	*3.2*
		13.1	**12.0**	**9.7**	**11.7**	**9.7**
		1.1	*1.7*	*4.6*	*1.9*	*2.9*
Schlick94		**9.4**	**10.6**	**10.7**	**10.9**	**11.5**
		2.7	*1.6*	*2.6*	*0.5*	*1.4*
		10.6	**10.3**	**9.7**	**9.8**	**11.1**
		2.0	*2.6*	*2.7*	*2.2*	*2.3*
Ward97		**10.6**	**11.6**	**12.3**	**11.4**	**11.3**
		2.1	*2.3*	*1.1*	*0.8*	*1.8*
		9.7	**9.0**	**8.9**	**8.7**	**8.9**
		2.9	*3.7*	*3.4*	*2.5*	*1.5*
Durand02		**3.6**	**6.0**	**4.8**	**5.6**	**7.7**
		2.0	*2.9*	*2.6*	*1.3*	*1.6*
		7.8	**10.4**	**7.0**	**9.0**	**10.1**
		3.4	*2.5*	*2.0*	*2.4*	*2.7*

Method	Image	Brightness	Contrast	Colors	Details	Overall Quality
LCIS99		**9.4**	**7.9**	**8.8**	**7.9**	**7.8**
		2.2	*1.6*	*1.5*	*2.0*	*1.7*
		7.8	**7.8**	**8.7**	**8.5**	**7.7**
		3.3	*3.2*	*2.6*	*2.6*	*3.0*
Pattanaik02		**6.1**	**4.1**	**3.4**	**2.1**	**2.1**
		4.1	*1.8*	*2.2*	*0.3*	*0.3*
		2.3	**2.1**	**4.8**	**3.1**	**2.3**
		0.8	*0.8*	*3.4*	*2.1*	*0.7*
Choudhury03		**8.2**	**5.8**	**7.9**	**7.3**	**6.0**
		1.7	*1.3*	*1.4*	*1.7*	*1.5*
		6.0	**6.1**	**7.8**	**7.1**	**7.7**
		2.4	*3.2*	*3.1*	*2.4*	*3.0*
Drago03		**3.6**	**4.1**	**5.1**	**5.4**	**6.9**
		2.1	*2.4*	*2.2*	*2.5*	*2.7*
		6.2	**6.8**	**6.1**	**5.0**	**5.7**
		2.3	*2.3*	*1.7*	*3.6*	*1.5*
Ashikhmin02		**8.7**	**6.9**	**7.7**	**6.8**	**4.9**
		2.5	*1.8*	*2.2*	*1.5*	*1.3*
		6.6	**6.0**	**7.9**	**7.1**	**5.2**
		2.4	*2.1*	*4.1*	*2.3*	*2.5*
Fattal02		**2.8**	**1.8**	**3.1**	**3.7**	**3.5**
		1.3	*1.1*	*1.4*	*1.9*	*1.2*
		4.4	**2.9**	**5.5**	**2.0**	**2.3**
		4.1	*0.9*	*4.7*	*0.7*	*0.9*
Chiu93		**4.4**	**3.5**	**3.1**	**1.1**	**0.3**
		3.8	*3.0*	*2.3*	*0.3*	*1.1*
		2.9	**2.6**	**5.3**	**4.4**	**1.8**
		2.2	*2.6*	*5.1*	*3.8*	*0.5*

In bold: average ranking scores (1st line) and average rating scores (3rd line); in italics: standard deviations for ranking (2nd line) and for rating scores (4th line). The higher values represent the higher reproduction quality.

Table 5
Strengths and weaknesses of evaluated TM methods—night scene

Method	Image	Brightness	Contrast	Details	Colors	Overall Quality	Method	Image	Brightness	Contrast	Colors	Details	Overall Quality
Linear Clip		**11.3** *3.6*	**12.8** *1.2*	**13.2** *1.3*	**12.2** *2.7*	**12.9** *1.0*	LCIS99		**6.5** *2.3*	**6.2** *1.5*	**5.8** *0.9*	**6.1** *2.7*	**5.7** *0.5*
		10.0 *3.0*	**9.8** *2.1*	**8.8** *3.3*	**10.4** *2.9*	**10.2** *3.1*			**7.6** *2.2*	**6.9** *2.9*	**7.5** *2.1*	**6.5** *2.0*	**7.7** *2.6*
Ward94		**10.6** *3.0*	**12.1** *1.8*	**12.1** *1.5*	**11.9** *2.2*	**12.5** *1.4*	Pattanaik02		**9.1** *2.0*	**10.0** *1.0*	**10.5** *1.8*	**9.6** *2.5*	**9.2** *1.6*
		10.2 *3.2*	**11.1** *3.2*	**9.6** *3.9*	**11.5** *3.0*	**11.9** *1.7*			**10.4** *3.9*	**12.2** *1.9*	**12.7** *1.3*	**11.7** *2.0*	**12.5** *1.8*
Tumblin99		**7.4** *2.3*	**7.6** *1.7*	**8.1** *1.0*	**8.4** *1.5*	**8.8** *1.3*	Choudhury03		**7.1** *2.6*	**6.6** *2.3*	**5.3** *0.9*	**5.9** *2.6*	**5.1** *1.0*
		8.8 *2.7*	**8.6** *3.1*	**6.8** *2.9*	**8.4** *3.4*	**9.4** *2.7*			**6.7** *1.9*	**5.6** *3.4*	**7.3** *2.9*	**6.5** *2.8*	**6.9** *2.5*
Reinhard02		**9.1** *3.6*	**9.0** *2.5*	**9.4** *1.3*	**9.7** *1.4*	**9.3** *1.3*	Drago03		**4.9** *4.2*	**4.9** *3.9*	**3.3** *0.6*	**3.8** *2.6*	**3.4** *1.2*
		6.8 *4.2*	**8.5** *2.4*	**9.8** *3.7*	**10.0** *2.3*	**8.7** *2.6*			**6.9** *3.9*	**4.9** *2.5*	**5.2** *3.0*	**3.9** *3.2*	**4.4** *1.9*
Schlick94		**8.8** *2.6*	**9.3** *2.1*	**9.7** *1.3*	**9.5** *1.8*	**10.4** *1.7*	Ashikhmin02		**5.1** *3.4*	**3.6** *1.1*	**3.6** *1.0*	**4.5** *2.7*	**3.3** *1.0*
		7.5 *3.1*	**10.9** *1.8*	**9.5** *2.5*	**9.3** *1.9*	**8.8** *1.9*			**8.2** *3.2*	**4.8** *2.2*	**8.5** *4.4*	**5.2** *2.6*	**4.8** *2.3*
Ward97		**8.7** *2.8*	**7.0** *1.8*	**7.8** *2.0*	**8.1** *1.4*	**7.7** *1.3*	Fattal02		**4.0** *2.5*	**2.4** *0.5*	**2.4** *0.7*	**2.6** *0.5*	**2.7** *0.6*
		3.7 *2.8*	**6.3** *2.0*	**6.5** *2.0*	**6.9** *2.0*	**5.9** *1.9*			**5.7** *4.8*	**3.7** *3.2*	**3.0** *3.0*	**2.6** *1.1*	**2.1** *0.4*
Durand02		**11.4** *2.5*	**12.5** *1.4*	**12.8** *0.4*	**11.7** *2.2*	**13.0** *0.6*	Chiu93		**1.0** *0.0*	**1.0** *0.0*	**1.0** *0.0*	**1.0** *0.0*	**1.0** *0.0*
		8.9 *3.7*	**10.9** *2.5*	**8.9** *2.9*	**10.9** *2.2*	**11.0** *2.4*			**3.9** *5.0*	**1.1** *0.2*	**1.2** *0.3*	**1.3** *0.6*	**1.1** *0.2*

In bold: average ranking scores (1st line) and average rating scores (3rd line); in italics: standard deviations for ranking (2nd line) and for rating scores (4th line). The higher values represent the higher reproduction quality.

Appendix B

Perception Motivated Hybrid Approach to Tone Mapping

M. Čadík. Perception Motivated Hybrid Approach to Tone Mapping. In *Winter School of Computer Graphics (WSCG Full Papers)*, pp. 129– 136, Pilsen, Czech Republic, 2007.

Perception Motivated Hybrid Approach to Tone Mapping

Martin Čadík
Department of Computer Science and Engineering, Czech Technical University in Prague
Karlovo náměstí 13, 121 35 Prague, Czech Republic
cadikm@fel.cvut.cz

ABSTRACT

We propose a general hybrid approach to the issue of reproduction of high dynamic range images on devices with limited dynamic range. Our approach is based on combination of arbitrary global and local tone mapping operators. Recent perceptual studies concerning the reproduction of HDR images have shown high importance of preservation of overall image attributes. Motivated by these findings, we apply the global method first to reproduce overall image attributes correctly. At the same time, an enhancement map is constructed to guide a local operator to the critical areas that deserve enhancement. Based on the choice of involved methods and on the manner of construction of an enhancement map, we show that our approach is general and can be easily tailored to miscellaneous goals of tone mapping. An implementation of proposed hybrid tone mapping produces good results, it is easy to implement, fast to compute and it is comfortably scalable, if desired. These qualities nominate our approach for utilization in time-critical HDR applications like interactive visualizations, modern computer games, HDR image viewers, HDR mobile devices applications, etc.

Keywords: Tone mapping, HDRI, dynamic range reduction.

1 INTRODUCTION

Merits of high dynamic range imaging (HDRI) are currently widely recognized in computer graphics, high-quality photography, computer vision, etc. However, HDRI becomes popular in interactive and real-time applications as well. Data visualization, computer games and other interactive applications gain new qualities thanks to HDRI. The reproduction of high dynamic (HDR) data on the low dynamic (LDR) output devices requires the reduction of dynamic range, commonly referred to as a tone mapping.

Many so-called tone mapping operators were proposed in history [Dev02, Rei05]. We can classify the existing approaches according to the transformation they apply to convert input luminances to the output values. *Global* tone mapping methods apply a tone reproduction curve - e.g. a function. Therefore, they transform particular value of the input luminance to one specific output value. *Local* tone mapping operators may on the other hand reproduce particular input luminance to different output values depending on the surrounding pixels.

Although many sophisticated local tone mapping operators were published, these are typically not very generic approaches and just a few of these methods is suitable for interactive and computationally weak applications. Even worse, hardly any of them can be marked as general and scalable. Besides, recent perceptual studies concerning the reproduction of HDR images [Yos05, Led05, Cad06, Cad07] have shown high importance of preservation of global image attributes (overall brightness, overall contrast).

Generally speaking, global methods reproduce overall image attributes well, they are fast to compute, and easy to implement, but may wash away important details. Local approaches excel in reproduction of local contrast (details), but they are computationally intensive and may reproduce overall image attributes poorly, see Figure 1. Motivated by the mentioned findings we present a fast and simple yet powerful general hybrid approach to tone mapping issue. This approach takes advantages of both global and local tone mapping approaches to overcome mentioned limitations. Moreover, since the aims of tone mapping can differ among particular applications, the proposed approach can be easily tailored to the miscellaneous goals.

The paper is organized as follows. In Section 2, we overview the previous work and we focus particularly on linear tone mapping methods. In Section 3, we introduce and describe generally the new hybrid tone mapping idea. In Section 4, we present two exemplary implementations of hybrid tone mapping approach and show the results. Finally, in Section 5, we conclude and give suggestions for future work.

2 RELATED WORK

The areas of high dynamic range imaging and the tone mapping are currently quite complex. Refer to the state of the art by Devlin [Dev02] or to the book by Reinhard et al. [Rei05] for an overview. Since we concentrate mainly on interactive applications and computationally

Figure 1: Global methods reproduce overall image attributes well, but wash away important details (left). Local approaches excel in reproduction of details, but may reproduce overall image attributes poorly (right).

efficient methods, we survey briefly the time-dependent methods suitable for interactive applications.

2.1 Time-dependent Tone Mapping

Adaptation mechanisms of human visual system show time-dependency. Human needs some time to adapt to low luminance levels after entering the dark cinema from the sunlit street, for example. However, involvement of time-dependency is profitable also for non-perceptual applications of tone mapping, since it avoids flickering and other time-dependent artifacts.

Probably the first time-dependent tone mapping method was presented by Pattanaik et al. [Pat00]. The method is based on statical color model by Hunt that is extended of time adaptation. Authors use global s-shaped curve for mimicking the response of both direct and inverse models of visual perception.

Durand and Dorsey [Dur00] used the global tone mapping operator by Ferwerda et al. [Fer96] for interactive tone mapping. Authors model visual adaptation course over the time for rods and cones.

Fairchild and Johnson [Fai03] adapted the iCAM model [Fai02] to account for time-dependent adaptation effects. The time-dependency provide two exponential filters that modify adaptation level. The used model is local (filtering using wide gaussian curve) and therefore computationally intensive. Since the filtering kernel is very large, the properties of the iCAM outputs resemble global tone mapping results.

Ledda [Led04] proposed strictly local time-dependent approach. The method is based on initial effort by Pattanaik, but adds local processing using bilateral filter. Time-dependency is modeled using exponential filters for rods and cones.

The above surveyed time-dependent approaches apply either global curves and thus they destroy subtle details or they apply local methods and thus they are computationally demanding. Moreover, interactive applications often need to do some sort of load balancing, however there is an unanswered question, how to scale the time-dependent methods properly.

2.2 Linear Tone Mapping

As we have noticed, global methods reproduce overall image attributes well, see Figure 1. The group of global methods comprise a subset of linear tone mapping curves. Despite of the simplicity of *linear* tone mapping curves, the approaches utilizing linear (or close-to linear) mapping have many advantages that deserve our attention.

Since the linear methods scale image intensities by a constant (scale factor), they do not change scene contrasts for display. This is probably the reason why these methods show [Cad06, Cad07] to perform well in perceptual reproduction of the overall image attributes.

Linear tone mapping methods transform the input HDR image to the output image values using the scale factor, $L_d = m \cdot L_w$, where m is the *scale factor*, L_w is the input luminance, and L_d is the output value in the interval of [0, 1].

The simplest linear approach is the *maximum luminance* mapping, where we map the maximal input luminance to the maximal output value (e.g. to 1): $m = \frac{1}{L_{wmax}}$, where L_{wmax} is the maximal input luminance. Since the maximal luminance is usually enormous in case of HDR images, this approach produces typically too dark and valueless results. *Mean value* mapping approach gives more reasonable outputs by mapping the *average* input luminance to the average output scale: $m = 0.5 \cdot \frac{1}{L_{avg}}$, where L_{avg} is the average input luminance.

Ward's contrast based scale factor [War94] focuses on the preservation of *perceived contrast*. The computation of the scaling factor is based on Blackwell's [CIE81] psychophysical contrast sensitivity model. Almost the same principle of contrast preservation is exploited also in the work of Ferwerda et al. [Fer96].

Another linear approach was introduced by Neumann et al. [Neu98]. They propose the minimum *information loss* method that tries to mimic the photographer's practice to lose a minimum amount of information. The method automatically selects ideal clipping interval to obtain a minimum of detail-lost areas. The automatic selection of the interval is done by means of logarithmic image histogram.

Mapping using *s-shaped curve* [Pat00] is formally not a linear approach, but practically it can produce results that are very close to the linearly mapped results. S-shaped curves resemble transfer curves of classical photographic media.

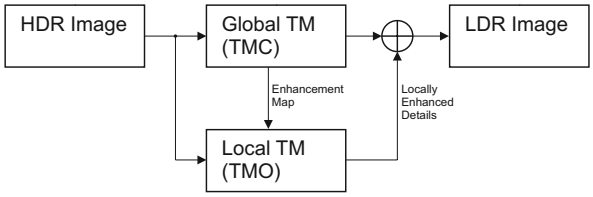


Figure 2: Overview of the hybrid tone mapping approach.

3 HYBRID TONE MAPPING

We propose a general idea of new hybrid tone mapping approach in this section, while two exemplary implementations are described in the next section. We do not concern the inherent physically-induced issues of real output devices (gamma correction, etc.), in this paper.[1] In the further text, we assume (without the loss of generality) that the output device range is linear and is limited to the interval of [0, 1]. Moreover, we do not concern the reproduction of colors here. All of our proposed methods aim to reduce dynamic range of the input image and the color information is left untouched. However, there is no obstacle to involve any method with a specific color processing into hybrid tone mapping approach if desired.

The overview of general hybrid tone mapping framework is shown in Figure 2. The process consists of three steps. First, the input high dynamic range image is transformed using global tone mapping curve. This mapping produces the base of the output low dynamic range image and (if necessary) outputs desired values for the construction of an *enhancement map*. In the second step, the enhancement map is constructed considering the aim of the tone mapping (see Figures 3, 6). Finally, the enhancement map is used to guide a local tone mapping method to reconstruct subtle details. The locally enhanced details are then added back to the globally transformed image to improve the final image while maintaining good overall reproduction of image attributes. Due to the enhancement map, the local tone mapping method is applied merely to the critical areas of the original image and therefore a lot of computational resources are saved.

3.1 Construction of Enhancement Map

Besides the choice of the involved methods, the manner of the construction of the enhancement map is an essential heart of the hybrid tone mapping. The enhancement map therefore has a profound effect on the properties of the hybrid tone mapping approach as a whole.

The enhancement map is generally a map of float numbers with the same dimensions as the original HDR image. In the examples shown in section 4, the enhancement map is constructed using a sort of thresholding of input luminance values. However, there is no obstacle to construct the enhancement map in a different way (for example based on input gradients, etc.). When we use a linear transform function in a global part of hybrid mapping, the construction of the enhancement map is very simple and effective. Since we know the analytical equation of the transfer curve, the clipping values are then directly known and the enhancement map is constructed effortlessly by thresholding. For the other approaches (s-shaped curves, log scaling, etc.), we propose to use the clip values of 5th and 95th luminance percentiles. However, the manner of the enhancement map construction can reflect a special goal of implemented method, as we show in section 4.3.

Figure 3: Example of an enhancement map (orange and green areas) for perceptually plausible hybrid approach. The luminances in orange-colored areas are potentially clipped due to linear mapping, while the details in dark parts, green-colored areas, are lost due to insufficient numerical precision.

Due to the abrupt changes of the global mapping to the local mapping (and vice versa), artifacts might appear on the borders of the enhancement map. To avoid the artifacts, we propose to construct and apply the enhancement map as follows (we show the process for only one clipping value L_{clip}):

$$EM(x,y) = \begin{cases} 0 & if\ TMC(L_w) < L_{clip} \\ \min(1, TMC(L_w) - L_{clip}) & otherwise, \end{cases}$$

where $EM(x,y)$ is the enhancement mask value, $TMC()$ is a global tone mapping method, L_w is the input luminance, and L_{clip} is the clipping luminance. The final output value L_d is then computed as a weighted sum of a global method and a local method outputs: $L_d(x,y) = EM(x,y) \cdot TMO(L_w) + (1 - EM(x,y)) \cdot TMC(L_w)$, where $TMO()$ is the involved local method.

If the computational time is not an issue, we can involve more sophisticated criteria to the construction of an enhancement map, e.g. the human visual system properties. Using the visual attention model, for example, we can pass the computational resources to the visually important areas of the image. Another possi-

[1] Nevertheless, all the tone mapping results presented in this paper have been gamma corrected finally, using the value of $\gamma = 2.2$ as usual.

bility is the usage of contrast sensitivity function (CSF) during the construction of the enhancement map. In this case, the effort of local method will be directed to the areas, where the detail is (at least potentially) visible for a human observer.

4 USE CASES AND RESULTS

The proposed idea of hybrid tone mapping is general – virtually any combination of existing (and potentially forthcoming) methods is possible. However, the choice of the involved methods has to reflect the intent of the resultant combination (f.e. an aesthetic view, a cognitively rich depiction, or a perceptually plausible reproduction). A combination of methods that is excellent in reproducing details can fail miserably when we aim in reproducing the perceptual experience of an observer. In this section, we show two different examples of the hybrid approach: perceptually plausible approach and cognitively rich approach. We show and discuss actual outputs of these methods and we also exhibit the performance values comparing to original local methods.

4.1 Fast Perceptually Plausible Approach

For such an interactive applications where the perceptually convincing reproduction is desired and where the computational resources are limited (e.g. in computer games), we propose fast and simple implementation of hybrid tone mapping approach as follows.

At the post of global method, we use the linear mapping by Ward [War94]. This method was proven to give reasonable results for natural scenes [Cad06, Cad07], and the computational demands of the method are minimal. Since the global part is purely linear, we can construct the enhancement map directly by thresholding of input luminance values. The exact clipping values are known: $L_{clipLO} = 0$ and $L_{clipHI} = 1$ for an original method. We can modify the approach by shifting the transform curve if desired (to allow the user to adjust brightness or contrast), but even in this case, the clipping values are easily found analytically. Pixels with luminance values outside of the linear interval would be clipped and therefore the information would be lost there. Therefore, these pixels form the enhancement map.

Having the enhancement map, we run the bilateral filtering method [Dur02] just on the areas marked in the map. Bilateral filter separates the original luminance map to the base layer and the detail layer. We use the detail layer to enhance the result of the global tone mapping method. For acceleration of the local filtering, we utilize graphics hardware (GPU) [Fia06]. Figure 4 compares the transforms of the original methods and the hybrid approach and Figure 5 shows the results in the form of output images.

4.2 Time-Dependent Hybrid Mapping

It is usually advantageous to model the course of visual adaptation over the time for interactive applications. Time-dependency of tone mapping is twofold profitable: it increases the perceptual quality on one hand, and it also avoids temporal image artifacts on the other hand.

The way of implementation of time-dependency is influenced by the goal of the whole tone mapping method and it is not necessary to realize it at all, in some cases. In accordance with other authors [Dur00, Pat00] we use an exponential decay function in our perceptually plausible hybrid approach (described in Section 4.1) to model the light adaptation. We omit the simulation of long-term dark adaptation due to its subtle and slow effect and due to efficiency reasons. We modulate the adaptation level $L_{a(w)}$ in the Ward's method [War94] by the exponential function for smooth transitions when tone mapping a dynamic environment:

$\frac{dL_{a(w)}}{dt} \approx \frac{L_{a(w)} - L_{a(w)(t)}}{\tau}$,

where $L_{a(w)}$ is the visual adaptation for static image, $L_{a(w)}(t)$ is the actual adaptation and $\tau = 0.1$ is a time constant that mimics the speed of adaptation.

Similar approach is amenable in many other hybrid tone mapping implementations, since we can usually smoothen the response of particular parameter of involved global tone mapping method. If the computational cost is the main limitation, the time-dependency may be omitted temporarily with reasonable loss of reproduction quality.

4.3 Cognitive Approach

As an example of cognitive (e.g. detail-oriented) hybrid tone mapping approach, we propose the combination of histogram adjustment global tone mapping operator [War97] enhanced by locally applied bilateral filtering [Dur02].

The histogram adjustment method grants most of the available device contrast to the areas of abundant luminance values in the input HDR image. Generally speaking, large areas in an input image are given more contrast (thus subtle details present at these areas may become visible) at the expense of tiny areas. This advanced 'distribution of contrast' is achieved thanks to cumulative function derived from formerly constructed image histogram. The cumulative function is then used as a tone reproduction curve to transform input luminance to output values.

In accordance with the choice of involved methods, we propose to construct also the *enhancement map* seeking the same goal of cognitively rich (detailed) output image. We detect the areas of small local contrast on the chart of cumulative function constructed in the previous step – these areas represent pixels, where the detail is potentially vanished. We use the second derivative of cumulative function for this detection (note that

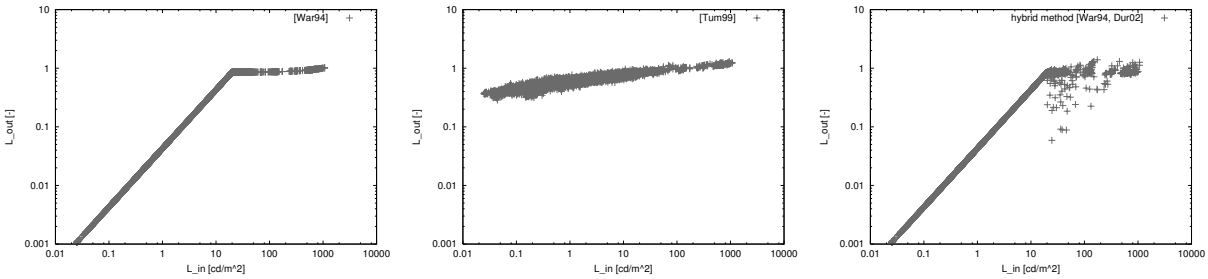


Figure 4: Tone mapping transforms, left: global method [War94] maps input HDR values via linear function – note the clipping of high luminances, middle: local method [Tum99] applies different transform to different pixels – the reproduction of overall image attributes is poor, right: hybrid approach [War94, Dur02] combines merits of both the global and local approaches.

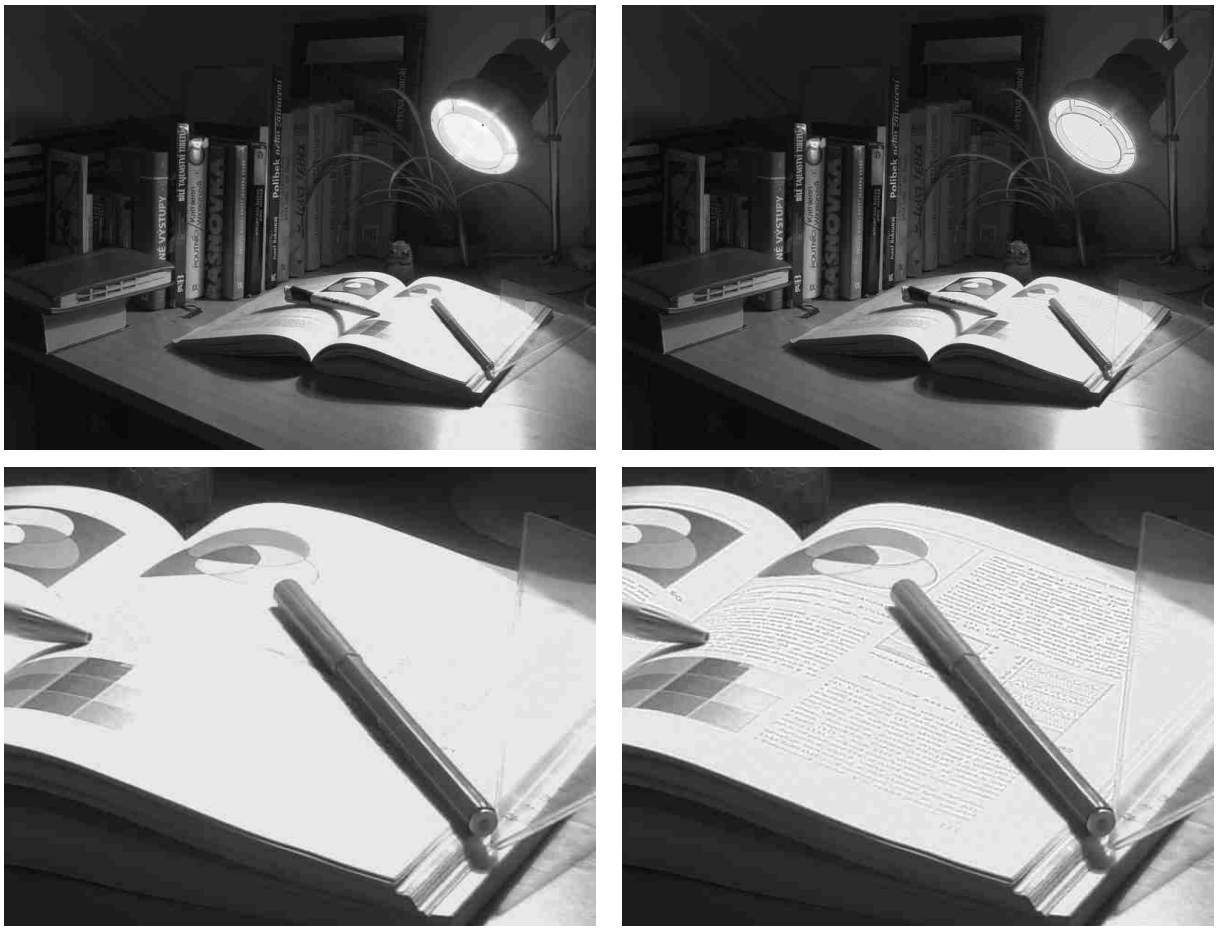


Figure 5: top left: result of purely global method [War94] exhibits well reproduced overall contrast, however shows the lack of subtle details, top right: result of the new hybrid approach [War94, Dur02] preserves the overall contrast accurately, and adds the lost details. Bottom: close-ups of the book, note the reproduced details in the hybrid approach result (bottom right).

Perceptual method (Sec. 4.1)		Cognitive method (Sec. 4.3)	
Enhancement map [% of image pixels]	Speedup [-]	Enhancement map [% of image pixels]	Speedup [-]
1.4e-3%	**118,5**	0.132 %	**41,74**

Table 1: Comparison of performances of two different implementations of hybrid tone mapping (average results over 10 HDR images). The speedup value shows the acceleration of hybrid approach against the original, completely local approach.

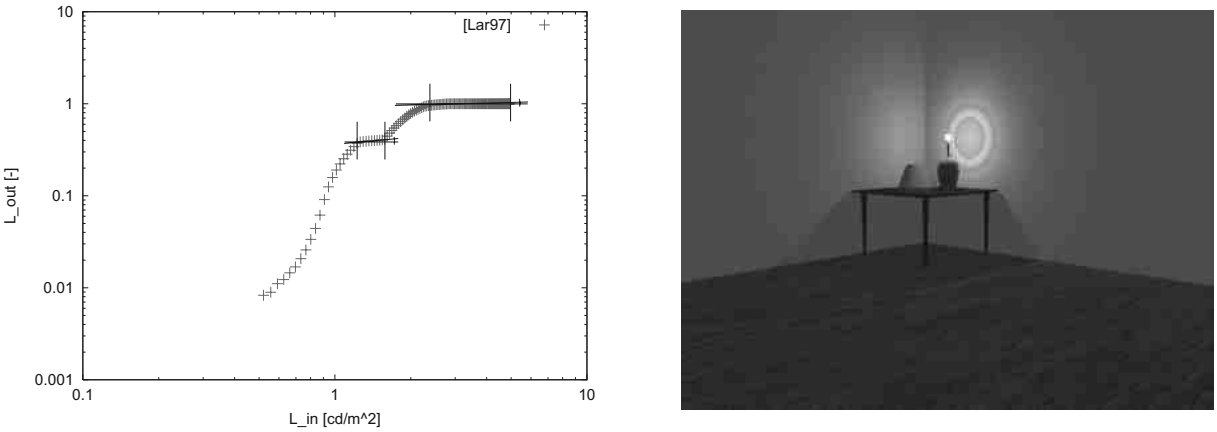


Figure 6: Cognitive hybrid approach. Left: cumulative function [War97] with detected areas for local enhancement. We use the threshold of 0.1 for detection, e.g. $L'_{out} < 0.1$. Right: corresponding enhancement map constructed from the cumulative function.

the the derivative is never a negative number in our case). Sections that show low slope (low values of the second derivation) are selected to construct the enhancement map (see Figure 6) – each pixel with the luminance value within the boundary of a selected section becomes a member of enhancement map.

Finally, we apply the local trilateral filtering method [Cho03] on the original input pixels that are present (or masked) in enhancement map. Similarly to the bilateral filter, the trilateral filter produces blurred image, but preserves significant luminance edges. We obtain local details by dividing the original image by the blurred image. Finally, we enrich the result of histogram adjustment method by these details. See Figure 7 (bottom line) where two renditions of one input image are shown to compare the two presented hybrid implementations.

4.4 Performance Results

Since one of the goals of the hybrid approach is the reduction of computational complexity, we present here the numerical performance results, see the Table 1[2]. The table shows average values measured at the group of 10 input HDR images. The perception-targeted fast hybrid approach is on average 3 times faster than the cognitive method and more that 118 times faster than the original bilateral filtering method. The reported speedup is gained thanks to the enhancement map. Since the enhancement map contains usually just a small portion of the original image pixels (as shown in the table), the time-demanding local approach is applied locally, to the small (necessary) part of the image.

Generally speaking, our technique places very small additional load to the system leaving large space for other computations. This is very profitable in interactive applications like the computer games, etc. However, note that besides the performance improvement, hybrid tone mapping can enhance the quality of the output image as well (see Figure 5).

In the imminent future, we can expect the need of dynamic range reduction even on various portable devices and on small and computationally elementary machines. The hybrid tone mapping will be reasonable in this case as well, thanks to its *good scalability*. If we face the lack of computational power, we can modify (soften) the criteria of the construction of the enhancement map. Depending on these criteria, we are able to continuously balance the computational load spanning from the complete locally enhanced method up to the factual omitting of the local enhancement (e.g. purely global tone reproduction). Finally, the other possibility to decrease the time consumption is to omit the time-dependent processing, as noted in Section 4.2.

[2] The performance values strongly depend on the selection and the implementation of the involved methods and we therefore present the average values for an overview.

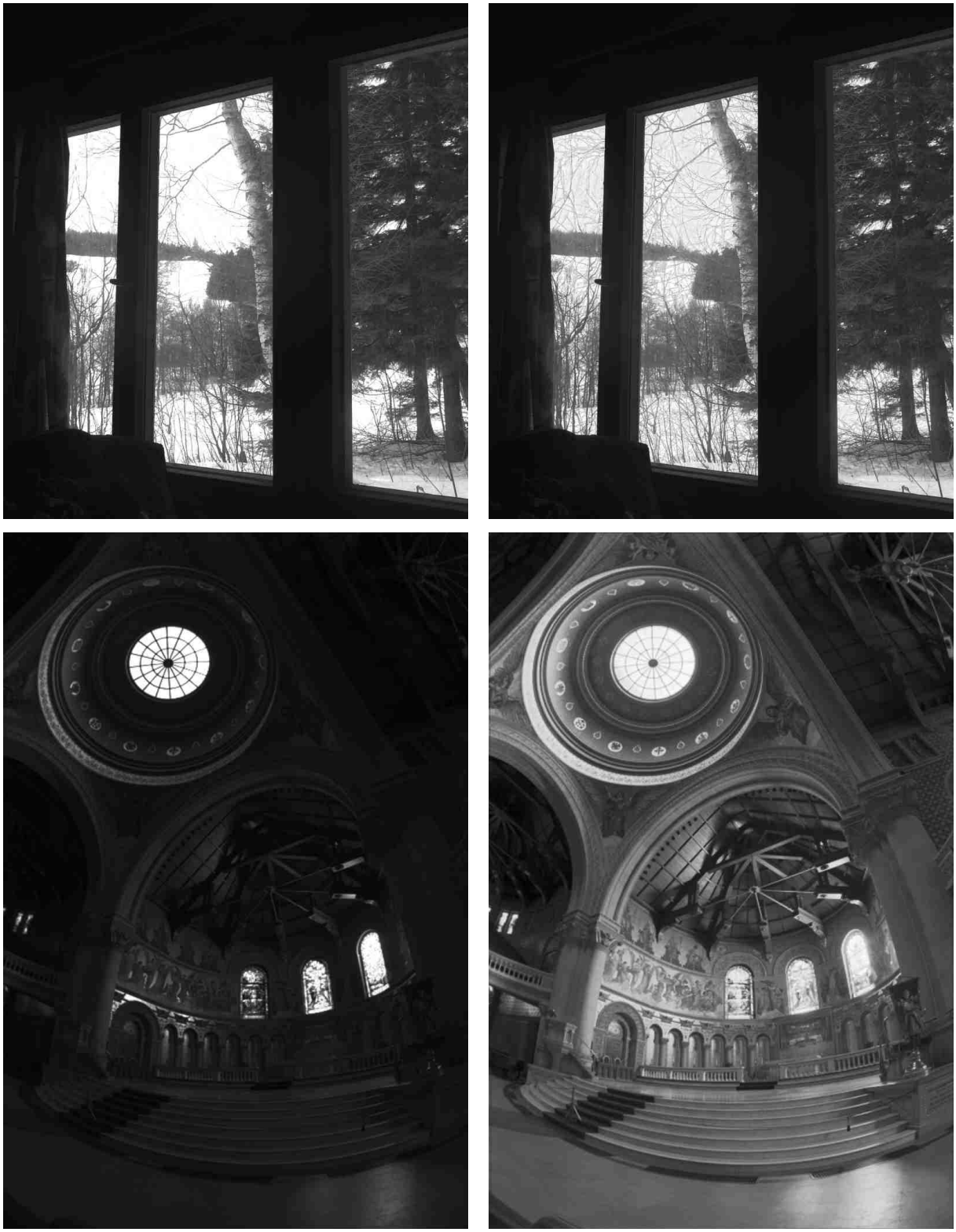


Figure 7: Further results of the new hybrid approach. Top left: pure global method [War94] without enhancement washes away details seen through the window. Top right: hybrid approach [War94, Dur02] enhances the image (note the revival of the birch twigs) without affecting the overall image attributes. Bottom left: hybrid approach [War94, Dur02] – perceptually plausible reproduction of well-known Memorial church image, bottom right: cognitive rendition of the same input image by hybrid combination of histogram adjustment and trilateral filtering [War97, Cho03].

5 CONCLUSIONS AND FUTURE WORK

We presented a novel hybrid approach to the issue of reproduction of high dynamic range images on devices with limited dynamic range of luminance (e.g. tone mapping issue). In our approach, we combine results of arbitrary global and local tone mapping operators. Recent perceptual studies concerning the reproduction of HDR images have shown high importance of preservation of overall image attributes. Motivated by these studies, we apply the global method first to reproduce overall image attributes correctly. At the same time, we construct an enhancement map to guide a local operator to the critical areas that deserve enhancement.

We do not invent another complex tone mapping method, but we rather propose a general framework that utilizes already known ideas and combines existing and potentially forthcoming methods. We have shown that the presented hybrid approach can be easily tailored to miscellaneous potential goals of tone mapping (e.g. to get perceptually plausible images, to get detail-rich depictions, etc.). Our experiences indicate that an implementation of proposed hybrid tone mapping approach typically produces reasonable results, it is easy to implement, fast to compute and it is comfortably scalable, if desired. These qualities nominate our approach for utilization in time-critical HDR applications like interactive visualizations, modern computer games, HDR image viewers on mobile devices, etc.

The perception of image attributes depends partially on the semantics of the input image or scene. Therefore every, even a subtle modification of an image can affect the quality of reproduction of an attribute (in both positive and negative sense). In the future, we will conduct subjective perceptual experiments to uncover and to quantify the effect of particular local enhancement method (in relation to the manner of enhancement map construction) on the quality of reproduction of image attributes.

ACKNOWLEDGEMENTS

This work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic under the research programs MSM 6840770014 and LC-06008 (Center for Computer Graphics). Thanks to Jiří Bittner for proof-reading the manuscript.

REFERENCES

[Cad06] M. Cadik, M. Wimmer, L. Neumann, and A. Artusi. Image attributes and quality for evaluation of tone mapping operators. In *Proceedings of Pacific Graphics 2006*, pages 35– 44, Taipei, Taiwan, 2006. National Taiwan University Press.

[Cad07] M. Cadik, M. Wimmer, L. Neumann, and A. Artusi. Evaluation of HDR tone mapping methods using essential perceptual attributes. *Submitted to Journal of Visual Communication and Image Representation*, 2007.

[Cho03] P. Choudhury and J. Tumblin. The trilateral filter for high contrast images and meshes. In *EGRW '03: Proceedings of the 14th Eurographics workshop on Rendering*, pages 186–196. Eurographics Association, 2003.

[CIE81] CIE. *An Analytical Model for Describing the Influence of Lighting Parameters upon Visual Performance*, volume 1: Technical Foundations, CIE 19/2.1. International Organization for Standardization, 1981.

[Dev02] K. Devlin, A. Chalmers, A. Wilkie, and W. Purgathofer. STAR: Tone reproduction and physically based spectral rendering. In D. Fellner and R. Scopignio, editors, *State of the Art Reports, Eurographics 2002*, pages 101–123. The Eurographics Association, September 2002.

[Dur00] F. Durand and J. Dorsey. Interactive tone mapping. In *Proceedings of the Eurographics Workshop on Rendering*. Springer Verlag, 2000. Held in Brno, Czech Republic.

[Dur02] F. Durand and J. Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 257–266. ACM Press, 2002.

[Fai02] M. D. Fairchild and G. M. Johnson. The iCAM framework for image appearance, image differences, and image quality. In *IS&T/SID Color Imaging*, 2002.

[Fai03] M. D. Fairchild and G. M. Johnson. Image appearance modeling. In *SPIE/IS&T Electronic Imaging Conference*, 2003.

[Fer96] J. A. Ferwerda, S. N. Pattanaik, P. Shirley, and D. P. Greenberg. A model of visual adaptation for realistic image synthesis. *Computer Graphics*, 30(Annual Conference Series):249–258, 1996.

[Fia06] O. Fialka and M. Cadik. FFT and convolution performance in image filtering on GPU. In *Tenth International Conference on Information Visualisation*, pages 609– 614, Los Alamitos, 2006. IEEE Computer Society Press.

[Led04] P. Ledda, L. P. Santos, and A. Chalmers. A local model of eye adaptation for high dynamic range images. In *AFRIGRAPH '04: Proceedings of the 3rd international conference on Computer graphics, virtual reality, visualisation and interaction in Africa*, pages 151–160, New York, NY, USA, 2004. ACM Press.

[Led05] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen. Evaluation of tone mapping operators using a high dynamic range display. In *ACM SIGGRAPH 2005, LA*. ACM Press, August 2005.

[Neu98] L. Neumann, K. Matkovic, and W. Purgathofer. Automatic exposure in computer graphics based on the minimum information loss principle. In *Proceedings of Computer Graphics International (CGI '98)*, pages 666– 679. IEEE Computer Society Press, 1998.

[Pat00] S. Pattanaik, J. Tumblin, H. Yee, and D. Greenberg. Time-dependent visual adaptation for fast, realistic image display. In *SIGGRAPH 2000 Conference Proceedings*, Annual Conference Series, pages 47– 54. ACM SIGGRAPH, Addison Wesley, 2000.

[Rei05] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. Morgan Kaufmann, 2005.

[Tum99] J. Tumblin and G. Turk. Low curvature image simplifiers (LCIS). In *SIGGRAPH 99 Conference Proceedings*, Annual Conference Series, pages 83–90. Addison Wesley, 1999.

[War94] G. Ward. A contrast-based scalefactor for luminance display. *Graphics Gems IV*, pages 415– 421, 1994.

[War97] G. Ward, H. Rushmeier, and C. Piatko. A visibility matching tone reproduction operator for high dynamic range scenes. *IEEE Transactions on Visualization and Computer Graphics*, 3(4):291–306, 1997.

[Yos05] A. Yoshida, V. Blanz, K. Myszkowski, and H. P. Seidel. Perceptual evaluation of tone mapping operators with real-world scenes. *Human Vision & Electronic Imaging X, SPIE*, 2005.

Appendix C

Visual Maladaptation in Contrast Domain

D. Pająk, M. Čadík, T. O. Aydın, K. Myszkowski, and H.-P. Seidel. Visual maladaptation in contrast domain. *Human Vision and Electronic Imaging XV*, Vol. 7527, No. 1, p. 752710, 2010.

Visual Maladaptation in Contrast Domain

Dawid Pająk[a,b], Martin Čadík[b], Tunç O. Aydın[b], Karol Myszkowski[b], Hans-Peter Seidel[b]
[a]WPUT; [b]MPI Informatik

ABSTRACT

In this work we simulate the effect of the human eye's maladaptation to visual perception over time through a supra–threshold contrast perception model that comprises adaptation mechanisms. Specifically, we attempt to visualize maladapted vision on a display device. Given the scene luminance, the model computes a measure of *perceived* multi–scale contrast by taking into account spatially and temporally varying contrast sensitivity in a maladapted state, which is then processed by the inverse model and mapped to a desired display's luminance assuming perfect adaptation. Our system simulates the effect of maladaptation locally, and models the shifting of peak spatial frequency sensitivity in maladapted vision in addition to the uniform decrease in contrast sensitivity among all frequencies. Through our GPU implementation we demonstrate the visibility loss of scene details due to maladaptation over time at an interactive speed.

Keywords: maladaptation, visual perception, contrast processing, human vision, temporal adaptation, high dynamic range

1. INTRODUCTION

It is consciously experienced by everyone that intense changes in illumination temporally cause a loss in visual sensitivity that is later recovered over a time period. In fact, considering the highly variant and temporally changing real world illumination, the human visual system (HVS) is virtually never fully adapted in practice. Due to this *maladaptation*, the visibility of some scene regions are reduced which would otherwise be perfectly visible if the HVS was fully adapted.

The temporal loss of visibility can often be tolerated in daily life, since a large fraction of sensitivity is recovered relatively fast in just a few seconds through neural mechanisms, and most real world objects are purposely designed to be strongly visible. However, some tasks require quick reaction times and undiverted attention. For those the rate of adaptation may not be sufficient. For instance, a car driver entering a forest highway after driving against the sun can be temporarily blinded for a short amount of time jeopardizing safety. In fact, computational methods have been proposed to determine the magnitude of vehicle display visibility under dynamic lighting conditions[1,2] enabling the validation of vehicle ergonomics and safety at design time. A more extreme case are fighter pilots who are exposed to much more drastic illumination changes, but regardless need to maintain near instant reaction capability at all times. On the other hand, the quickly recovered sensitivity may not be sufficient in environments containing low contrast objects. As an example, people often struggle to find their seats if they enter a movie theatre after the session started, while during the course of the movie the obstacles in the room become gradually visible due to the additional sensitivity recovery through the slower adaptation mechanisms based on chemical processes.

The aforementioned examples can benefit greatly from faithfully simulating the effect of maladaptation on visibility. Such a model should predict the visibility magnitude of both near– and supra–threshold scene details. Recent luminance based models[3–5] tend to explicitly focus on modeling maladaptation while ignoring other HVS aspects such as contrast sensitivity and visual masking. The modeling of the latter mechanism[6] requires a contrast based approach involving a transducer function. Current contrast domain frameworks, however, often do not account for luminance adaptation and contrast sensitivity, as well as the overall sensitivity loss and shift in peak sensitivity due to maladaptation.

Further author information: e-mail: {dpajak, mcadik, tunc, karol, hpseidel}@mpi-inf.mpg.de
[a] Westpomeranian University of Technology, Szczecin; [b] Max Planck Institut Informatik, Saarbrücken

An important consequence of maladaptation is the locality of resulting visibility loss in a scene. For instance, looking outside the window in a dark room on a sunny day, one will eventually adapt to the bright illumination outdoors and start seeing objects clearly. If at that instance the gaze is directed towards the interior, the observer will not be able to discriminate objects that are visually less apparent. Thus, at any given time details in some scene regions are less visible than others, as dictated by the current level of maladaptation. The problem is that it is often not possible to predict the direction the gaze will be shifted towards, and thus the illumination levels that will be observed in the next timestep. Similar to real–world scenes, the emerging HDR displaying technology coupled with the ever increasing size of display devices is also prone to such local losses of visibility. From an application perspective it is beneficial to simulate how the entire scene would look like under current adaptation conditions, which is not possible using current methods relying on a single adaptation level for the entire scene.

We present a system that renders a series of images of a scene as it would be seen by a maladapted eye over time. Each separate image corresponds to the visual perception of the scene at a time step while the sensitivity is recovered. The time course of adaptation is modeled by considering both neural mechanisms and pigment bleaching and regeneration. Our framework operates in contrast multi–scale domain and models supra–threshold effects like visual masking, while also accounting for contrast sensitivity and luminance (mal)adaptation usually considered only in luminance domain frameworks. We also model the shifting of peak frequency sensitivity in maladapted vision, which has not been considered by previous models. In the rest of the paper we first discuss related work (Section 2), followed by a new model for simulation of human maladaptation in contrast domain (Section 3). Next, we present, analyze and discuss the results of our system (Section 4) and finally we conclude and suggest ideas for future research (Section 5).

2. BACKGROUND

Previous models of time-course adaptation often operate on luminance and are not able to simulate visual phenomena locally. In this work, our goal is to simulate local maladaptation in *contrast domain* to account for supra–threshold mechanisms of vision as well as near–threshold. There were a few elaborate models of contrast perception proposed in history, but a vast majority of those were not concerned with the simulation of the time–course of maladaptation.

Ferwerda et al.[7] presented a computational model of visual adaptation. Their model captures the changes in threshold visibility, color appearance, visual acuity, and sensitivity over time using Ward's scaling tone mapping approach.[8] Ward's mapping is enriched by an offset parameter that is a function of time. The visual acuity is approximated by removing higher frequencies according to Schaler's measurements. This is a simplistic approach because human sensitivity to contrast also decreases for lower frequencies. A photoreceptor-based global time–dependent tone mapping method presented by Pattanaik et al.[4] is built on parts of an advanced Hunt's model of color vision.[9] By means of the adaptation model the method accounts for time dependency of retinal adaptation mechanisms for both cones and rods. However, as this adaptation model and the method itself are global they can simulate neither local adaptation mechanisms nor human contrast sensitivity. Irawan et al.[5] devised a model of low vision that is able to simulate the performance of an impaired or aged human visual system. The model is based on the combination of histogram adjustment[10] and Pattanaik et al.'s[4] global tone mapping methods. Due to the maladapted threshold-versus-intensity function (*tvia*), it can mimic the viewer's changing adaptation. The method is able to simulate the effect of maladaptation, but only at the threshold level and only globally for the whole image. However we are also interested in the supra–threshold effects of maladaptation in addition to visual perception around the threshold level.

Pattanaik et al.[3] proposed an advanced multiscale model of adaptation and spatial vision. As the model is based on spatial decomposition it can predict spatial contrast sensitivity behavior. The authors proposed gain functions that should be valid both for near– and supra–threshold luminance levels. However, the model does not comprise the time course of adaptation and is therefore unable to simulate effects of maladaptation. More recently, Mantiuk et al.[11] proposed a multiscale framework for perceptual processing of contrast. The method simulates supra–threshold perception (compression) of contrasts on multiple scales using transducer functions. However, contrasts still need to be compressed in a response space and yet another and more artificial compression is accomplished by the optimizer due to its weighting coefficients. The output of the optimization

problem solver is therefore hard to calibrate for the correct reproduction of luminance perception. Moreover, the method in fact does not simulate human adaptation.

HVS models involving maladaptation have been also been proposed in the context of detecting the visibility loss on display devices due to dynamically changing illumination.[12] While in this work temporal maladaptation is modeled in the contrast domain, they consider global adaptation and only output a "visibility map" that depicts distortions in the image structure similar to image quality assessment metrics, instead of rendering images of the scene appearance in a maladapted state. Furthermore, they don't model the change in spatial frequency sensitivity due to maladaptation, which we discuss in detail in Section 3.2.

2.1 Human Contrast Sensitivity in Maladapted State

Vision literature concerning the modeling human spatial *contrast sensitivity in an adapted state* usually through a *contrast sensitivity function* (CSF) is rich.[13] Much work has also been done on *temporal contrast sensitivity*,[14] i.e. the sensitivity of HVS to the spatial frequencies over time, as this (and so called critical flicker frequency) was crucial in the design of first CRT display devices. However, measurements of CSF in maladapted states are hardly that obvious, perhaps due to the complicated testing and evaluation process. Maladapted luminance intensity thresholds are measured only for simple stimuli without any variation of spatial frequency.[15] Consequently, in the rest of this section we discuss findings on the shape of CSF in maladapted conditions.

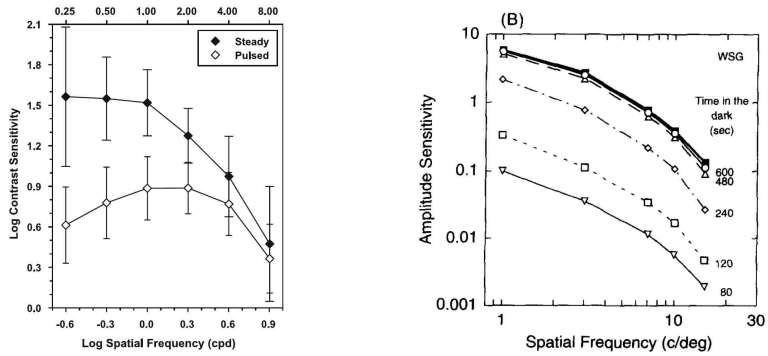


Figure 1. Measurements of maladapted contrast sensitivity. Left: the shape of CSF for steady (adapted) state and for briefly pulsed (maladapted) stimulus, adapted from[16], right: amplitude sensitivity functions during dark adaptation, adapted from [17] (the right image shows the amplitude sensitivity functions (ASF); one can obtain CSF from ASF by multiplying with the background luminance).

The encoding of contrast within the human visual system is thought to be mediated by two processing streams: the magnocellular (M) and parvocellular (P) pathways.[18] To investigate the effect of the pathways, Lenova et al.[19] and Alexander et al.[16] measured contrast sensitivity using two different paradigms. In the steady–pedestal paradigm, they briefly presented a test stimulus against a continuously presented adaptation field. In the pulsed–pedestal paradigm, the test stimulus was presented simultaneously with the adapting field. The steady–pedestal paradigm favors the M pathway, while the pulsed–pedestal paradigm favors the P pathway. The measured mean contrast sensitivity function for control subjects for steady–pedestal has a low–pass shape, while for the pulsed–pedestal it has a band–pass shape, see Fig. 1 (left).

On the other hand, Hahn et al.[17] found the CSF to be invariant in shape during dark adaptation. Differently from Leonova et al.[19] who presented stimuli only briefly to observers during experimentation, Hahn et al. measured a longer time course of dark adaptation ranging from seconds to hundreds of seconds, see Fig. 1 (right). This suggests that the transition from original to destination stimuli is very fast in terms of sensitivity to spatial frequencies (as modulated by P pathway), but much slower in terms of overall sensitivity to contrast. In other words, the shift in frequency sensitivity happens almost instantly and is retained during the time course of adaptation to the destination stimulus.

In our method, we use Daly's CSF[20,21] and we were tempted to simulate the aforementioned transition behavior by using current adaptation luminance as input parameter L_a of the maladapted observer. This

approach, combined with the use of *maladaptation ratio* resulted in reasonable time course shape of the CSF. (The maladaptation ratio is approximated as $cvi(L_b)/cvia(L_b, L_a)$, where *cvi* and *cvia* are the contrast versus intensity functions for adapted and maladapted eye, respectively, L_a is the current adaptation luminance, and L_b is the current background luminance[12]). However, this method implies the assumption that the spatial frequency sensitivity characteristics of the HVS remains constant in maladapted states, since the CSF we use was measured for the adapted eye. We can neither calibrate nor justify this approach as we were not able to find a sufficient amount of maladapted CSF experimental measurement data.

Therefore, in our model we incorporate the shift in frequency sensitivity due to maladaptation to the maladaptation ratio approach. Following an abrupt illumination change, we instantly modify the shape of the CSF to reflect the spatial frequency sensitivity in the target state, and then increase the sensitivities globally using the maladaptation ratio over the time course of adaptation. Our method is supported by experimental evidence: the sensitivity after sudden illumination change drops down drastically and when it is (at least partially) regenerated the curve already has the invariant shape of the target (compare Fig. 1 right with Fig. 5 right).

3. SIMULATION OF VISUAL MALADAPTATION

The data flow of the proposed model for human contrast perception in maladapted states is illustrated in Fig. 2 for the steady–state. We assume that the input HDR image is calibrated in cd/m^2 units. First, we construct background luminance and local adaptation maps (L_b, L_a), which are used both for contrast processing and final display purposes. The adaptation map is modified over time to model the temporal adaptation. Simultaneously, we decompose the input image into the contrast representation (C) using the Laplacian pyramid.[22] We then process physical contrast by a model of *maladapted scene observer* depending on sensitivity to spatial frequencies as well as on the current adaptation state to get the perceptual contrast responses (R). The response values are transformed by inverse *adapted display observer* model to obtain physical display contrast. All contrast processing steps are performed on multiple scales simultaneously. Consecutively, the physical contrast is converted to display luminance map (L_d) and colors are processed (I_{corr}). Finally, the inverse display model produces the output code values (I_{out}) that are shown on the display device.

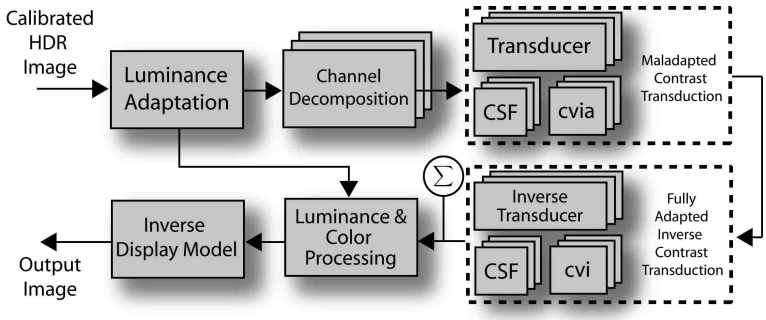


Figure 2. Flow chart of the proposed method. See text for details.

3.1 Adaptation Map

The adaptation map L_a represents the actual state of local adaptation of the observer. The construction of the *local adaptation* map is based on the actual *background luminance map* L_b and on the previous course of local adaptation (see Section 3.4 for the details on temporal adaptation). Background luminance L_b is the actual stimulus of an observer and is calculated for each input frame as the blurred image of input luminance (as the contrast sensitivity function was measured for foveated vision we blur the luminance conformably to one visual degree $(1°)$[10]). To accomplish this we use the Gaussian filter with the kernel size $K = \frac{2d}{p}\tan(\frac{\pi}{360})$, where p is the pixel size (in meters) and d is the observer's distance from the display (in meters).

Similarly to Irawan et al.[5] we model the adaptation due to human rods and cones separately. To obtain a single response value, Hunt [9, Sec. 31.8.2] proposed to sum the achromatic cone and rod responses up. The

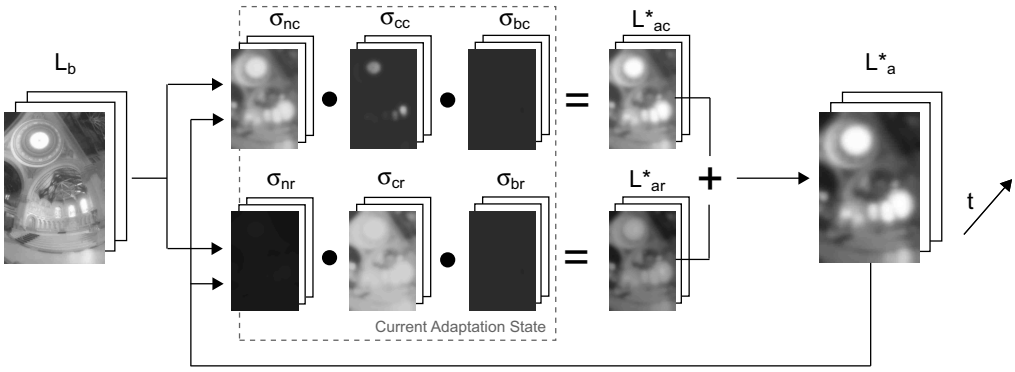


Figure 3. Visual illustration of the local adaptation processing. Temporal behavior is modeled for rod L^*_{ar} and cone L^*_{ac} adaptations separately, which are computed from adaptation maps $(\sigma_{bc}, \sigma_{cc}, \sigma_{nc}, \sigma_{br}, \sigma_{cr}, \sigma_{nr})$ that model various adaptation mechanisms. These adaptation maps are updated at each timestep using the current background luminance map L_b. In maladaptation computation (see Sec. 3.2), a compound adaptation map L^*_a is obtained by adding adaptation maps for rods and cones. HDR image courtesy of Paul Debevec.

current adaptation luminance L^*_a is therefore obtained (as also illustrated in Fig. 3) as a sum of cone (L^*_{ac}) and rod (L^*_{ar}) adaptations: $L^*_a = L^*_{ac} + L^*_{ar}$, where $L^*_{ac} = \sigma_{bc} \cdot \sigma_{cc} \cdot \sigma_{nc}$ and $L^*_{ar} = \sigma_{br} \cdot \sigma_{cr} \cdot \sigma_{nr}$. Factor σ_b accounts for the *photopigment bleaching and regeneration*: $\sigma_b(L_b) = 1/p(L_b)$, where $p(L) = I_0/(I_0 + L)$ and $I_0 = 10^4$ cd/m^2. To model *neural adaptation mechanisms*, we calculate σ_n (fast neural adaptation) and σ_c (slow neural adaptation) for rods and cones using the equations proposed by Irawan et al.[5] Note however that our implementation of human adaptation is local (i.e., we have the adaptation map) and all of the factors mentioned above $(L^*_{ac}, \sigma_{bc}, \sigma_{cc}, \sigma_{nc}, L^*_{ar}, \sigma_{br}, \sigma_{cr}, \sigma_{nr})$ are not single values, but complete maps spanning the whole image.

For the subsequent processing (CSF filtering, *cvi*, *cvia* functions), we need to convert the adaptation values $L^*_{\cdot}$ scaled in hypothetical *perceptual adaptation* units back into the physical units. In other words, we are searching for an adaptation map L_a in physical luminance units that would evoke the actual maladapted state L^*_a in the observer's visual system. To do this, we numerically invert the function L^*_a and set $L_a = L^{*-1}_a(L^*_a(L^*_{ac}, L^*_{ar}))$. Note that for the fully adapted observer this results in $L_a = L_b$ as expected, but for the maladapted observer, the behavior of this function is more complex (see Section 3.4).

Figure 4. Comparison of the effect of global and local adaptation. Left: global adaptation (using global values L_{ag} and L_{bg}), right: local adaptation (using local L_b and global L_{ag}). Notice that local background luminance map allows to simulate different sensitivity to spatial contrasts according to varying illumination in the scene.

For experimental visual analysis and illustration purposes we allow the use of the *global adaptation* value L_{ag} instead of the local adaptation map. We can calculate the global background luminance L_{bg} as a geometric mean of the input luminance L for each pixel: $L_{bg} = (\Pi^n L)^{1/n}$ and similarly we obtain the global adaptation

luminance L_{ag}. Global L_{ag} is useful for the analysis of static images, where it would be hard to change local adaptation map L_a manually if a refence HDR image depicting the adaptation state is not present. Note that in the rest of the figures the background luminance (L_b) is still local even in global adaptation (L_{ag}) case (see Fig. 4-right), with the exception of Fig. 4-left where we illustrate global adaptation L_{ag} with global background L_{bg} luminance for comparison.

3.2 Maladapted Spatial Sensitivity to Contrast

To account for sensitivity to spatial frequencies, we utilize the contrast sensitivity function (CSF) proposed by Daly.[20,21] The corresponding spatial frequency ρ (in c/deg) for each level l (starting from 1) of Laplacian pyramid is obtained as $\rho = K/2^{(l-1)}$. The size of the image (in $X \times Y$ pixels) in visual degrees is $i^2 = \max(X, Y)/K$. Given spatial frequency ρ in c/deg, observer distance d in meters, image size i^2 in visual degrees, and current background luminance level L_b (in cd/m^2), and neglecting orientation and eccentricity we can calculate the sensitivity S_a for contrast magnitudes C (in Weber's units) for each pixel at each level l of pyramid:

$$S_{al} = CSF(\rho, \theta, L_{bl}, i^2, d, C_l), \tag{1}$$

where the coarser background luminance map $L_{b(l+1)}$ is a downsampled from the finer scale map L_{bl}. We account for maladaptation by computing the maladaptation fraction as given in:[12]

$$S_{ml} = S_{al} \cdot \frac{cvi(L_{bl})}{cvia(L_{bl}, L_{al})}, \tag{2}$$

where S_{ml} is the sensitivity in the maladapted state, S_{al} is the sensitivity at the fully adapted state, L_{bl} is the current background luminance and L_{al} is the current adaptation luminance. The subscript l indicates the scale of each map.

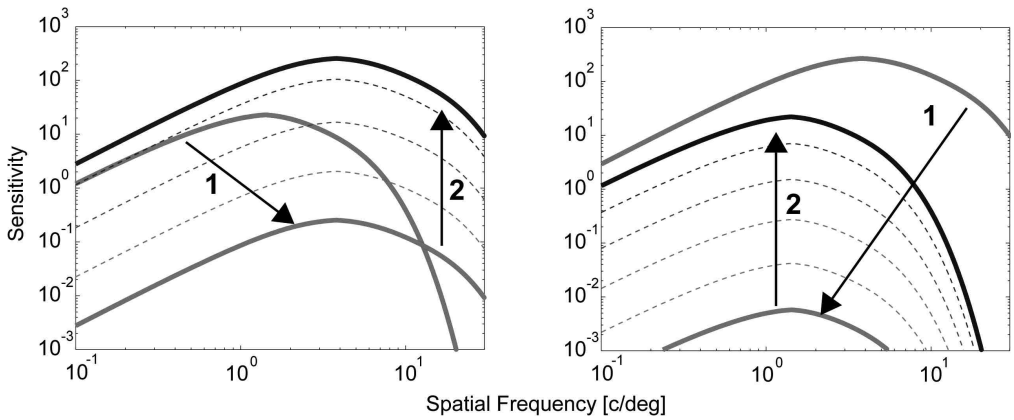


Figure 5. Simulation of time-course of contrast sensitivity for a maladapted observer. Left: transition from dark to bright, Right: transition from bright to dark environments (dark adaptation).

Fig. 5 shows the change in shape of CSF between two adapted states. In the left image, a subject is adapted to a dark environment. Accordingly, her sensitivity to contrast is low and shifted to low spatial frequencies (blue curve). After the exposition to a bright environment, the sensitivity rapidly shifts towards higher frequencies (arrow 1), but due to the maladaptation (as one is blinded by strong light for some time) the sensitivity is still very low (green curve). However, sensitivity is restored over time (arrow 2) to reach the final fully adapted state for the bright environment (red curve). The process is similar for a subject adapted to the bright environment (red curve in Fig.5 right). First, the sensitivity drops rapidly (arrow 1), shifts to the low frequencies (green curve) and consecutively it regenerates (arrow 2) to the final dark–adapted state (blue curve). The described behavior is in accord with psychophysical experiments conducted by Hahn and Geisler [17, Fig. 5, 6], who measured that the CSFs are nearly identical throughout the course of dark adaptation. Naturally, the two processes differ in the speed of the sensitivity regeneration and we describe our implementation of the temporal aspects of adaptation below.

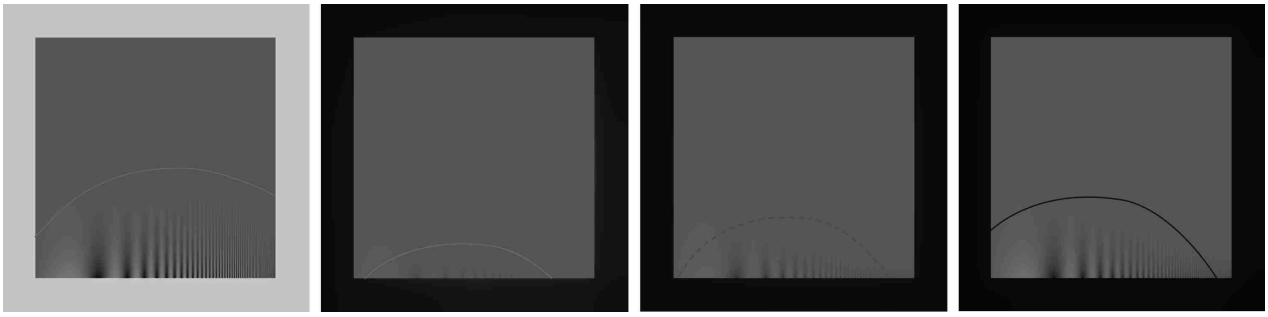


Figure 6. Classical Campbell-Robson contrast sensitivity chart for dark adaptation. From left to right: (1) fully adapted state in a relatively bright environment (adaptation luminance 112 cd/m^2), (2) background luminance was decreased to 2 cd/m^2, the contrast sensitivity moves to lower frequencies, 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 thresholds observed from approximately 30 centimeters at original paper size.

3.3 Contrast Transduction

The visual sensitivity to a contrast patch of a certain spatial frequency decreases with the presence of other similar frequency contrast. Daly's Visible Differences Predictor (VDP)[20] accounts for this effect known as *visual masking* using a threshold elevation map. This approach trades off supra–threshold contrast interval for near–threshold precision. Such a trade–off is not suitable to our purposes, as real–world scenes are expected to comprise contrast well above the visibility threshold. Thus, in our model we employ the *transducer* function T described in[23] based on the premise that it is tuned for both near–and supra threshold precision. The contrast C at each scale is processed separately as follows:

$$R_{al|ml} = T(C, S_{al|ml}) = \frac{3.291 \cdot [(1 + (S_{al|ml}C)^3)^{1/3} - 1]}{0.2599 \cdot (3.433 + S_{al|ml}C)^{0.8}}, \tag{3}$$

where $R_{al|ml}$ is adapted or maladapted human perceptual response to contrast, $S_{al|ml}$ is the sensitivity at either maladapted or fully adapted state. The constants are taken from Wilson's work without any change. The monotonically increasing behaviour of the transducer function enables a fast inversion through the use of a lookup table stored in GPU memory.

3.4 Temporal Adaptation

Temporal adaptation can be modeled through two separate exponential decay functions; one for *pigment bleaching and regeneration* and another for *neural adaptation*.[5] For simplicity, we describe the adaptation process generically, but recall (Sec. 3.1) that the final adaptation map L_a^* is combined from six values that possess different time constants.

The time course of the neural adaptation mechanism from perceived luminance L_0^* at time $t = 0$, to $L_\cdot^*$ (where $L_\cdot^*$ is $\sigma_{cc}, \sigma_{nc}, \sigma_{cr}, \sigma_{nr}$) is modeled as follows:

$$L_\cdot^* = L_b^* + (L_0^* - L_b^*)\, e^{\frac{-t}{t_0}}. \tag{4}$$

The contribution of neural adaptation to temporal recovery of visual sensitivity is modeled by updating the *cvia* function at each time step using the current L_a^*. We set t_0 to 0.08 seconds for cones, and 0.15 seconds for the rods.[5]

Pigment bleaching and regeneration (modeled by σ_{bc} and σ_{br}), unlike neural adaptation, are slow and not symmetric for dark and bright adaptation. Assuming that the amount of signal transmitted by receptors is proportional to $p \cdot L$, the fraction of unbleached pigments p is computed as in Equation 5:

$$p = p(L_b) + (p_0 - p(L_b))\, e^{\frac{-t}{t_0 \cdot p(L_b)}}. \tag{5}$$

In the steady state, $p(L)$ is $I_0/(I_0 + L)$ where I_0 is 10^4 cd/m^2. The time constant t_0 is set to 110 and 400 seconds for cones and rods, respectively.

3.5 Luminance and Color Processing

The inverse transducer converts maladapted contrast responses R_m to the luminance values L_m. By summing all the levels of the Laplacian pyramid we obtain the maladapted luminance map. This map represents the hypothetical output of the display device that would evoke the same perception of *contrast* in a fully-adapted display observer as the original HDR scene in the maladapted observer. However, to account also for the luminance sensitivity, we transform L_m using the S–shaped function as follows:

$$L_d = \frac{L_m}{L_m + \bar{L_a^*}}; \qquad \bar{L_a^*} = (\Pi^n L_a^*)^{1/n}, \tag{6}$$

where L_d is the output display luminance, L_m is the luminance value we obtained from of inverse observer model, and $\bar{L_a^*}$ is a geometric average of the current adaptation luminance map. Note that the value of L_a^* accounts for the current (mal)adaptation state and therefore the S–shaped function results in dark images for dark adaptation scenario and bright images for adaptation to bright scenes, and the sensation will improve according to the temporal adaptation as described above.

As our aim is the simulation of maladaptation in contrast domain (and not the simulation of color vision phenomena), we perform only very simplified color processing. Simply put, all the above described processing happens on achromatic channel only. However, as the local adaptation map L_a is calculated as a combination of the human rod and cone responses (L_{ar}, L_{ac}, refer to Section 3.4) we can utilize them for color processing. In the absence of a model of color perception specialized in maladapted HVS states, we desaturate the colors as follows: $I_{\text{corr}} = I^s$ (for each color channel I separately) using the following saturation coefficient $s = L_{ac}/(L_{ac} + L_{ar})$, where L_{ac} and L_{ar} are current cone and rod adaptation map values, respectively.

3.6 Inverse Display Model

Our simple display model consists of three parameters, the maximum and minimum display luminance, and a gamma value which we set to 1/2.2. The gamma corrected values I_{corr} are fitted to the luminance range of the display by a simple linear mapping. Our interface allows the user to control the display luminance range, this way a variety of display types can be approximated. For a more precise simulation of specific displays the linear mapping can be replaced by the display response function.

4. RESULTS

In this section we discuss our results, implementation details, and the other possible uses of the model. Please refer to supplemental materials (`http://mpi-inf.mpg.de/~mcadik/maladaptation`) for further results.

A visual verification of maladapted contrast sensitivity behavior (described in Sec. 3.2) is presented in Fig. 6. We augmented the Campbell-Robson chart with a frame of uniform luminance and generated two different HDR images (an initial and a final) to simulate the dark adaptation. The results show both the shift in peak frequency and the drop and regeneration of absolute sensitivity as expected. As we want to illustrate only the contrast processing in this figure, we simplify the equation (6) to $L_d = L_m/(L_m + k)$, where k is set to 100 cd/m^2. Otherwise the maladapted images (the two middle images in Fig. 6) would be too dark to visualize. Compare the output of our model in Fig. 6 with Fig. 5 and 1 (right).

In Fig. 7 we compare our results to the approach of Irawan et al.[5] The reference method (upper row) is based on global tone mapping function and global background luminance (L_{bg}), while our approach (bottom row) operates on contrasts and utilizes local background luminance (L_b). In both cases global adaptation luminance L_{ag} is assumed. Therefore our method accounts for diverse perception of bright and dark areas of the scene. One can notice a difference in the fully adapted state as well (Fig. 7-rightmost images): in our model, the stained glass window is reproduced sharply and all the details are visible, while the dark area below the desk is blurred, which is the expected behavior. Note that by considering local background luminance (L_b) we ignore changes in the state of adaptation due to attending different image regions as a consequence of the saccadic eye motion. We rather visualize the image appearance under the condition that the eye attends locally each respective region without any gaze change. Thus, Fig. 7 (bottom row) presents a synthetic summary how each specific region will be seen under this assumption, but the overall image appearance may not be presented precisely. Irawan

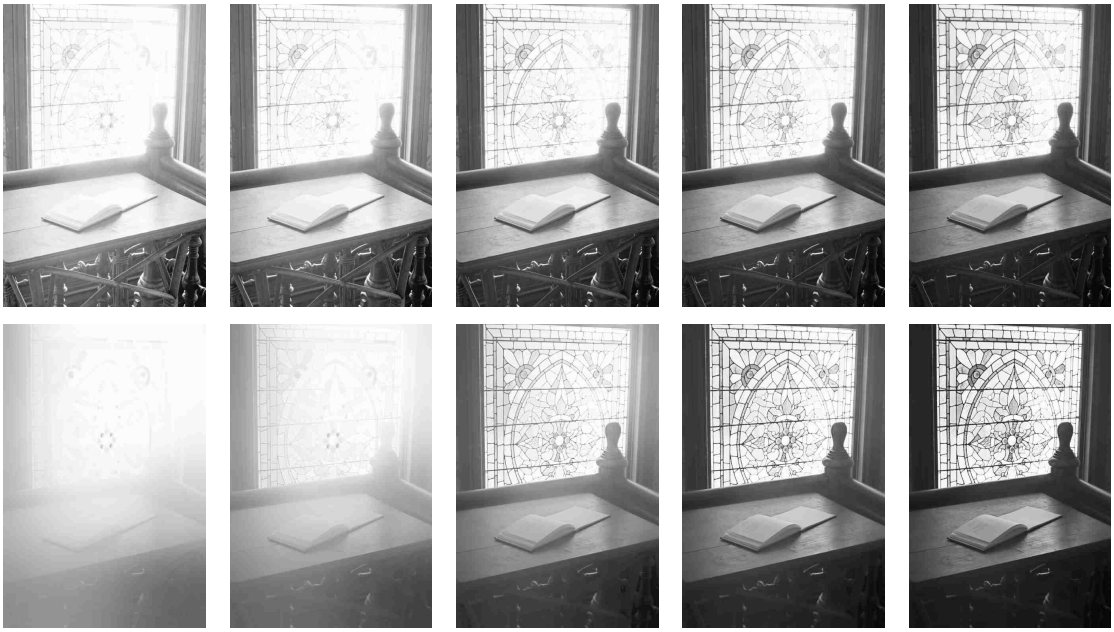


Figure 7. Comparison of our method to the approach of Irawan et al.[5] simulating the fast adaptation from a dark environment (10^{-4} cd/m^2) to the stained glass (17 cd/m^2). Top row: method of Irawan et al. can not simulate differences in perception of contrasts in bright and dark parts of the scene, however our method can (bottom row). 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). HDR image courtesy of OpenEXR.

used another extreme approach by considering global background L_{bg} (Fig. 7 – upper row), in which case it is implicitly assumed that through the gaze direction changes the eye adaptation tends to some average luminance in the scene. Since the most dramatic changes in light adaptation take place during the time required just for a couple of fixations this assumption is also not realistic in particular for video, while it is commonly used. Thus, Fig. 7 (upper row) gives perhaps a better prediction of overall image impression (except that no frequency processing is accomplished), but the local detail visibility might be better predicted in Fig. 7 (bottom row).

In Fig. 8 we illustrate an application of our approach to analysis of the visibility of a display and controls on the panel in a flight control room. Left column shows fully adapted state, where all the details are well visible. After the adaptation to the bright sky however, those details are not noticeable for some seconds.

Our system enables also to simulate even more complex scenario (see Fig. 9) where in lack of opposing evidence we consider local adaptation L_a and background L_b maps, in which case each local region in the image

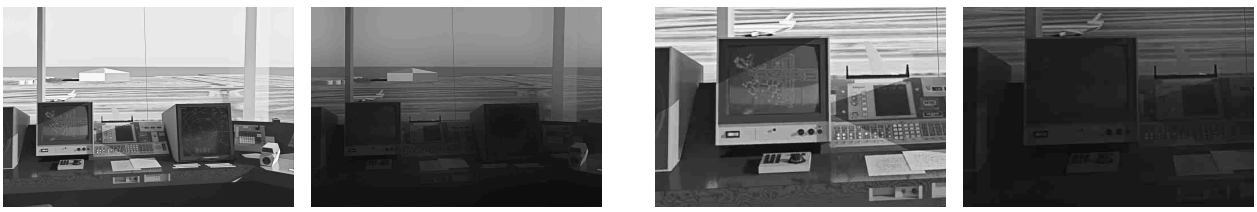


Figure 8. Rendering of the interior of an airport control tower. Left: fully adapted state (178 cd/m^2). Right: maladaptation due to a previous exposition to the bright sky (10^4 cd/m^2), $t = 0.5$s. Compare the visibility of displays and controls in close-ups (right pair). HDR image courtesy of Greg Ward.

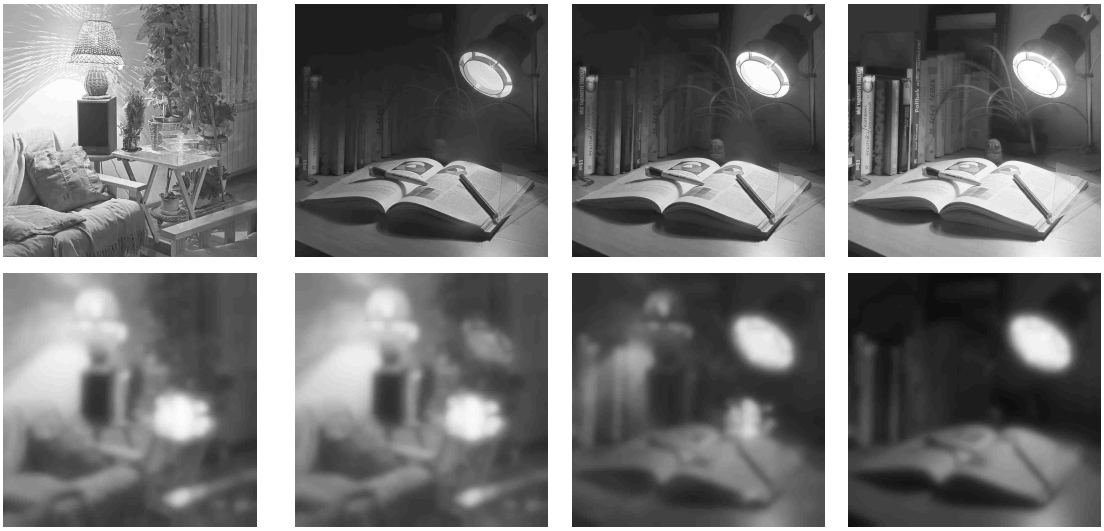


Figure 9. Simulation of maladaptation in a complex hypothetical scenario. Top row, from left to right: (1) fully adapted state in a relatively bright living room (adaptation luminance 70 cd/m^2), (2) rapid movement to a work-room (8 cd/m^2), time $t = 0.02$s, (3) sensitivity regenerates and aftereffects diminish ($t = 0.1$s), (4) in $t = 0.5$s the observer reached nearly fully–adapted state, but some dark details are washed out, due to the dim illumination of the work–room. Bottom row: states of adaptation maps L_a corresponding to the upper images. As the values in adaptation maps are HDR, they were tone mapped using the global version of Reinhard's[24] TMO for the display purpose.

has also corresponding local adaptation. Imagine an observer who is adapted to relatively bright illumination of a living room and then she instantly moves to her desk in a dim work–room. For a moment, while her sight is being regenerated, she does not see the details in some parts of the scene due to the previous adaptation to much brighter environment. The vision reaches the fully adapted state in some seconds, but due to the low illumination of the work-room, the vision is still not sharp in dark parts of the scene.

4.1 Fast GPU Implementation

In order to get real–time performance, we moved perceptual (contrast transduction, calculation of maladaptation, local cones and rods adaptation) and image processing (laplacian pyramid, tone-mapping) parts of the algorithm to the GPU. For the purpose of real–time HDR movie processing we also had the radiance format decompression realized on the GPU. Because of that, we are able to achieve interactive frame rates on mainstream hardware. Our test system is based on Intel Core2 3.0Ghz CPU, 4GB of RAM and NVidia GTX260 GPU. The average performance is around 75 fps for 1024x1024 HDR image. After processing the data we manually copy the resulting texture to GUI surface. Displaying the image directly would improve the speed even further.

4.2 Simulating Maladaptation in LDR Images

Another possible application of our model is the simulation of maladaptation effects on contrast perception in an ordinary (LDR) image, see Fig. 10. Let us assume that only an LDR image is available, but we want to know how its appearance will be affected due to the maladaptation. It is possible to perform inverse tone mapping[25] and derive a reasonable approximation of adaptation map. Having a scene referred HDR image as the adaptation pattern, we can simulate appropriate HVS reaction for arbitrary LDR image as follows: we run the model for the HDR image and we keep the contrast responses R_a for fully adapted observer and for a particular maladapted state R_m, then we linearize the LDR image using inverse gamma correction and decompose it using Laplacian pyramid. To simulate maladaptation in the LDR image, we multiply the values in the Laplacian pyramid as

follows:

$$C'_{i,j,l} = C_{i,j,l} \cdot \frac{R^m_{i,j,l}}{R^a_{i,j,l}}, \tag{7}$$

where C is the current LDR contrast value for pixel i, j and level l of Laplacian pyramid. The final LDR image with the simulated maladaptation effect is obtained by adding C' at all the levels of modified Laplacian pyramid. In this special case we simulate only the effect of maladaptation to the perception of contrasts, as we omit the luminance processing (i.e. we do not involve equation (6)).

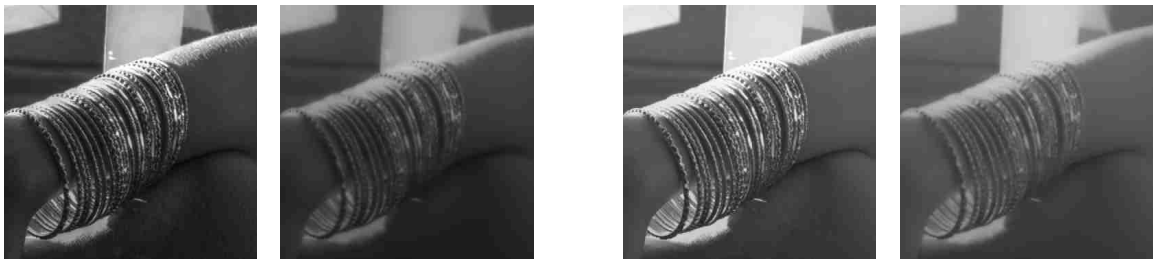


Figure 10. Simulation of maladaptation in two different LDR images. In each pair: left: original LDR image. Right: maladaptation simulation using the background luminance from the HDR image (200cd/m^2) obtained by the inverse tone mapping. Simulated adaptation luminance: 20cd/m^2. HDR image courtesy of Allan Rempel et al.[25]

5. CONCLUSION

We presented an efficient, real–time visual maladaptation framework capable of rendering images of a scene as perceived by a maladapted observer. Our model operates on contrast domain and accounts for supra–threshold HVS mechanisms such as visual masking, as well as luminance adaptation and contrast sensitivity as a function of spatial frequencies that have often been neglected by previous contrast domain methods. We also model the shift in spatial frequency sensitivity due to maladaptation, which we found to have a significant effect on scene visibility. We discuss a fast GPU implementation that enables interactive rendering of maladapted images. Our system can potentially be used to simulate human vision in illumination conditions causing extreme maladaptation in real–world scenarios such as driving.

5.1 Limitations and Future Work

As the model is not targeted for the simulation of HVS *color processing*, it mainly operates on the achromatic channel only. Therefore it does not account for chromatic adaptation, color aftereffects and other phenomena of color vision; but we believe those can be pertinently included, if necessary.

The model assumes to input a calibrated HDR image and by modeling of the HVS features it is accordingly able to perform the *HDR tone mapping* task (for a calibrated HDR image). However, as the primary goal of the model is the correct simulation of the HVS contrast processing, the results for some extremely high dynamic range or not calibrated images can not outperform the results of specifically tuned tone mapping operators. Note however, that the HVS is also unable to see all the details in the scene simultaneously for extremely high dynamic ranges. From this point of view, the results of many "successful" tone mapping operators are not perceptually correct, as indicated by recent experimental studies.[26,27]

REFERENCES

[1] Krantz, J. H., Silverstein, L. D., and Yeh, Y.-Y., "Visibility of transmissive liquid crystal displays under dynamic lighting conditions," *Hum. Factors* **34**(5), 615–632 (1992).

[2] Silverstein, L., "Display visibility in dynamic lighting environments: Impact on the design of portable and vehicular displays," (2003). IDMC'03, Taipei, Taiwan.

[3] Pattanaik, S. N., Ferwerda, J. A., Fairchild, M. D., and Greenberg, D. P., "A multiscale model of adaptation and spatial vision for realistic image display," in [*SIGGRAPH '98*], 287–298, ACM Press (1998).

[4] Pattanaik, S. N., Tumblin, J., Yee, H., and Greenberg, D. P., "Time-dependent visual adaptation for fast realistic image display," in [*SIGGRAPH '00*], 47–54 (2000).

[5] Irawan, P., Ferwerda, J. A., and Marschner, S. R., "Perceptually based tone mapping of high dynamic range image streams," in [*EGSR'05*], 231–242 (2005).

[6] Ferwerda, J. A., Pattanaik, S. N., Shirley, P. S., and Greenberg, D. P., "A model of visual masking for computer graphics," in [*SIGGRAPH'97*], 143–152 (Aug. 1997).

[7] Ferwerda, J. A., Pattanaik, S. N., Shirley, P., and Greenberg, D. P., "A model of visual adaptation for realistic image synthesis," in [*SIGGRAPH '96*], 249–258, ACM, New York, NY, USA (1996).

[8] Ward, G., "A contrast-based scalefactor for luminance display," *Graphics Gems IV* , 415– 421 (1994).

[9] Hunt, R. W. G., [*The reproduction of colour*], Fountain Press, 5. ed. ed. (1995).

[10] Larson, G. W., Rushmeier, H., and Piatko, C., "A visibility matching tone reproduction operator for high dynamic range scenes," *IEEE TVCG* **3**, 291–306 (1997).

[11] Mantiuk, R., Myszkowski, K., and Seidel, H.-P., "A perceptual framework for contrast processing of high dynamic range images," in [*APGV '05*], 87–94, ACM Press (2005).

[12] Aydın, T. O., Mantiuk, R., and Seidel, H.-P., "Predicting display visibility under dynamically changing lighting conditions," *EUROGRAPHICS* **28**(3) (2009).

[13] Barten, P. G., [*Contrast sensitivity of the human eye and its effects on image quality*], SPIE (1999).

[14] Kelly, D. H., "Spatio-temporal frequency characteristics of color-vision mechanisms," *Journal of the Optical Society of America (1917-1983)* **64**, 983–+ (July 1974).

[15] Dowling, J., [*The Retina: An Approachable Part of the Brain*], Harvard Univ. Press (1987).

[16] Alexander, K. R., Barnes, C. S., Fishman, G. A., Pokorny, J., and Smith, V. C., "Contrast Sensitivity Deficits in Inferred Magnocellular and Parvocellular Pathways in Retinitis Pigmentosa," *Invest. Ophthalmol. Vis. Sci.* **45**(12), 4510–4519 (2004).

[17] Hahn, L. W. and Geisler, W. S., "Adaptation mechanisms in spatial vision–i. bleaches and backgrounds," *Vision Research* **35**(11), 1585 – 1594 (1995).

[18] Lennie, P., "Roles of M and P pathways," in [*Contrast Sensitivity*], Shapley, R. and Lam, D. M.-K., eds., ch. 11, 201– 213, MIT Press (1993).

[19] Leonova, A., Pokorny, J., and Smith, V. C., "Spatial frequency processing in inferred pc- and mc-pathways," *Vision Research* **43**(20), 2133 – 2139 (2003).

[20] Daly, S., "The visible differences predictor: An algorithm for the assessment of image fidelity," in [*Digital Images and Human Vision*], Watson, A. B., ed., 179– 206, MIT Press (1993).

[21] Mantiuk, R., Daly, S., Myszkowski, K., and Seidel, H.-P., "Predicting visible differences in high dynamic range images – model and its calibration," in [*HVEI X, IS&T/SPIE'05*], **5666**, 204–214 (2005).

[22] Burt, P. J. and Adelson, E. H., "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications* **Com–31**, 532– 540 (1983).

[23] Wilson, H., "A transducer function for threshold and suprathreshold human vision," *Biol. Cybernetics* **38**, 171–178 (1980).

[24] Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J., "Photographic tone reproduction for digital images," in [*SIGGRAPH '02*], 267–276, ACM Press (2002).

[25] Rempel, A. G., Trentacoste, M., Seetzen, H., Young, H. D., Heidrich, W., Whitehead, L., and Ward, G., "Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs," in [*SIGGRAPH '07: ACM SIGGRAPH 2007 papers*], 39, ACM, New York, NY, USA (2007).

[26] Akyüz, A. O., Fleming, R., Riecke, B. E., Reinhard, E., and Bülthoff, H. H., "Do HDR Displays Support LDR Content? A Psychophysical Evaluation," *SIGGRAPH'07* **26**(3), 38 (2007).

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

Appendix D

An Efficient Perception-based Adaptive Color to Gray Transformation

L. Neumann, M. Čadík, and A. Nemcsics. An Efficient Perception-Based Adaptive Color to Gray Transformation. In *Proceedings of Computational Aesthetics 2007*, pp. 73– 80, Eurographics Association, Banff, Canada, 2007.

Computational Aesthetics in Graphics, Visualization, and Imaging (2007)
D. W. Cunningham, G. Meyer, L. Neumann (Editors)

An Efficient Perception-based Adaptive Color to Gray Transformation

L. Neumann[†1] and M. Čadík[‡2] and A. Nemcsics[§3]

[1]ICREA, Barcelona, and VICOROB, University of Girona, Spain
[2]Department of Computer Science and Engineering, Czech Technical University in Prague, Czech Republic
[3]Technical University of Budapest, Hungary

Abstract

The visualization of color images in gray scale has high practical and theoretical importance. Neither the existing local, gradient based methods, nor the fast global techniques give a satisfying result. We present a new color to grayscale transformation, based on the experimental background of the Coloroid system observations. We regard the color and luminance contrasts as a gradient field and we introduce a new simple, yet very efficient method to solve the inconsistency of the field. Having a consistent gradient field, we obtain the resultant image via fast direct integration. The complexity of the method is linear in the number of pixels, making it fast and suitable for high resolution images.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Picture Image Generation]: Display Algorithms, Viewing Algorithms; I.4.3 [Image Processing and Computer Vision]: Enhancement-Filtering

1. Introduction

Nowadays, except for a few artistic and scientific applications, the vast majority of captured images are color photographs. On the other hand, many laser printers are still black-and-white, and most of the pictures in daily newspapers published in the world are predominantly gray-scale images. Thereby the practical importance of color to grayscale transformations is clear. The theoretical challenge is also evident. The color to gray transformation is a mapping of a 3D set with spatial coherences to a one dimensional (1D) space and it necessarily leads to some information loss. What is the best way? Which way gives the highest perceptual equivalence? Does there exist a universal approach?

The solution requires the preservation of chromatic contrasts during the conversion to luminance contrasts and the associated evaluation of the luminance and chrominance changes (gradients) and values. The problem combines various aspects of color vision and spatial vision. How does the visual effect of chrominance and luminance contrasts depend on spatial frequencies?

The above questions do not have simple solutions, as adaptive color to gray transformations are not generally found in nature. We believe that global transformation approaches cannot give a full answer to the above questions, although they appear to offer some fast and acceptable results. The adaptive local methods hold the promise of a much better solution, although they suffer from theoretical problems in perceptual modeling and practical difficulties in numerical calculations. In this paper, we present a perceptually-based adaptive approach using the experimental background of the Coloroid observations. We investigated the relative equivalent luminance differences for a set of chromatic differences at a given spatial frequency, using 10×10 cm solid color samples. However, a comprehensive spatio-chromatic analysis still demands further investigation.

The paper is structured as follows. We review previous work on color to gray image transformation in Section 2. In Section 3, we describe the Coloroid color system and we present our new observations based on the Coloroid. Sec-

† lneumann@silver.udg.es
‡ cadikm@fel.cvut.cz
§ nemcsics.antal@t-online.hu
Full color versions of the images and other materials are online:
http://www.cgg.cvut.cz/~cadikm/color_to_gray

tion 4 introduces an efficient gradient-based color to gray transformation algorithm powered by a new gradient inconsistency correction method. Then, in Section 5, we show and discuss the results of the presented transformation algorithm. Finally, in Section 6, we conclude and suggest some ideas for future research.

2. Related work

There are several approaches available in the literature that aim to convert color images to grayscale. Strickland et al. [SKM87] proposed a local color image enhancement technique used to sharpen images based on saturation feedback. Zhang and Wandell [ZW96] devised a spatial extension of the CIELab color model (S-CIELab) that is useful for measuring color differences between images. Using the pattern-color separable transformation, the S-CIELab difference measure reflects both spatial and color sensitivity. Bala and Eschbach [BE04] presented spatial color to gray transformation that locally preserves the chrominance edges by introducing high-frequency chrominance information into the luminance channel. The method applies a spatial high-pass filter to the chromatic channels, weighs the output with a luminance dependent term, and finally adds the result to the luminance channel. Grundland and Dodgson [GD05] proposed the decolorize algorithm for contrast enhancing, color to grayscale conversion. The method applies a global color to grayscale conversion by expressing grayscale as a continuous, image dependent, piecewise linear mapping of the RGB color primaries and their color saturation. The authors calibrate the behavior of their method by using three parameters to control contrast enhancement, scale selection, and noise suppression. The authors suggest image independent default values for these parameters. Gooch et al. [GOTG05] presented a Color2Gray algorithm which iteratively adjusts the gray value of each pixel to minimize an objective function based on local contrasts between pixels. The method applies three free parameters (θ, α, μ), but the authors do not provide image independent default values. Moreover, the complexity of the method is $O(N^4)$. Hence, the method is very slow and it is difficult to apply to high resolution images. Rasche et al. [RGW05] presented a color to gray technique that aims to preserve the contrast while maintaining luminance consistency. Authors approach the problem by means of constrained multidimensional scaling which scales badly with the number of colors, and therefore the color quantization is suggested. However, due to the necessary quantization of colors the method produces quantization-like artifacts. Therefore, the usage of this method is very questionable for images with continuous tones (e.g. real-world photos). Moreover, the time-demands are enormous (even a low-res image transformation takes minutes) and depend on the number of colors.

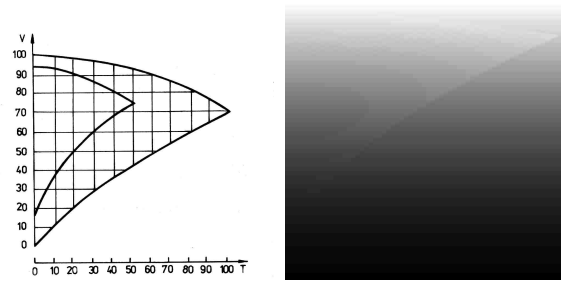


Figure 1: *Left: shape of the Coloroid gamut at a fixed hue value. Right: turquoise hue plane of the Coloroid space.*

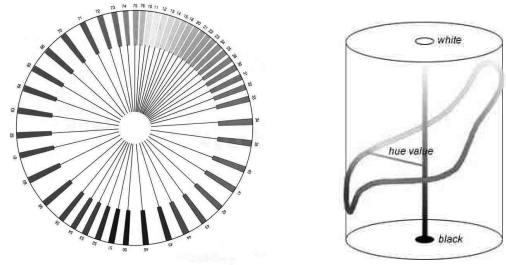


Figure 2: *Left: 48 limit colors of the Coloroid system. Right: curve of Coloroid limit colors in 3D.*

3. The Coloroid system

The Coloroid is a color-order system and color space with conversion formulas to and from CIE XYZ system. The Coloroid system is based on huge number of observations [Nem01] and represents perhaps the most adequate tool or "natural language" to describe harmony relationships and other psychometric attributes between colors [NNN05]. The experimental arrangement of the observations of the color harmony relationships is an ideal tool to study the basic questions of the color to gray transformation, especially to find the chrominance-luminance equivalent attributes on a relative scale.

Conditions of observations and basic concept of the Coloroid system differ from other color order systems. In typical Coloroid experiments, the observer is given a wide field of view to observe a large set of often non-neighboring color samples, and must give their responses relative quickly. These conditions make it similar to an observation of a complex image in the real life. Under these viewing conditions the human vision system can distinguish a reduced number of colors, especially in the darker regions. The colors in Coloroid can be obtained by additive mixture of the black, white and the limit-color, by ratios s, w, and p, respectively, where $s + w + p = 1$. The limit-colors were the most saturated *solid-colors* instead of spectral colors. Due to the very great number of observations and also to the obtained good correlations, we consider the basic concepts of Coloroid to

be *axioms*, which are valid for the above mentioned view-conditions:

1. Surfaces of a constant Hue (A) form a plane, containing the neutral axis and a hue dependent limit-color, unlike most of the other systems that have curved hue-surfaces (e.g. the Munsell system).
2. Saturation $T = const * ratio(p)$ of the limit-color, where the constant depends on the hue.
3. Lightness $V = 10 * Y^{1/2}$. Unlike the ds line-element based spaces, the Coloroid does not contain a 3rd root or a logarithmic formula here.

Fig. 2 (left) shows the circle of 48 limit-colors, while Fig. 2 (right) shows the continuous 3D limit-color line. Fig. 1 (left) demonstrates the typical shape of the Coloroid gamut at a fixed hue value. The lightness of the most saturated point depends on the hue according to Fig. 2 (right). The two Coloroid gamuts represent two limit-color selections. The larger one corresponds to the spectrum and purple limit-colors and smaller to the most saturated solid-colors, which will be used in our paper. Concepts and formulas of Coloroid can be found in several basic publications [Nem80, Nem87, Hun92], a deep survey of application areas can be found in [Nem04].

3.1. Observations based on Coloroid

After the above short survey we present the experiments in color to gray conversion. First, we studied the relative *luminance difference of the hue pairs* in an average sense. We first selected seven basic colors, one from each Coloroid hue group. For these seven basic colors, Table 1 shows the values of the Coloroid hues A, their characteristic wavelengths λ and their angular degrees φ for the D65 white point in the CIE xy chromacity diagram. The observers' task was to rank the color and gray samples, or to fit missing samples into color series with short adaptation time (semi-adapted eye) [Nem04]. The obtained 7×7 matrix (see Table 2) contains the relative gray-equivalent differences. The matrix has zero diagonal and it is anti-symmetric. The maximal value is scaled to 10. The largest perceived difference is from the A=50 blue hue to the A=10 yellow hue. For arbitrary (A1, A2) saturations, we applied 4 linear interpolations, which preserved the anti-symmetric property of the perceived differences. In the Section 4.1.2, we describe how to generalize the above hue-pair based gray change for arbitrary saturations.

The second observation series aimed to formulate the *gray-equivalence of the saturation increase*. We investigated the effect of the saturation increase for all of the above hues (A=10, 20, 30, 40, 50, 60, 70) at different constant luminance levels (V=45, 65, 85). As above, all of the observations were scaled to maximal value of 10. The unexpected fact is that the equivalent gray difference changes *non-monotonously*! For example, on the A = 60 hue page at the Coloroid-lightness V = 85 in the realistic range of saturations (T) for solid and monitor colors we obtained:

A	λ	φ
10	570.836	58.040
20	582.640	32.898
30	602.717	5.533
40	-504.836	-46.209
50	450.000	-116.628
60	490.371	-174.503
70	536.295	103.890

Table 1: *Definition of the seven basic Coloroid hues*

A/A	10	20	30	40	50	60	70
10	**0.0**	-2.5	-5.0	-7.0	-10.0	-5.0	-2.0
20	2.5	**0.0**	-2.5	-5.0	-8.0	-3.0	1.5
30	5.0	2.5	**0.0**	-3.0	-5.0	-3.0	3.5
40	7.0	5.0	3.0	**0.0**	-2.5	1.0	4.0
50	10.0	8.0	5.0	2.5	**0.0**	4.0	8.5
60	5.0	3.0	3.0	-1.0	-4.0	**0.0**	3.0
70	2.0	-1.5	-3.5	-4.0	-8.5	-3.0	**0.0**

Table 2: *Relative gray-equivalent differences of the basic hue pairs*

Relative saturation difference	1	2	3	4	5
Relative Δ-gray	1	2	4	0	-5

In the above example, low saturation differences lead to positive gray differences while high saturation differences appear to lead to negative gray differences – a highly saturated bright turquoise can be visualized by a gray decrease, while the middle saturated one of nearly the same value requires a gray increase. This relationship holds for the entire gamut. We performed the interpolation between the seven selected hues using the maximal absolute solid-color saturation values of the 48 Coloroid pages. Thereafter, we apply *relative saturations* at every hue and luminance level. The relative saturation is defined to take on the maximal value of 5 at the Coloroid gamut border [Nem04]. The relative saturation is obtained using the Coloroid limit-colors. For an arbitrary color, trilinear interpolation was applied, taking proper account of the zero saturation of the black and white points.

The two gray changes mentioned above are scaled on relative scales, independently. We made dozens of additional observations, where two attributes were changed simultaneously, to calibrate the relative scales to each other using linear regression. For example, here are two color pairs from this set of observations (where Δ-gray is the observed equivalent absolute gray difference):

Δ-gray=-1.0 (A1= 70.0, T1= 15.0, V1= 67.0)
(A2= 24.0, T2= 15.0, V2= 67.0)

Δ-gray=+4.0 (A1= 30.0, T1= 32.0, V1= 47.0)
(A2= 50.0, T2= 32.0, V2= 47.0)

4. Adaptive color to gray transformation

Our adaptive color to gray transformation method consists of three steps. In the first step, we regard the color and luminance contrasts as a gradient field which we construct using formulas described in section 4.1. Then, instead of using a Poisson solver or similar computationally-demanding approach, we correct the gradient field using a newly introduced fast and effective gradient inconsistency correction method based on an orthogonal projection (section 4.2). Finally, we integrate the corrected gradient field and transform the values to the display range to get the resulting image.

4.1. Formulas for building the gradient field

We propose two formulas for construction of the gradient field. The first formula is simple to implement and operates directly on the CIELab color data, while the advanced second one takes the full advantage of the Coloroid color space.

4.1.1. A simple new CIELab based formula

In advance to the Coloroid formula, we studied an extension of Color2Gray method [GOTG05] to avoid the artifacts and to reduce the computational costs. Based on the CIELab values, the mentioned method computes the warm-cold hue transient value multiplied by the chroma and finally modified by a stretched tanh (the "crunch") function to obtain the chrominance. The used signed gray difference is either the chrominance or the luminance value selected according to the max function of their absolute values. However, this approach can result in a strongly non-consistent gradient field, e.g. a large negative value can appear immediately after a large positive one. To "blur" this kind of artifacts the method requires a large neighborhood, and practically the complexity of $O(N^4)$.

To overcome the mentioned shortcomings, we introduce a non-max based, continuous function using the CIELab space. Being the max the $n = +\infty$ power-norm, we use the 3rd power norm, which both preserves somewhat from the max feature, but it is also near to the square-root. Let $A = w_a * a$ and $B = w_b * a$, where w_a and w_b in interval [0.2..0.6] are weight factors to reduce the chrominance-luminance ratio. The equivalent luminance has to be smaller than a CIE color difference value, which can go over 200. Our new formula is as follows:

$$\Delta = \left(\Delta L^3 + \Delta A^3 + \Delta B^3\right)^{1/3}. \quad (1)$$

Formula (1) conveys directly the sign of the Δ gray difference. In the worst "diagonal" colors, the difference from the color difference value - using square-root of a and b - is negligible for the wanted purpose. See one of results of this approach on Figure 3; please note that in contrast to Gooch et al., it takes just a fraction of a second to process this image by our method and the result exhibits more details. The gradient field was corrected with the method described in the paragraph 4.2, using the 1-pixel neighborhood. The above method can be simply extended with 4 weight factors, different for positive and negative a and b (red-green and yellow-blue) channels.

Figure 3: *Comparison of our method using the CIELab formula with the CIE Y equivalent and Gooch et al. Top left: original color image, top right: CIE Y equivalent, bottom left: the result of Gooch et al., bottom right: our adaptive color to gray transformation result.*

4.1.2. The Coloroid based formula

The XYZ color system coordinates to Coloroid coordinates transformation and the Coloroid (ATV) based local gray-change (gradient) formula have central importance in the proposed method. Unfortunately, they cannot be given in a closed form, since they contain tables of observations with the appropriately accurate interpolation rules. Therefore, we describe the structure of the formula and explain the meanings of the terms here.

As the relative gray-equivalency of the *hue changes* is given only for 7 basic hues (by the Tables 1, 2), we apply a bilinear interpolation. In particular, we linearly interpolate the φ values to derive the color hues [Hun92]. For an arbitrary hue-pair (A_1, A_2), we obtain this way a H value in the interval of [-10,10]. The hue-term is additive in this model and it depends sub-linearly on the saturation. In the gradient term, the H occurs with a weight factor of the following form:

$$h(A_1, T_1, A_2, T_2) = w_h \times H(A_1, A_2) \times \sqrt{u(T_{1rel}) \times u(T_{2rel})}, \quad (2)$$

where T_{rel} is the relative saturation scaled to [0,5] for every hue plane and at every luminance level, computed from the maximal solid-color saturation. Equation (2) contains the geometrical mean of the two u-factors, and therefore will be

zero if at least one of the two colors is neutral. The $u(x)$, where $x = 2 \times T_{rel}$, is defined as follows:
$u(x) = 0.5 \times x$, iff $x < 0.5$
$u(x) = \sqrt{x} - 0.5$, otherwise.

The *saturation dependent* gray-equivalent change is more complicated, since it depends on hue and on luminance too. We have to evaluate the relative gray-change of both colors. We made observations for the 7 basic hues (Table 1) using the perceptually uniform Coloroid V values, at the luminance levels of 45, 65 and 85. In the black and white point the change is zero ($V = 0$ and $V = 100$). The suggested gray change of a color (A,T,V) due to the saturation term is scaled also to [-10,10], but with a different weighting. We made additional observations to fit the two different scalings to each other. The effect of growing saturation can result in a positive or negative gray change for a fixed hue and luminance. We use the φ values of the 7 basic hues and the data of the most saturated solid colors [Nem04], that is a version of the 48 limit-colors, and furthermore the above mentioned 5 luminance levels. Let us notate $S(A,T,V)$ the gray-change-effect of the saturation of one color. To compute S, we have to obtain the T_{rel} value first, as in the case of the hue. Then we apply a trilinear interpolation using the neighboring φ, T_{rel}, and V values. For two colors, the signed gray-change can be obtained in the form:

$$S(A_1,T_1,V_1,A_2,T_2,V_2) = w_s \times [S(A_2,T_2,V_2) - S(A_1,T_1,V_1)]. \quad (3)$$

The evident part of the gray gradient is the *luminance difference*, without weighting:

$$dL(L_1,L_2) = L_2 - L_1. \quad (4)$$

The color difference (gradient) is then obtained by adding the luminance (5), the saturation (3) and the hue (2) formulas:

$$\Delta_{1,2} = dL(L_1,L_2) + S(A_1,T_1,V_1,A_2,T_2,V_2) + h(A_1,T_1,A_2,T_2). \quad (5)$$

4.2. Gradient inconsistency correction method

Gradient domain imaging methods generally change the original gradient field of an image, or generate an artificial gradient field from a set of images. The key issue of that approach is the backward transformation – e.g. to find an image having the prescribed gradient field. An exact solution of the problem does not exist in general, there are only best approximations. The set of manipulated artificial gradient vectors is not a conservative consistent gradient field, thereby the appropriate unknown image does not exists, and we cannot obtain it via a 2D integration method.

Which image has the nearest gradient field to the given inconsistent one? This question is behind the existing methods. The well known and widely used multigrid Poisson solver, FFT method, or different iterative methods minimize the sum of elementary quadratic error terms containing the finite difference of unknown pixels and as constant the appropriate given horizontal or vertical gradient values. Perhaps the most efficient and elegant technique is the conjugate gradient method with locally adapted hierarchical basis preconditioning [Sze06].

In this section, we approach the problem of inconsistent gradient field with a new question. *What is the nearest consistent gradient field to an existing non-consistent one?* Having the nearest consistent gradients, the image can be obtained by a simple *two dimensional integration* requiring only one addition per pixel. For the sake of simplicity, we will present here the non multi-resolution basic version with an efficient over-projection.

The unknowns of the classical methods are the pixel luminance values. Our new approach uses two times more unknowns, namely all of X and Y components of each gradient vector (grad). The consistency has a simple pictorial meaning: going around a pixel, the total gradient changes have to be zero, see Figure 4 (left). Thereby, every pixel with 4 of the gradient components defines an equation. The total number of these equations is equal to the number of pixels ($N \times M$). The number of unknown gradient terms is $L = (N-1) \times M + (M-1) \times N$, which is approximately $2 \times N \times M$.

The possible inconsistent gradient terms can be described in the $L \approx N \times M$ dimensional space, while the nearest consistent field is searched in the $N \times M$ dimensional linear subspace of the consistent gradient fields. The metric is simply the Euclidean one, which defines the most natural way of "nearest point". The problem in higher dimension is similar to searching of the nearest point of a line or plane from an outer point in the 3D space. Summing the appropriate set of elementary equations with 4 gradient terms, we can obtain the equation of arbitrary closed curves. On all of these curves, or on all closed "Manhattan-lines" containing vertical and horizontal elementary intervals, the sum of the gradients has to be zero.

We remark an important feature of the new technique: nearly all of gradient methods face the problem of the *contradiction of gradients* on different resolution levels. E.g. in HDRI, Fattal et al. [FLW02] constructed an artificial gradient field in a multi-scale way.However, this approach changes also the larger low dynamic range image parts, which would have to remain invariant. Gooch et al [GOTG05] applied "every pixel to every others" comparison in $O(N^4)$ time to avoid the resolution-contradiction problem for a highly inconsistent gradient field and to obtain a pleasant global appearance. Our new consistency-correction method with the simple 1-neighbor gradients gives the wanted appearance, and solves implicitly the resolution-contradiction problem in a new way.

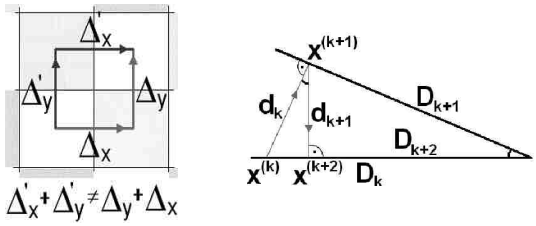


Figure 4: *Left: the sum of gradients around a pixel should be zero, but it is not in the case of inconsistent gradient field. Right: scheme of the orthogonal projection in 2D case.*

4.2.1. The algorithm

To find the nearest point in higher dimensional space, or to project a point orthogonally in a sub-space is generally a time-demanding algorithm. In a lower dimension, the Gram-Schmidt orthogonalization gives a closed form solution. Fortunately, for sparse matrix problems an iterative method tends to the nearest point with very simple elementary projection steps. Figure 4 (left) shows the pixel (middle point) and two different ways around the pixel. The total changes on these two ways have to be the same. After image manipulation or for artificially prescribed gradient field this consistency does not hold. The sum of gradients (E) on the closed curve containing the blue and red parts has to be zero. If it is non-zero, but $|E| > \varepsilon$, we have to change the values for all of closed "1-pixel-ways", until all of E-values converge to zero. In the app. $2 \times N \times M$ dimensional space of gradient components all of the 1-pixel-ways define an equation ($N \times M$), containing only 4 non-zero coefficients:
$g_x(i,j) + g_y(i+1,j) - g_y(i,j) - g_x(i,j+1) = E \neq 0$,
$\mathbf{N} = (0,..,0,+1,+1,-1,-1,0,..,0)$,

$$\mathbf{g}_{new} = \mathbf{g} - 1/4 \times E \times \mathbf{N},$$

(see Algorithm 1), where $\mathbf{g}$ is the vector describing the whole gradient field. The orthogonal projection method converges to the nearest point of a sub-space, in our case one of the consistent gradient fields. This subspace is the common part of all hyper-planes defined by the 1-pixel-way equations. If we select the equation with maximal error, and project the current gradients in the direction of $\mathbf{N}$ normal until reaching this plane, or fulfilling the equation, we go nearer to the wanted point according to the Figure 4 (right). The new distance D_{k+1} can be expressed with the old one and with the d_k distance of the projection characterizing the local inconsistency:

$$D_{k+1}^2 = D_k^2 - d_k^2.$$

With maximum error (E) selection, the method is more efficient, than with a cyclical correction of all of pixels, but this latter does not require a structure, e.g. to build a Fibonacci-heap to the quick max selection. On the other hand, the over-projection (e.g. parameter ω in Alg. 1), which is less efficient locally, significantly increases the overall convergence also in the non-multiresolution form of Algorithm 1 (the value of $\omega = 1.8$ is convenient for most images, while for $\omega = 1$ we get the original convergence rate).

Algorithm 1 Inconsistency correction

```
correct (gradient_field grad, double ω, double ε) {
repeat
  max_err=0;
  for y = 1 to YRES-1 do
    for x = 1 to XRES-1 do
      err=grad.X[x][y]+grad.Y[x+1][y]-
          -grad.Y[x][y]-grad.X[x][y+1]
      if |err|>max_err then
        max_err=|err|;
      end if
      s = 1/4 × err × ω;
      grad.X[x][y]=-s + grad.X[x][y];
      grad.Y[x+1][y]=-s + grad.Y[x+1][y];
      grad.Y[x][y]=s + grad.Y[x][y];
      grad.X[x][y+1]=s + grad.X[x][y+1];
    end for
  end for
until max_err < ε
}
```

Having the consistent gradient field, the final image is constructed via simple 2D integration, as shown in Algorithm 2. We believe the reader can implement this new simple, but efficient method very easily.

Algorithm 2 Double integration

```
integrate (gradient_field grad, output_image out) {
out[1][1] = 0;
for y = 1 to YRES do
  if y>1 then
    out[1][y] = out[1][y-1] + grad.Y[1][y-1];
  end if
  for x = 2 to XRES do
    out[x][y] = out[x-1][y] + grad.X[x-1][y];
  end for
end for
}
```

5. Results and discussion

We demonstrate the performance of our new color to gray transformation on a variety of color images and photographs. Figure 5 illustrates the mandatory color to gray transformation test containing largely isoluminant colors. We can observe from this figure how the classical approach results in a constant luminance (see Figure 5 - center). On the contrary, our approach (see Figure 5 - right) transforms the chrominance difference into well noticeable luminance differences.

The low contrast is due to the small visible differences in the original color image, preserving the overall appearance.

Figure 6 exhibits how, beyond other changes, the bluish image parts of the color image obtain a more realistic darker appearance in the resulting graylevel image after applying the proposed adaptive method. On the other hand, Figure 7 (top row) presents an obvious improvement of the final appearance in the details and visibility of the sky area. The sun is well visible in our result, while it nearly disappears in the classical gray conversion. Further examples are illustrated in Figure 7 and in color plates.

The processing time of a color to gray transformation using our approach is in the order of seconds even for high-res images (appr. 5 - 10 seconds per Megapixel), an appropriate value of parameter ε (see Alg. 1) is 0.001 for most images. We are currently working on the accelerated real-time version applying the multi-scale solution.

Figure 5: *An artificial isoluminant image. Left: original color image, middle: CIE Y equivalent, right: our adaptive color to gray transformation result.*

6. Conclusions and future work

We presented a new fast and efficient perceptual color to gray transformation method, based on a large number of experiments and observations of the local luminance-chrominance equivalency. Our method describes the luminance-equivalent nature of the whole gamut in a gradient domain, which (as we observed) often has an unexpected behavior with smooth changes. We propose two different formulas for the construction of the gradient field, first one operating in the CIELab color space, while the advanced second one takes the full advantage of the Coloroid color space.

Figure 6: *Left: original color image, middle: CIE Y equivalent, right: our adaptive color to gray transformation result.*

Moreover, we introduced a new gradient inconsistency correction method for solving the gradient field translated problem. The method has a linear complexity in the number of pixels and thereby it is suitable also for high resolution images. The method finds the most natural solution for a given inconsistent gradient field, e.g. the nearest one in the linear subspace of consistent gradient fields. The final image is then obtained via simple and fast 2D integration and clipping of the values.

In the future, we will systematically assess the algorithm performance and we will provide more extensive experimentation including subjective testing (the best transformation requires judgment of photographers and painters). Moreover, we will involve the multiscale processing to make the proposed method real-time.

Acknowledgements

This work has been partially supported by the Ministry of Education, Youth and Sports of the Czech Republic under the research programs MSM 6840770014 and LC-06008; and through the MOMARNET EU Research and Training Network project (MRTN-CT-2004-505026). Special thanks to Olivier Delaunoy for his fruitful comments.

References

[BE04] BALA R., ESCHBACH R.: Spatial color-to-grayscale transform preserving chrominance edge information. In *Color Imaging Conference* (2004), pp. 82–86.

[FLW02] FATTAL R., LISCHINSKI D., WERMAN M.: Gradient domain high dynamic range compression. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques* (2002), ACM Press, pp. 249–256.

[GD05] GRUNDLAND M., DODGSON N. A.: *The Decolorize Algorithm for Contrast Enhancing, Color to Grayscale Conversion*. Tech. Rep. UCAM-CL-TR-649, University of Cambridge, 2005.

[GOTG05] GOOCH A. A., OLSEN S. C., TUMBLIN J., GOOCH B.: Color2gray: salience-preserving color removal. *ACM Trans. Graph. 24*, 3 (2005), 634–639.

[Hun92] HUNT R. W. G.: *Measuring Colour*, 2nd ed. Ellis Horwood Series in Applied Science and Industrial Technology, 1992.

[Nem80] NEMCSICS A.: Coloroid Color System. *Color Research and Application 5* (1980), 113– 120.

[Nem87] NEMCSICS A.: Color space of the coloroid color system. *Color Research and Application 12* (1987), 135– 146.

[Nem01] NEMCSICS A.: Recent experiments investigating the harmony interval based colour space of the coloroid colour system. In *AIC 9th Congress Rochester* (2001).

[Nem04] NEMCSICS A.: *Colour Dynamics, Environmental Colour Design*, 2nd ed. Akadémiai Kiadó, Budapest, 2004.

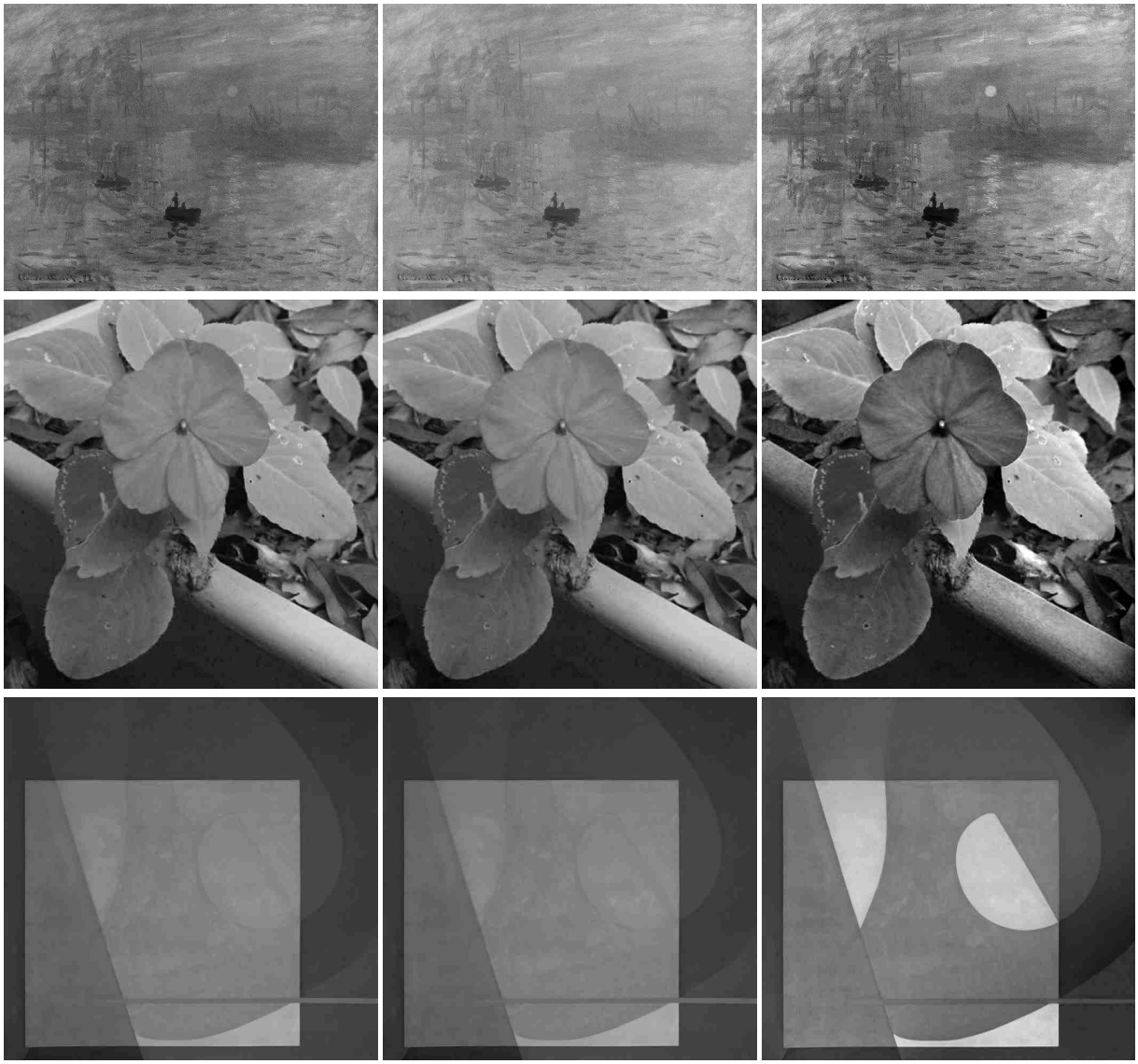


Figure 7: *Left column: original color image, middle column: CIE Y equivalent, right column: our adaptive color to gray transformation result.*

[NNN05] NEUMANN L., NEMCSICS A., NEUMANN A.: Computational color harmony based on coloroid system. In *Computational Aesthetics in Graphics, Visualization and Imaging 2005* (2005), Neumann L., Sbert M., Gooch B., Purgathofer W., (Eds.), pp. 231– 240.

[RGW05] RASCHE K., GEIST R., WESTALL J.: Re-coloring Images for Gamuts of Lower Dimension. *Computer Graphics Forum 24*, 3 (2005), 423–432.

[SKM87] STRICKLAND R. N., KIM C.-S., MCDONNELL W. F.: Digital color image enhancement based on the saturation component. *Optical Engineering 26* (July 1987), 609–616.

[Sze06] SZELISKI R.: Locally adapted hierarchical basis preconditioning. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers* (New York, NY, USA, 2006), ACM Press, pp. 1135–1143.

[ZW96] ZHANG X., WANDELL B.: A spatial extension of cielab for digital color image reproduction. In *Proc. Soc. Inform. Display 96 Digest* (San Diego, 1996), pp. 731– 734.

Appendix E

Perceptual Evaluation of Color-to-Grayscale Image Conversions

M. Čadík. Perceptual Evaluation of Color-to-Grayscale Image Conversions. *Computer Graphics Forum*, Vol. 27, No. 7, pp. 1745–1754, 2008.
IF=1.595

Pacific Graphics 2008
T. Igarashi, N. Max, and F. Sillion
(Guest Editors)

Volume 27 (2008), Number 7

Perceptual Evaluation of Color-to-Grayscale Image Conversions

M. Čadík[†1]

[1]Czech Technical University in Prague, Czech Republic

Abstract

Color images often have to be converted to grayscale for reproduction, artistic purposes, or for subsequent processing. Methods performing the conversion of color images to grayscale aim to retain as much information about the original color image as possible, while simultaneously producing perceptually plausible grayscale results. Recently, many methods of conversion have been proposed, but their performance has not yet been assessed. Therefore, the strengths and weaknesses of color-to-grayscale conversions are not known. In this paper, we present the results of two subjective experiments in which a total of 24 color images were converted to grayscale using seven state-of-the-art conversions and evaluated by 119 human subjects using a paired comparison paradigm. We surveyed nearly 20000 human responses and used them to evaluate the accuracy and preference of the color-to-grayscale conversions. To the best of our knowledge, the study presented in this paper is the first perceptual evaluation of color-to-grayscale conversions. Besides exposing the strengths and weaknesses of the researched methods, the aim of the study is to attain a deeper understanding of the examined field, which can accelerate the progress of color-to-grayscale conversion.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation–Display algorithms, viewing algorithms; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism–Color, shading, shadowing, and texture; I.4.3 [Image Processing and Computer Vision]: Enhancement–Filtering; J.4 [Social and Behavioral Sciences]: Psychology

1. Introduction

Converting color images to grayscale is used for various reasons, like for reproducing on monochrome devices, subsequent processing, or for aesthetic intents. Color-to-grayscale conversions perform a reduction of the three-dimensional color data into a single dimension, seen in Figure 1. It is evident that some loss of information during the conversion is inevitable, so the goal is to save as much information from the original color image as possible. At the same time, the aim is also to produce perceptually plausible grayscale results. Recently, various approaches to the color to grayscale conversion problem have been proposed. While the problem's complexity is currently recognized, the performance of existing solutions is not. Even though researchers frequently claim that their methods advance the field with respect to previous ones, it is important to evaluate the performance of these algorithms in comparative, subjective experiments and analyze their strengths and weaknesses. However, until now, there has not been an evaluation of color-to-grayscale conversions involving a representative number of subjects and input stimuli.

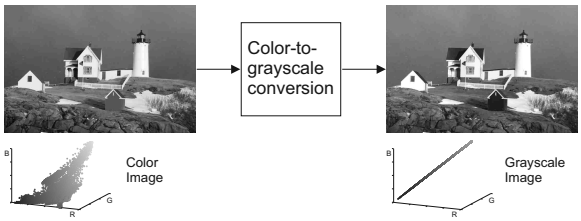


Figure 1: *The color to grayscale image conversion.*

In this paper, we present the results of two subjective perceptual experiments (preference and accuracy), for which

† cadikm@fel.cvut.cz http://www.cgg.cvut.cz/~cadikm

seven state-of-the-art color-to-grayscale conversions were evaluated by 119 human subjects. The set of inputs consisted of 24 various color images. By means of statistical analysis of the subjective experimental data, we assess the strengths and weaknesses of the conversions, with respect to the preference and accuracy of color reproduction. The overall results show that the best score for *accuracy* is achieved by the approach of Smith et al. [SLJT08], while the most *preferred* method is Decolorize [GD05]. The method of Bala and Eschbach [BE04] was ranked the worst in both the accuracy and preference experiments. Furthermore, we aim to attain a deeper insight into the color-to-grayscale conversion field.

The rest of this paper is structured as follows: In Section 2 we first survey the related work. In Section 3 we introduce the two perceptual experiments that we have conducted. In Section 4 we present, analyze and discuss the results of the experiments. Finally, in Section 5 we conclude and suggest some ideas for future research.

2. Related Work

In this section, we give an overview of current state-of-the-art, color-to-grayscale conversions. Most of the described methods are evaluated in our perceptual study (please, refer to Section 3.1 and Table 1). We also survey existing evaluations of color-to-gray conversions and related studies.

2.1. Color-to-Grayscale Image Conversions

The simplest and widely used approach to converting color to grayscale is based on neglecting of the chrominance channels, e.g. taking a luminance channel as a grayscale representation of the original color image. One of the possibilities is to utilize the Y channel of the CIE XYZ [Fai05] color space. This approach is simple and computationally efficient, but it may fail for specific images, such as those with isoluminant colors.

Bala and Eschbach [BE04] propose a spatial approach to color-to-grayscale conversion. They preserve chrominance edges locally by introducing high-frequency chrominance information into the luminance channel. A spatial high-pass filter is applied to the chromatic channels, the output is weighted with a luminance-dependent term, and the final result is added to the luminance channel.

Grundland and Dodgson [GD05] propose the *Decolorize* algorithm for contrast enhancement as well as converting color to grayscale. They perform a global grayscale conversion by expressing grayscale as a continuous, image-dependent, piecewise linear mapping of the primary RGB colors and their saturation. Three parameters are used to control contrast enhancement, scale selection and noise suppression, and image-independent default values for these parameters have been proposed [GD05].

A different approach was taken by Gooch et al. [GOTG05], who introduced the local algorithm known as *Color2Gray*. In this gradient-domain method, the gray value of each pixel is iteratively adjusted to minimize an objective function, which is based on local contrasts between all the pixel pairs. The computational complexity of this method is high ($O(N^4)$), and can be improved by limiting the number of considered differences (e.g. by color quantization). Mantiuk et al. [MMS06] show an application of their contrast processing framework to accelerate the Color2Gray [GOTG05] method. In their approach, the close neighborhood of a pixel is considered on fine levels of a pyramid, whereas the far neighborhood is covered on coarser levels. The authors claim that this enables them to convert bigger images and perform computations faster.

Another conversion was introduced by Rasche et al. [RGW05]. Their method aims to preserve contrast while maintaining consistent luminance. The authors formulate an error-function based on matching the gray differences to the corresponding color differences. The goal is minimizing the error function to find an optimal conversion. The authors propose using color quantization to reduce the considerable computational costs of the error-minimization procedure.

Queiroz and Braun [dQB06] have proposed an *invertible* conversion to grayscale. The idea is to transform colors into high-frequency textures that are applied onto the gray image and can be later decoded back to color. The method is based on wavelet transformations and on the replacement of subbands by chrominance planes.

Alsam and Kolas [AK06] introduced a conversion method that aims to create sharp grayscale from the original color rather than enhancing the separation between colors. The approach resembles the method of Bala and Eschbach [BE04]: first, a grayscale image is created by a global mapping to the image-dependent gray axis. Then, the grayscale image is enhanced by a correction mask in a way similar to unsharp masking [GW02].

Neumann et al. [NČN07] proposed two local, gradient-based, color-to-grayscale conversions. The first is a generalization of the CIELab formula [Fai05], which introduces a signed power function to give a signum to the weighted Lab components. The second technique aims to obtain the best perceptual gray gradient equivalent by exploiting the Coloroid system and its experimental background. The gradient field constructed using one of the techniques is corrected using a gradient inconsistency correction method. Finally, a 2D integration yields the grayscale image.

A recent method by Smith et al. [SLJT08] combines global and local conversions in a way similar to Alsam and Kolas [AK06]. The method first applies global "absolute" mapping based on the Helmholtz-Kohlrausch effect, and then locally enhances chrominance edges using adaptively-weighted multiscale unsharp masking. While the global

conversion	reference	G/L	implementation	parameters
CIE Y	[Fai05]	G	own implementation, C++	—
Bala04	[BE04]	G + L	own implementation, C++	N=3, K=1, B1=15, B2=40
Decolorize	[GD05]	G	www.eyemaginary.com, Matlab	effect=0.5, scale=25, noise=10^{-3}
Color2Gray	[GOTG05]	L	www.color2gray.info, command_line, C++	colors=256, θ=45, α=10, μ=full
Rasche05	[RGW05]	L	www.fx.clemson.edu/~rkarl, c2g_i, C	colors=256, exp=2, threshold=15
Neumann07	[NČN07]	L	www.cgg.cvut.cz/~cadikm/color_to_gray, own impl., C++	$\varepsilon = 10^{-5}$
Smith08	[SLJT08]	G + L	www.mpi-inf.mpg.de/resources/ApparentGreyscale, 1-scale, Gimp	rad=5, amount=0.15, gamma=1

Table 1: *Summary of the evaluated color-to-grayscale conversion methods. G and L stands for global and local, respectively.*

mapping is image independent, the local enhancement reintroduces lost discontinuities only in regions that insufficiently represent the original chromatic contrast [SLJT08]. The goal of the method is perceptual accuracy, not the exaggeration of discriminability.

2.2. Evaluations of Color-to-Grayscale Conversions

Apart from simple evaluations of the proposed methods surveyed below, we are not aware of any subjective perceptual evaluation study of color-to-grayscale conversions.

Bala and Eschbach [BE04] performed a small preference experiment to evaluate the qualitative performance of their conversion. The authors used three input color images that were converted using their novel method and by the simple conversion that retains the luminance component. The grayscale results were presented as hardcopy prints to six observers. The subjects preferred the novel spatial conversion approach (16 positive decisions out of total $6\times3 = 18$ comparisons).

Rasche et al. [RGW05] performed an accuracy experiment (with reference images) to assess their color-to-grayscale conversion. Six color images converted by the standard mapping of luminance to gray and by Rasche's method were presented to a group of 17 observers. The results revealed that for one group of input images the performance of the evaluated conversions was comparable, while for the second group of images, Rasche's method outperformed the traditional conversion.

3. Perceptual Experiments

In this section we describe the specific details of perceptual experiments that we have conducted to evaluate tested color-to-grayscale image conversions. We utilized the psychophysical technique of paired comparisons [Dav88], namely the two-alternatives forced choice (2AFC) experiment paradigm. We performed two experiments: in the first experiment (for accuracy), the grayscale images were presented along with the original (reference) color image, and in the second experiment (for preference), the subjects saw two grayscale images without any reference.

3.1. Evaluated Color-to-Grayscale Conversions

In total, we evaluated seven color-to-grayscale conversions, summarized in Table 1. When available, we utilized the codes provided by the authors for a particular conversion, but otherwise we implemented the conversion personally. All the conversions were run using default (constant) parameter settings (please, refer to Table 1 for numerical values). We decided to use constant parameters over all the input images for several reasons: first, to ensure comparable conditions for all the conversion methods involved; second, to reduce the number of images that are presented to subjects; and lastly not to bias the results by choice (tweaking of parameters) of an experimenter or an author (as different people may have a different sense of what is the best grayscale image).

3.2. Input Images

One of the advantages of a good-quality color-to-grayscale conversion is to give compelling results over a wide range of input images. We used 24 input color images in our study, with various motifs, origins, gamuts, etc, (the collection of these images is shown in Table 4 on Page 9). The images depict plants (images 9, 13, 23), foliage (22), fruits & vegetables (1, 10), portraits (11, 16), various photos (3, 4, 14, 15, 19), paintings (6, 20), cartoons (5, 21), color testing images (2, 7, 8, 12, 17), and computational images (18, 24). All the images were rescaled to maximally span 390×390 pixels for presentation purposes (to fit on the screen with the reference image) and also for the computational demands of several conversions.

3.3. Experimental Design

The evaluated images were displayed on a characterized and calibrated monitor EIZO S1910, a 19-inch LCD display, in native resolution 1280×1024 pixels. Calibration was performed by X-Rite GretagMacbeth Eye-One Display 2 colorimeter to D65, 120 cd/m^2, and colorimetrically characterized by measured ICC profiles. The experimental images were presented on a neutral gray background with a luminance of 18% of the white point. The experimentation room was neutrally painted, darkened (measured light level: 4 lux), and observers sat approximately 70 cm from the display. All testing was performed approximately in the same

time of day (before noon) to avoid fatigue or other factors. The total of 121 observers took part in our experiments. The observers were both male and female between the ages of 18 to 41, and all of them reported to have normal, or corrected-to-normal vision. Each subject was verbally introduced to the problem before the experiment, as described in the following section.

3.4. Experimental Procedure

The design of the experiments followed the 2AFC approach [Dav88]. Specifically, we utilized the software 'Ranker' which is available at `ranker.sourceforge.net`. Every grayscale image was compared with every other grayscale image (see Table 5 on Page 10), i.e. for each input color image, we have $n(n-1)/2 = (7 \times 6)/2 = 21$ comparisons, where $n = 7$ evaluated conversions. With 24 input color images, we would need $24 \times 21 = 504$ trials, which would be prohibitive for each subject. Therefore, we ran a pilot study to assess the reasonable amount of trials for one observer (and to verify the setup as well). The pilot study indicated that eight sets of grayscale images (21 comparisons in each), i.e. the 168 trials, is an acceptable quantity for one observer without experiencing exhaustion and loss of concentration. With eight randomly selected sets (balanced design), the whole experiment took approximately 20 minutes per observer. The sequence of images and the position of images on the display (left or right) were randomized. The type of the experiment (accuracy or preference) was also randomized, however for a given observer it remained constant.

Experiment with a reference (accuracy): every time, two grayscale images were displayed along with the color original in the middle. Observers were asked to select the one of the two grayscale images that was closer in appearance to the original color image, i.e. to select the image that better reproduced the original. More specifically, the instructions stated: "Your task is to select the grayscale image that better matches the colors of the original color image."

Experiment without a reference (preference): every time, two grayscale images were displayed. Observers were instructed to select the grayscale image that they preferred. Specifically, the instructions stated: "Your task is to select the preferred grayscale image from the presented pair." Generally, accuracy (with reference) experiments were slightly more time-demanding with comparison to preference (without reference) experiments, and took 20 to 30 minutes per observer.

4. Results and Discussion

A total of 121 observers completed 20328 observations (pair-wise comparisons). Based on a post-test questionnaire, the results of two observers were excluded as outliers because of color vision deficiencies. In the following, we present the results based on the observations of 60 participants who performed the accuracy experiment and 59 subjects who took part in the preference experiment. For each trial, the grayscale image chosen by an observer was given a score of 1, the other a score of 0. The data were stored in a 7×7 frequency matrix for each observer, where the value in column i and row j represents the score of grayscale conversion i compared with conversion j. We used Thurstone's Law of Comparative Judgments, Case V, to convert the data into interval z-score (standard score) scales [Thu27, Eng00].

Source of Variation	SS	$d.f.$	MS	F	p
conversion	105.6	6	17.6	185.4	≈ 0
experiment	0	1	0	0	≈ 1
input image	0	23	0	0	≈ 1
conversion × experiment	2.8	6	0.5	4.9	10^{-4}
conversion × input image	260.1	138	1.9	19.9	≈ 0
experiment × image	$\approx$ -0	23	$\approx$ -0	$\approx$ -0	≈ 1
Residual	13.1	138	0.1		
Total	381.5	335			

Table 2: *The results of multi-factorial ANOVA test (where SS denotes Sum of Squares, $d.f.$ means Degrees of Freedom, MS denotes Mean Square, F is F value, and p is p-value for the null hypothesis [TF07]).*

As the z-scores calculated from the observation data using Thurstone's law are normally distributed, we can utilize classical parametric statistics in the further analysis. To inquire the significance of the *input images*, the *experiments* (accuracy and preference), and the *conversions* (i.e. the factors) on the observation data, it is profitable to apply the multi-factorial analysis of variance (ANOVA) test [TF07]. Multi-factorial (n-way) ANOVA is able to consider all the factors at once. The results of the n-way ANOVA are summarized in an ANOVA table [MR99] (Table 2). The results show that the only significant *main effect* is the conversion (because the p-value is below the threshold of 0.05), which means that there are significant differences in the performances of the inquired conversions. Neither the experiment type, nor the input image can alone explain the variability in the data. However, two statistically significant *interaction effects* imply that the observed scores depend on the combination of the conversion and the input image, and (with the smallest probability, but still with a statistical significance) on the combination of the conversion and the type of the experiment. This result suggests that the performances of the conversions depend on input images and on experiment type, and it makes sense to show the results separately for each input image and for each experiment. Finally, we performed a multiple comparison test (Tukey's honestly significant differences [HT87]) over all the subjective data. This test re-

Decolorize	Smith08	CIE Y	Color2Gray	Rasche05	Neumann07	Bala04
0.544	0.487	0.158	0.149	-0.203	-0.317	-0.819

Figure 2: *Overall performances of the inquired conversions. Results of the multiple comparison across all input images in both experiments. The best result is the leftmost, any conversions that are underlined are considered perceptually similar.*

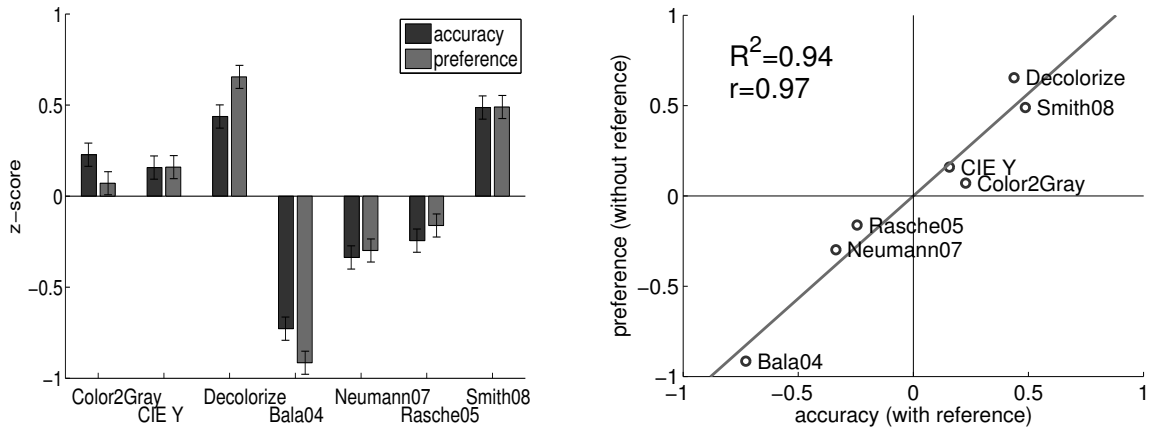


Figure 3: *Overall results separately for the two experiments. Left: overall scores for both the accuracy and preference experiments. Error bars show intervals of 95% confidence. Right: comparison of accuracy and preference experiments.*

turns an overall ranking of the conversions with the indication of the statistical significance of the differences between them (please, see Figure 2). The results show that the best ranked conversion in our study is Decolorize, but it performs statistically similar to Smith08; the worst ranked is Bala04.

4.1. Overall Accuracy and Preference Results

The overall scores were obtained by averaging the percentage matrices over all input images separately for the accuracy and preference experiments (please, see Figure 3). We can see from the overall results that altogether the best score in the accuracy experiment was achieved by Smith08, while Decolorize produces the most preferred grayscale images. Bala04 was ranked the worst in both the accuracy and preference experiments.

Comparing the overall accuracy and preference scores, we see similar trends in the results of the experiments. The calculated Pearson correlation coefficient [MR99] $r = 0.97$ and the coefficient of determination [MR99] $R^2 = 0.94$ (Figure 3 right) indicate high similarity of the preference and overall accuracy experiments. Notice that the CIE Y and Smith08 methods exhibit almost unchanged performance in both experiments. On the other hand, the rest of the methods show certain differences in accuracy and preference experiments. Specifically, Decolorize, Neumann07, and Rasche05 perform better in the preference experiment than in the accuracy experiment. On the contrary Color2Gray and Bala04 perform better in the accuracy experiment than in the preference experiment. Please refer to Section 4.3 for further analysis of accuracy and preference.

4.2. Results for Individual Images

Next, we examined the experimental data for all the color images individually (please see the summarized results in Table 3). We converted the observation data into z-scores independently for each input image using the Thurstone's Law of Comparative Judgments. The ranking reported in Table 3 is based on the calculated z-scores. The coefficient of agreement between subjects u ranges from $u = -1/(s-1)$, where s is the number of subjects, (which indicates no agreement between subjects) to $u = 1$ (all subjects responded the same). We show the results of the χ^2 test on the coefficient u, and the obtained p-values. The coefficient of consistency of subject's responses ζ ranges from $\zeta = 0$ (no consistence) to $\zeta = 1$ (ideally consistent responses), we report the average ζ over the subjects for a given input image. The values of u, ζ, χ^2, and p were calculated in a similar way to Ledda et al. [LCTS05].

The results of the χ^2 test show that there is some agreement between observers, seen by the reported statistical significance (all the p-values of the null hypothesis are clearly below the threshold). This means that there are differences the in performances of the conversions, which is also revealed by the ANOVA test reported above. The high values of ζ suggest that each subject was fairly consistent in their judgments. On the other hand, the agreement u amongst subjects varies from high values (images 2, 8) to lower agreement (for images 3, 9, 11), which indicates that the complexity of judgments differ depending on the input image.

Table 3 shows that no conversion produces universally satisfying results for all involved input images. Each of the

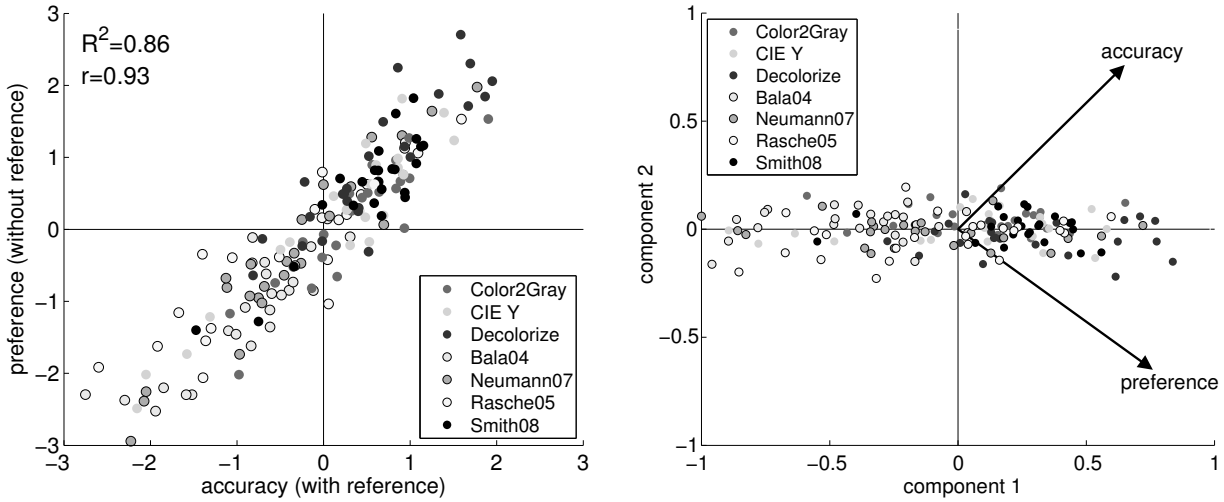


Figure 4: *The results for individual input images. Left: Accuracy-preference scores for all input images. Each point represents a score of one color-to-gray conversion method for a particular input image. Right: Principal component analysis. The axes represent the principal components and the points represent the principal component scores of one conversion for one input image. The vectors show the values of principal component coefficients for the accuracy and preference variables.*

seven tested conversions was ranked as worst for at least one input image and, apart from Bala04, each conversion was ranked as best for some input image. It is interesting to notice that Decolorize exhibits exceptionally good results for those input images that have rather narrow color gamuts or a limited number of colors (i.e. the images 2, 5, 7, 8, 10, 17, 20), refer to Table 4. For such images it is possible that, the image-dependent global mapping of Decolorize performs very well. Contrarily, Smith08 excels at colorful images with extensive color gamuts (4, 9, 15, 19, 22, 23), where the locally enhanced mapping based on the Helmholtz-Kohlrausch effect outperforms other conversions. Of note, however, is that the simple CIE Y conversion also performs quite well for these input images and it is remarkably good in images 5, 16, 18, 19, and 21.

4.3. Accuracy vs. Preference

We calculated values of the correlation coefficient r and the coefficient of determination R^2 [MR99] to determine the relationship between the accuracy and preference scores. The high values of r and R^2 for *overall* accuracy and preference scores (Figure 3, right) as well as for the scores for *individual* images (Figure 4, left) imply that there is a strong correlation between peoples' judgments of the color-to-grayscale conversion accuracy and the grayscale image preference. This suggests that one aspect dominates subjective judgment – let us call it an overall perceptual quality of color-to-grayscale conversion. The high values of correlations are interesting, as one would expect tricky judgments for grayscale pairs without the reference of some input images (e.g. 6, 7, 12, 17). The values of u and ζ, however, imply that the subjects were rather consistent in their opinions.

The principal component analysis [TF07] results in two principal components, where the first principal component explains 96.4% of the data variance (Figure 4, right). As illustrated, the first component (perhaps the overall quality of the conversion) lies nearly perfectly in the axis of accuracy and preference vectors. This result supports the above idea that only one dimension prevails in our subjective data.

4.4. Comparison to Previous Work

We believe that the presented study is much more credible than the two simple evaluations described in Section 2.2, as the number of subjects, input images and evaluated conversions is much higher. However, it is interesting and fair to compare the results obtained with the results of the previous evaluations.

In the preference experiment of Bala and Eschbach [BE04], Bala04 performed better than the mapping retaining the luminance. The authors used three input images (two of them are very similar to this study's image14 and image21). In our preference experiment, CIE Y and Bala04 performed similarly for image14, and Bala04 performed worse than CIE Y for image21. In our overall results, Bala04 performed worse than CIE Y, which is not consistent with findings of Bala and Eschbach. Besides the higher number of observers in our experiment, the discrepancy in the two studies is perhaps due to the different experimental setups, since Bala and Eschnach presented hardcopy prints and we utilized an LCD monitor.

Rasche's [RGW05] results show that for four input images, the performance of Rasche05 is comparable to the standard mapping of luminance. For another three images,

	u	ζ (avg)	χ^2	p (21 d.f.)	best		ranking of scores				worst
image1 (accuracy)	0.191	0.833	101.2	p<0.001	C	Y	S	B	D	N	R
image1 (preference)	0.290	0.832	136.8	p<0.001	C	Y	S	D	R	B	N
image2 (accuracy)	0.713	0.966	290.4	p<0.001	N	D	R	C	S	Y	B
image2 (preference)	0.804	0.981	324.9	p<0.001	D	N	R	C	S	Y	B
image3 (accuracy)	0.103	0.673	64.2	p<0.001	R	Y	B	S	N	C	D
image3 (preference)	0.134	0.696	74.4	p<0.001	S	R	Y	B	N	C	D
image4 (accuracy)	0.326	0.827	158.0	p<0.001	S	Y	N	D	C	B	R
image4 (preference)	0.585	0.893	254.4	p<0.001	S	Y	N	D	C	B	R
image5 (accuracy)	0.489	0.929	226.5	p<0.001	Y	D	S	C	B	R	N
image5 (preference)	0.561	0.946	245.0	p<0.001	D	S	Y	R	C	B	N
image6 (accuracy)	0.468	0.876	197.7	p<0.001	R	D	N	C	S	Y	B
image6 (preference)	0.550	0.891	228.9	p<0.001	R	N	D	C	S	Y	B
image7 (accuracy)	0.258	0.876	118.6	p<0.001	D	Y	R	C	B	N	S
image7 (preference)	0.425	0.925	199.5	p<0.001	D	R	Y	C	B	N	S
image8 (accuracy)	0.567	0.929	235.2	p<0.001	D	N	R	C	S	B	Y
image8 (preference)	0.667	0.977	273.1	p<0.001	D	N	R	S	C	B	Y
image9 (accuracy)	0.106	0.771	60.9	p<0.001	S	D	R	Y	B	N	C
image9 (preference)	0.199	0.737	96.3	p<0.001	S	D	Y	R	N	B	C
image10 (accuracy)	0.162	0.853	82.4	p<0.001	D	S	R	Y	N	B	C
image10 (preference)	0.484	0.861	204.1	p<0.001	D	S	Y	R	N	B	C
image11 (accuracy)	0.138	0.703	73.1	p<0.001	S	R	Y	D	C	N	B
image11 (preference)	0.186	0.737	91.2	p<0.001	S	R	D	B	Y	N	C
image12 (accuracy)	0.564	0.940	234.4	p<0.001	C	N	D	S	R	B	Y
image12 (preference)	0.552	0.956	252.8	p<0.001	C	D	N	S	R	Y	B
image13 (accuracy)	0.307	0.846	137.1	p<0.001	N	Y	S	C	D	B	R
image13 (preference)	0.146	0.803	82.1	p<0.001	D	C	Y	S	N	B	R
image14 (accuracy)	0.288	0.756	129.9	p<0.001	S	Y	C	D	N	B	R
image14 (preference)	0.173	0.671	90.2	p<0.001	D	C	S	N	Y	B	R
image15 (accuracy)	0.256	0.801	117.7	p<0.001	S	R	D	C	Y	B	N
image15 (preference)	0.247	0.786	124.8	p<0.001	S	R	Y	C	D	N	B
image16 (accuracy)	0.217	0.827	112.2	p<0.001	C	S	Y	R	D	N	B
image16 (preference)	0.372	0.868	169.4	p<0.001	Y	S	C	D	N	R	B
image17 (accuracy)	0.333	0.908	161.0	p<0.001	D	S	R	N	Y	B	C
image17 (preference)	0.391	0.929	177.2	p<0.001	D	S	N	Y	R	B	C
image18 (accuracy)	0.231	0.762	118.0	p<0.001	Y	S	C	D	B	N	R
image18 (preference)	0.247	0.736	119.6	p<0.001	Y	S	C	D	R	N	B
image19 (accuracy)	0.273	0.842	124.1	p<0.001	Y	S	C	D	B	N	R
image19 (preference)	0.409	0.867	192.6	p<0.001	S	Y	C	D	B	R	N
image20 (accuracy)	0.520	0.861	217.5	p<0.001	D	C	S	N	Y	R	B
image20 (preference)	0.530	0.895	243.7	p<0.001	D	S	C	N	Y	R	B
image21 (accuracy)	0.462	0.951	195.6	p<0.001	Y	S	D	C	N	B	R
image21 (preference)	0.538	0.977	224.3	p<0.001	Y	D	S	C	N	B	R
image22 (accuracy)	0.484	0.861	204.1	p<0.001	S	C	R	D	Y	N	B
image22 (preference)	0.491	0.880	206.6	p<0.001	S	R	C	Y	D	N	B
image23 (accuracy)	0.406	0.840	191.5	p<0.001	S	C	Y	B	R	D	N
image23 (preference)	0.390	0.832	176.8	p<0.001	S	Y	C	R	B	D	N
image24 (accuracy)	0.303	0.797	135.4	p<0.001	S	C	Y	D	R	B	N
image24 (preference)	0.296	0.837	145.4	p<0.001	D	Y	C	S	R	B	N

Table 3: *The results for individual input images. Used abbreaviations: avg=average, d.f.=degrees of freedom, **C** =Color2Gray, **Y** =CIE Y, **D** =Decolorize, **B** =Bala04, **N** =Neumann07, **R** =Rasche05, **S** =Smith08, notice that the used colors are equivalent to the colors in Figure 4.*

the second one from the evaluated adjustments of parameters of Rasche05 outperforms the traditional conversion. In our experiment, the overall accuracy score of Rasche05 is close to CIE Y, but it is worse than CIE Y with statistical significance. Rasche05 outperforms CIE Y only for 11 of 24 input images (i.e. for images 2, 3, 6, 8, 9, 10, 11, 12, 15, 17, 22). The reason why Rasche05 performs worse in our study than in the Rasche experiment is due to the fact that we applied Rasche05 with constant parameters (alike the other conversions, seen in Table 1). We admit that Rasche05 could be ranked better after a thorough parameter tuning for each image (and other conversions, too), however this was not the objective of our study (please, refer to the discussion in Section 3.1).

5. Conclusions and Future Work

We presented a perceptual evaluation of color-to-grayscale image conversions. In two experiments, a total number of 119 subjects assessed the accuracy and the preference of grayscale images produced by seven state-of-the-art conversion methods. The inputs of the evaluated conversions represented the set of 24 color images of varying characteristics, motifs, and acquisitions.

The results show that the Decolorize [GD05] and Smith04 [SLJT08] conversions are overall the best ranked approaches, and the approach of Bala04 [BE04] performed the worst. However, the analysis of individual images reveal that no conversion produces universally good results for all the involved input images. Specifically, each of the seven inquired conversions was ranked the worst for at least one input image and, apart from Bala04, each conversion was ranked the best for some input image. These results suggest that there still exist areas for improvement of current conversions, especially in the robustness over various inputs. Furthermore, we found a high degree of correspondence between the accuracy and preference scores. Specifically, the results indicate that one dimension prevails in the subjects' judgment of the quality of the grayscale results. We believe that this is of particular importance and it is necessary to conduct experimental subjective studies, such as the one presented, to validate and evaluate color-to-grayscale conversions properly in order to expose their strengths and weaknesses, and to attain a deeper understanding of the examined field.

The presented study does not reflect computational demands, implementation difficulties, and other factors, which can play an important role for practical use. Notice that our results are valid for images presented on a screen, and the tested conversions may perform differently for hardcopy printouts or other media. Moreover, the desirable properties of the color-to-grayscale conversion may sometimes depend on the chosen application. In future work, we plan to implement all the conversions in the same platform to assess their computational demands and their actual usefulness. We will also research how to involve more input parameters of the conversions so as to explore the parameter space.

Acknowledgements

This work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic (research programs MSM 6840770014 and LC-06008), and by the Aktion OE/CZ grant no. 48p11. Special gratitude to M. Kalouš who coded the 'Ranker', Z. Míkovec, I. Malý, and O. Poláček for help in carrying out the experiments, J. Křivánek and J. Bittner for valuable comments, and to all the participants in the experiments for time and patience.

References

[AK06] ALSAM A., KOLAS O.: Grey colour sharpening. In *Proc. of 14th Color Imaging Conf.* (2006), IS&T & SID, pp. 263– 267.

[BE04] BALA R., ESCHBACH R.: Spatial color-to-grayscale transform preserving chrominance edge information. In *Color Imaging Conference* (2004), IS&T, pp. 82–86.

[Dav88] DAVID H. A.: *The Method of Paired Comparisons*, 2nd ed. Oxford University Press, 1988.

[dQB06] DE QUEIROZ R. L., BRAUN K. M.: Color to gray and back: color embedding into textured gray images. *IEEE Trans. on Image Processing 15* (June 2006), 1464–1470.

[Eng00] ENGELDRUM P. G.: *Psychometric scaling: a toolkit for imaging systems development*, 1st ed. Imcotek Press, 2000.

[Fai05] FAIRCHILD M. D.: *Color Appearance Models*, 2nd ed. Wiley-IS&T, Chichester, UK, 2005.

[GD05] GRUNDLAND M., DODGSON N. A.: *The Decolorize Algorithm for Contrast Enhancing, Color to Grayscale Conversion*. Tech. Rep. UCAM-CL-TR-649, University of Cambridge, 2005.

[GOTG05] GOOCH A. A., OLSEN S. C., TUMBLIN J., GOOCH B.: Color2Gray: salience-preserving color removal. *ACM Trans. Graph. 24*, 3 (2005), 634–639.

[GW02] GONZALEZ R. C., WOODS R. E.: *Digital Image Processing*, 2nd ed. Prentice-Hall, 2002.

[HT87] HOCHBERG Y., TAMHANE A. C.: *Multiple Comparison Procedures*, 1st ed. Wiley, 1987.

[LCTS05] LEDDA P., CHALMERS A., TROSCIANKO T., SEETZEN H.: Evaluation of tone mapping operators using a high dynamic range display. *ACM Trans. Graph. 24*, 3 (2005), 640–648.

[MMS06] MANTIUK R., MYSZKOWSKI K., SEIDEL H.-P.: A perceptual framework for contrast processing of high dynamic range images. *ACM Trans. Appl. Perc. 3*, 3 (2006), 286–308.

[MR99] MONGOMERY D. C., RUNGER G. C.: *Applied Statistics and Probability for Engineers*, 2nd ed. John Wiley & Sons, 1999.

[NČN07] NEUMANN L., ČADÍK M., NEMCSICS A.: An efficient perception-based adaptive color to gray transformation. In *Proc. of Computational Aesthetics 2007* (2007), Eurographics Association, pp. 73– 80.

[RGW05] RASCHE K., GEIST R., WESTALL J.: Re-coloring Images for Gamuts of Lower Dimension. *Computer Graphics Forum 24*, 3 (2005), 423–432.

[SLJT08] SMITH K., LANDES P.-E., JÖELLE THOLLOT K. M.: Apparent greyscale: A simple and fast conversion to perceptually accurate images and video. *Computer Graphics Forum 27*, 3 (2008).

[TF07] TABACHNICK B. G., FIDELL L. S.: *Using multivariate statistics*, 5th ed. Pearson Education, 2007.

[Thu27] THURSTONE L. L.: A law of comparative judgement. *Psychological Review 34* (1927), 273–286.

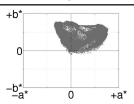
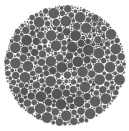
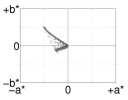
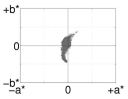
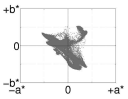
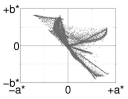
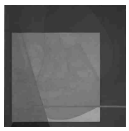
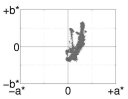
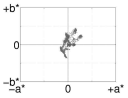
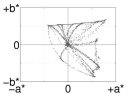
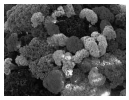
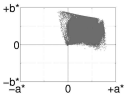
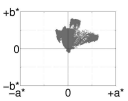
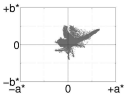
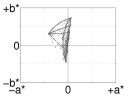
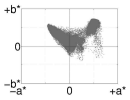
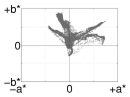
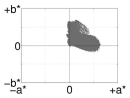
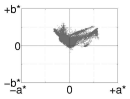
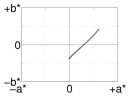
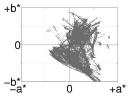
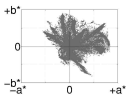
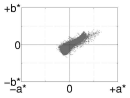
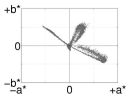
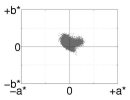
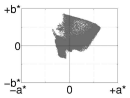
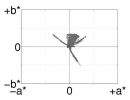
id	color image	color gamut	gamut [min, max]
image1			L* = [0.06151, 100] a* = [-35.32, 77.37] b* = [-10.55, 81.26]
image2			L* = [38.67, 100] a* = [-51.06, 0.6223] b* = [-19.64, 52.48]
image3			L* = [0, 100] a* = [-15.18, 20.87] b* = [-45.69, 38.37]
image4			L* = [0.7095, 99.71] a* = [-54.93, 40.19] b* = [-66.35, 53.9]
image5			L* = [0, 100] a* = [-71.94, 84.66] b* = [-92.34, 83.02]
image6			L* = [14.67, 96.38] a* = [-5.309, 38.68] b* = [-41.72, 68.38]
image7			L* = [64.75, 100] a* = [-16.63, 30.27] b* = [-36.28, 45.12]
image8			L* = [42.24, 57.86] a* = [-42.78, 79.86] b* = [-87.36, 69.06]
image9			L* = [0.4412, 100] a* = [-21.38, 79.57] b* = [-14.87, 91.16]
image10			L* = [0, 100] a* = [-25.91, 64.11] b* = [-11.55, 81.48]
image11			L* = [0, 100] a* = [-28.37, 66.11] b* = [-40.77, 52.23]
image12			L* = [62.76, 71.36] a* = [-41.11, 7.765] b* = [-46.83, 73.57]
image13			L* = [0.9717, 93.59] a* = [-46.35, 71.04] b* = [-38.17, 73.74]
image14			L* = [1.72, 98.19] a* = [-44.35, 74.49] b* = [-29.15, 90.95]
image15			L* = [5.421, 99.72] a* = [-6.689, 64.22] b* = [-17.33, 69.62]
image16			L* = [0, 99.3] a* = [-32.98, 60.36] b* = [-17.39, 61.53]
image17			L* = [56.07, 60.27] a* = [-1.697, 61.24] b* = [-38.39, 42.2]
image18			L* = [0, 100] a* = [-46.81, 82.9] b* = [-112.1, 88.87]
image19			L* = [0, 100] a* = [-55.68, 83.98] b* = [-81.79, 90.77]
image20			L* = [8.564, 81.58] a* = [-26.9, 65.64] b* = [-33.65, 40.39]
image21			L* = [3.012, 100] a* = [-55.65, 78.98] b* = [-47.86, 64.1]
image22			L* = [3.13, 100] a* = [-23.04, 31.68] b* = [-22.23, 37.5]
image23			L* = [0.7857, 98.24] a* = [-35.58, 63.81] b* = [-34.31, 87.03]
image24			L* = [0, 99.9] a* = [-34.36, 34.24] b* = [-50.49, 35.87]

Table 4: *The set of input images. Images courtesy of e-cobo.com (1), A. Gooch (2, 7, 8, 17), R. E. Barber (3), K. Rasche (4, 13, 22), imagekingdom.com (5), L. Neumann (6, 9, 12), Kodak (11, 14), UT Austin (15), Sony (16), Fujifilm (19), artcyclopedia.com (20), M. Čadík (21), and K. Odhner (24).*

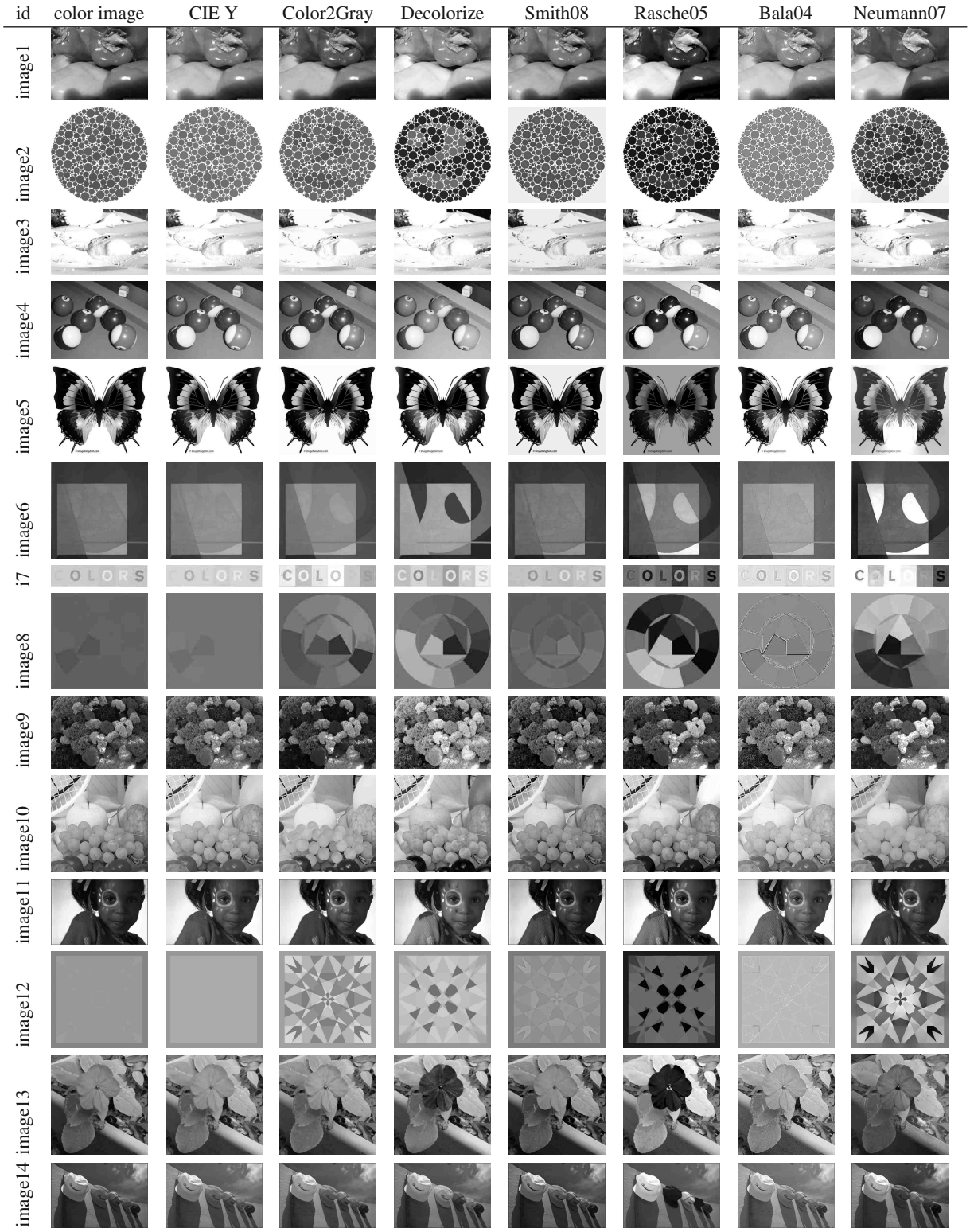


Table 5: *The results of the evaluated color-to-grayscale conversion methods. Please, refer to the accompanying webpage:* `http://www.cgg.cvut.cz/~cadikm/color_to_gray_evaluation` *for the complete set of the full-resolution images.*

Appendix F

Video Quality Assessment for Computer Graphics Applications

T. O. Aydın, M. Čadík, K. Myszkowski, and H.-P. Seidel. Video Quality Assessment for Computer Graphics Applications. *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia)*, Vol. 29, No. 6, pp. 161:1–161:12, Dec. 2010.
IF=3.725

Video Quality Assessment for Computer Graphics Applications

Tunç Ozan Aydın* Martin Čadík* Karol Myszkowski* Hans-Peter Seidel*
MPI Informatik

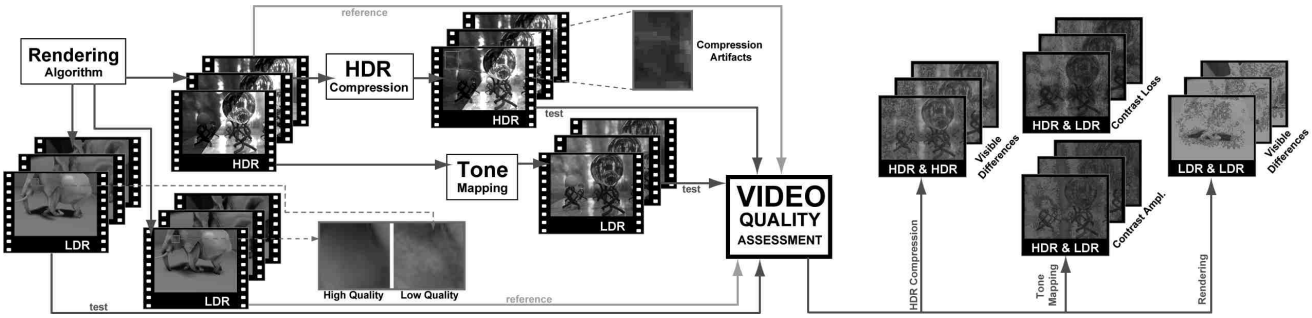


Figure 1: *The proposed metric predicts the perceived quality of natural as well as rendered video sequences with respect to a reference, even if the input videos have different dynamic ranges. Our work enables new applications including objective evaluation of video tone mapping and HDR compression.*

Abstract

Numerous current Computer Graphics methods produce video sequences as their outcome. The merit of these methods is often judged by assessing the quality of a set of results through lengthy user studies. We present a full-reference video quality metric[1] geared specifically towards the requirements of Computer Graphics applications as a faster computational alternative to subjective evaluation. Our metric can compare a video pair with arbitrary dynamic ranges, and comprises a human visual system model for a wide range of luminance levels, that predicts distortion visibility through models of luminance adaptation, spatiotemporal contrast sensitivity and visual masking. We present applications of the proposed metric to quality prediction of HDR video compression and temporal tone mapping, comparison of different rendering approaches and qualities, and assessing the impact of variable frame rate to perceived quality.

CR Categories: I.3.0 [Computer Graphics]: General; I.3.3 [Picture/Image Generation]: Display Algorithms—Viewing Algorithms

Keywords: video quality metrics, high dynamic range video, human visual perception, temporal artifacts, subjective video quality assessment

*e-mail: {tunc, mcadik, karol, hpseidel}@mpi-inf.mpg.de

[1]A web service that implements the metric described in this paper can be freely accessed at `http://drim.mpi-sb.mpg.de`.

1 Introduction

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 human visual system (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. 2008]. However, simply using image quality metrics to evaluate each frame of a video sequence fails to reflect the temporal aspects of Human Visual System's (HVS)

mechanisms, typically resulting in underestimating the visibility of temporal artifacts such as flickering (Sections 4, 5).

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 3). Due to the absence of a visual attention model, the metric predictions are conservative in the sense that they correspond to the perception of an observer who scrutinizes the entire video sequence. The results in Section 4 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 (Section 5). 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 6).

2 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.

2.1 Video Quality Assessment

Video quality assessment metrics often draw ideas from the more developed image quality assessment field. It has been quickly observed that simple statistics like signal-to-noise ratio are not necessarily correlated with human vision, which motivated HVS-based image quality metrics. Commonly used image quality metrics focus on near-threshold detection [Daly 1993], supra-threshold discrimination [Lubin 1995], or functional differences [Ferwerda and Pellacini 2003]. The proposed video quality metric makes use of a near-threshold human visual system model to comply with the needs of computer graphics applications.

The focus of the early work on video metrics has been extending image quality assessment metrics with temporal models of visual perception, resulting from the fact that frame-by-frame application of image quality metrics is not sufficient. 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 16th 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 4 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. 2008] (henceforth referred as DRIVDP).

2.2 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 prop-

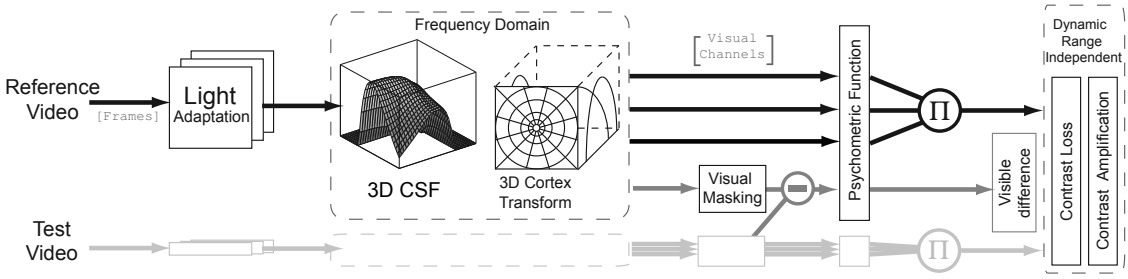


Figure 2: *The computational steps of our metric. Refer to text for details.*

erly. 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.

2.3 Temporal Aspects of Human Visual System

Temporal Visual Channels
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 $\sim 10\%$) cells. LGN, in turn contains *parvocellular* (small cell bodies) and *magnocellular* (large 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.

Practical Implications
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 3 we describe how the proposed metric addresses these issues.

3 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. The HDR capability is a result of the light adaptation computation through the JND space transformation and the 3D contrast sensitivity function, both explained in more detail later in this section. 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 2. 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 **Light adaptation**. In our metric we apply the nonlinearity described in [Mantiuk et al. 2005], 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[2].

[2] All externally referred derivations and formulas in the rest of the paper are recollected in supplementary material for easy reference.

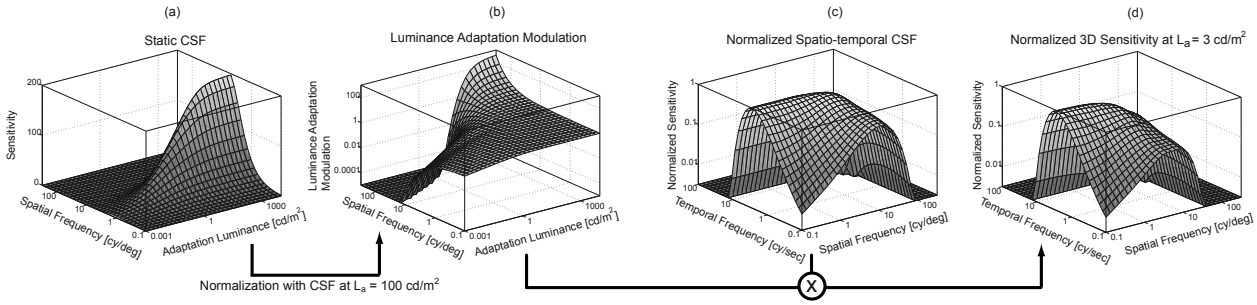


Figure 3: *Computation of the CSF^{3D}. The static $CSF^S(\rho, L_a)$ (a) is divided to $CSF^S(\rho, L_a = 100cd/m^2)$ to obtain scaling coefficients (b) that account for luminance adaptation in CSF^{3D}. The specific adaptation level is chosen to reflect the conditions where the spatiotemporal CSF^T 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 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 3c shows the human contrast sensitivity for variations of ρ and ω at a fixed adaptation luminance. At a retinal velocity v of 0.15 deg/sec, the CSF^T is close to the static CSF^S [Daly 1993] (Figure 3a) at the same adaptation level (the relation between spatio-temporal frequency and retinal velocity is $\omega = v\rho$ assuming the retina is stable). 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:

$$CSF^T(\rho, \omega) = CSF^T(\rho,\ max(v, 0.15) \cdot \rho). \qquad (1)$$

In addition to the drift movement, one could consider integrating a visual attention model-based smooth pursuit eye motion (SPEM) estimate [Yee et al. 2001] (which may not always be precise), or actual eye tracking data to our metric, at the cost of introducing user input and thus loosing objectivity of the approach.

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 3d) by first computing a *Luminance Modulation Factor* (Figure 3b) 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:

$$CSF^{3D}(\rho, \omega, L_a) = \frac{CSF^S(\rho, L_a)}{CSF^S(\rho, 100)} nCSF^T(\rho, \omega). \qquad (2)$$

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 2.3). The effect of luminance adaptation to spatiotemporal contrast sensitivity can approximately be modeled by a multiplier (Figure 3b) 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. For this purpose we extend the **Cortex Transform** [Watson 1987] 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 4-left, are given as Equation 3 and its second derivative, respectively [Winkler 2005]:

$$f(t) = e^{-\frac{\ln(t/0.160)}{0.2}}. \qquad (3)$$

The corresponding frequency domain filters are computed by applying the Fourier transform to both impulse responses and are shown in Figure 4-right.

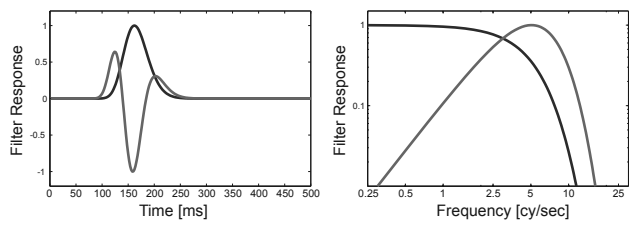


Figure 4: *Impulse (left) and frequency (right) responses of the transient (red) and sustained (blue) temporal channels. The frequency responses comprise the extended 3D Cortex Transform's channels in temporal dimension.*

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** is computed using the following nonlinearity [Daly 1993]:

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

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.

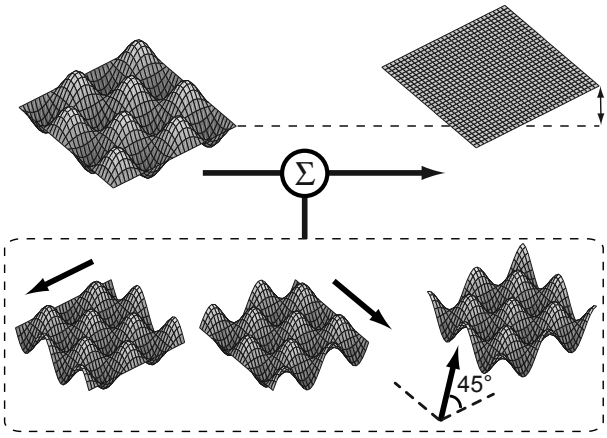


Figure 5: *Practical illustration of achieving phase uncertainty in 2D. The Hilbert transform should be applied in multiple orientations to obtain a phase independent signal.*

The dependency of the visual channels to signal phase contradicts with the observation that the phase sensitivity of the HVS is very limited. Phase uncertainty, while often not explicitly mentioned, is a crucial component of many quality assessment metrics. If one uses a decomposition consisting of spatially even filters, the filter responses would contain zero crossings at step edge locations. This contradicts with human perception which exhibits a strong response to step edges. Analogously, in the temporal dimension sudden changes in pixel intensity are perceived strongly. The effect of phase uncertainty on complex stimuli is often a reduced amount of detected distortions, due to the increased visual masking in step edge locations. 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 5). 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:

$$h^S(\rho_x, \rho_y) = i\ sgn(p\ \rho_x + q\ \rho_y), \quad (5)$$

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:

$$h^T(\omega) = i\ sgn(\omega). \quad (6)$$

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

$$H^{S|T}\{C^{k,l,m}\} = \mathscr{F}^{-1}\{h^{S|T}\ \mathscr{F}\{C^{k,l,m}\}\}. \quad (7)$$

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 6.

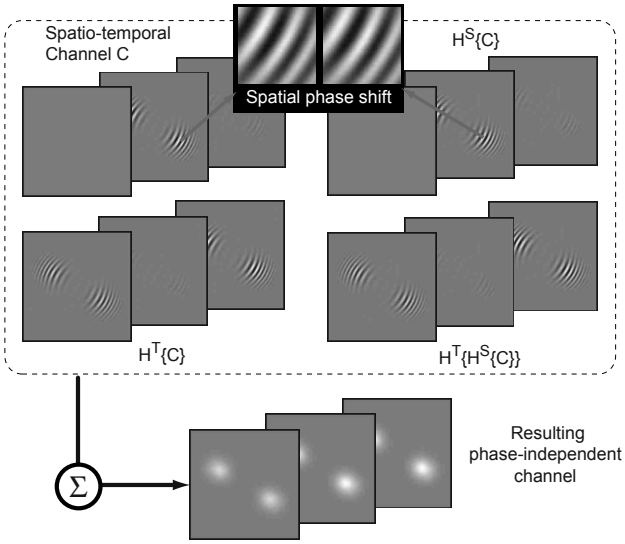


Figure 6: *3D phase uncertainty on a frequency plate image modulated in temporal domain using a sinusoid function. The spatiotemporal channel C obtained by 3D Cortex Transform is used to compute $H^S\{C\}$, $H^T\{C\}$ and $H^T\{H^S\{C\}\}$, the phase shifted response in spatial, temporal and both dimensions, 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**, separately for the reference and test images:

$$P(C) = 1 - \exp(-|C|^3). \quad (8)$$

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:

$$\hat{P}_{\Delta|\searrow|\nearrow} = 1 - \prod_{k=1}^{K}\prod_{l=1}^{L}\prod_{m=1}^{M}\left(1 - P^{k,l,m}_{\Delta|\searrow|\nearrow}\right). \quad (9)$$

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 6.4).

4 Results

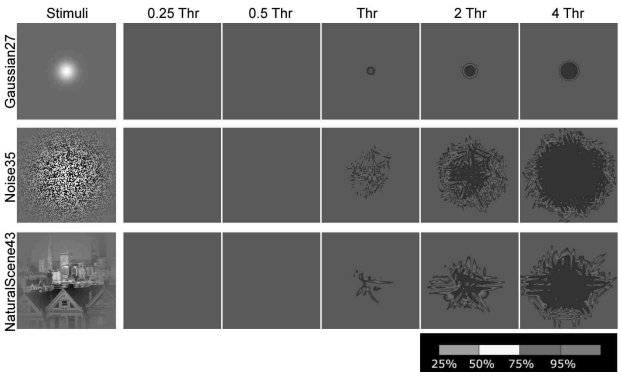


Figure 7: *Predicted visible differences between selected stimuli from the Modelfest data set and the background luminance, where the stimuli is scaled at* $\frac{1}{4}, \frac{1}{2}, 1, 2$ *and* 4 *times the threshold contrast (The same color coding is used throughout the paper for visualizing distortion detection probabilities, unless noted otherwise).*

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. 2008], 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 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[3].

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 the artifacts that appear in rendered videos in the absence of temporal coherency. The magnitude of the noise

[3]Refer to supplementary material for responses to all Modelfest stimuli.

Figure 8: *Approximate perception of the reference and test scenes*

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 8, tone mapped using Pattanaik's operator [2000], depict the approximate appearance of the scene.

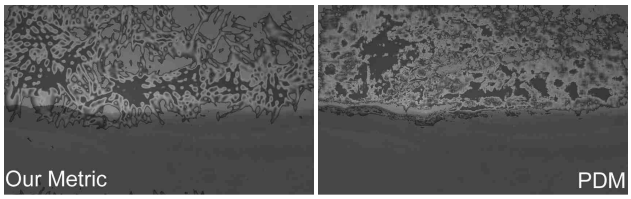


Figure 9: *Metric comparison for LDR test and reference videos*

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[4]. In this case the outcome of our metric and the PDM are similar (Figure 9).

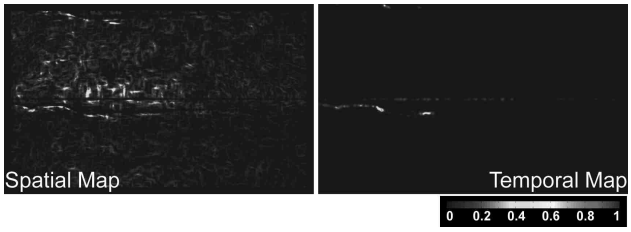


Figure 10: *MOVIE index for LDR videos. Note the different color coding*

The output of the MOVIE index on the other hand are a series of spatial and a temporal distortion maps that are computed at every 16th frame. In Figure 10 we show the spatial distortion map at the 3rd 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 10-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

[4]All original video sequences and corresponding distortion maps are presented in the supplementary video.

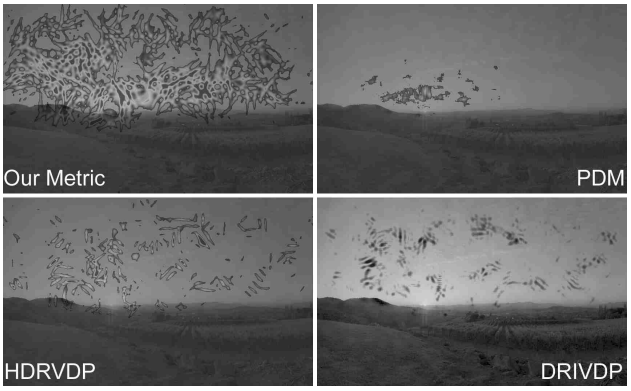


Figure 11: *Metric comparison for HDR test and reference videos. The contrast amplification in DRIVDP is color coded with blue.*

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 11). 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.

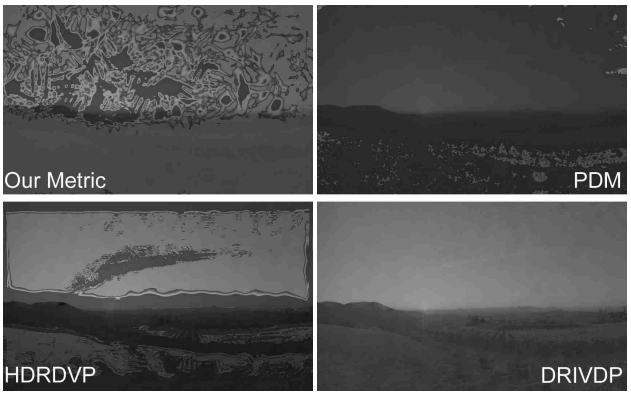


Figure 12: *Metric comparison for HDR reference and LDR test videos*

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 12). Here, both PDM and and HDRVDP's distortion maps are dominated by the contrast difference due to the different dynamic 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.

5 Validation

We performed a subjective study to validate the prediction performance of the metric[5]. The metric's capability of working on video pairs with different dynamic ranges, as well as the outcome in the form of distortion maps containing spatial information, demanded the creation of a new data set, since current public video quality databases are limited to LDR videos, and the measured subjective data is a single number indicating overall quality without any information on spatial distribution of visible distortions. To that end, a test set of 9 reference-test video pairs (1 LDR-LDR, 2 HDR-LDR, and 6 HDR-HDR) were generated by adding temporally and spatially varying artifacts (such as random noise, compression, tone mapping and luminance modulation) to 6 different HDR scenes. A BrightSide DR37-P HDR display was employed to properly display the scene luminance of both HDR and LDR videos. The participants of the study were 16 subjects between ages 23 and 50, all with near perfect or corrected vision. They were shown all video pairs side by side on the HDR display, and were asked to mark the visible differences (detail loss and amplification for HDR-LDR stimuli) on a 16×16 grid displayed over the video using a graphical user interface (Figure 13).

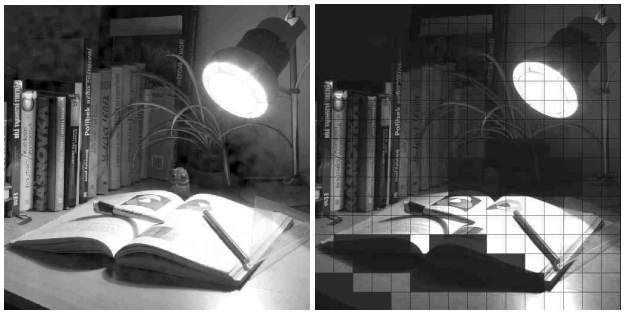


Figure 13: *The graphical user interface displays the test video (left) side-by-side with the corresponding reference video. The subjects mark regions where they notice visible differences on a 16×16 grid (right). Both video frames are tone mapped, and the distortions in the left frame are exaggarated for illustration purposes.*

The marked regions for each trial were stored as distortion maps, 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 all frames, and downsampled to the 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. The correlations varied from 0.733 to 0.883, averaging to 0.809. The high correlation 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 proposed metric provides a reliable estimate of the video quality as a function of spatial location. For comparison, we also evaluated the test set with PDM, HDRVDP and DRIVDP (Figure 14). For almost all stimuli our metric's predictions were more accurate with respect to the subjective data, and the average correlations over all stimuli were found as 0.257 for PDM, 0.528 for HDRVDP, and 0.563 for DRIVDP.

[5]Refer to the supplementary material for a detailed discussion of the experiment.

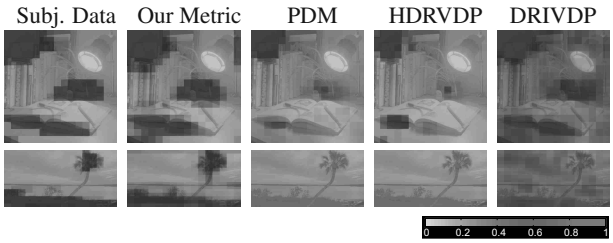


Figure 14: *The comparison of the subjective data averaged over participants with the predictions of our metric, PDM, HDRVDP and DRIVDP for stimulus #2 and #4 in our test set (refer to the supplementary material for the complete set of results).*

6 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.

6.1 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 15 shows that our metric can be used to detect compression artifacts in a video sequence compressed [Mantiuk et al. 2004] at various quality settings.

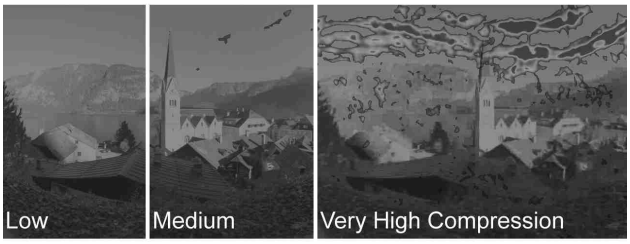


Figure 15: *Visible differences between frames from the HDR video and the corresponding compressed frames shown in three compression settings (Low – q=1, Medium – q=5, Very High – q=31). The banding artifacts become clearly visible under extreme compression. Near the foliage at the bottom, banding artifacts are present but not visible due to the low luminance*

6.2 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 16 we show the results from selected frames of a tone mapped HDR sequence computed with global [Drago et al. 2003] and gradient based [Fattal et al. 2002] tone mapping methods.

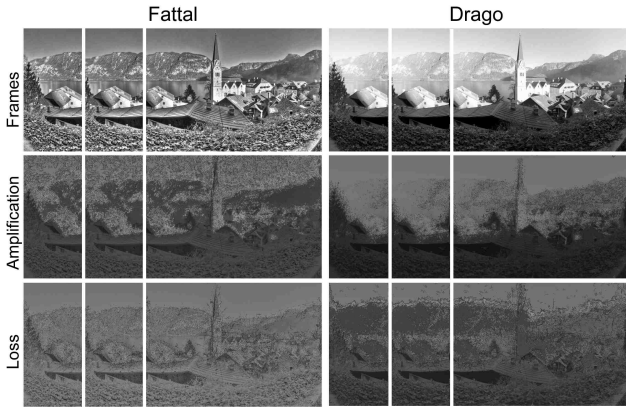


Figure 16: *Selected frames from the tone mapped HDR sequences and corresponding contrast amplification and loss maps. Each frame of the reference HDR video is tone mapped separately. Fattal's gradient based operator enhances perceived contrast notably, thus leading to highly detectable contrast amplification 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 17 we show such an example where tone mapping adversely affects perceived quality of the compressed HDR video, which is correctly detected by our metric.

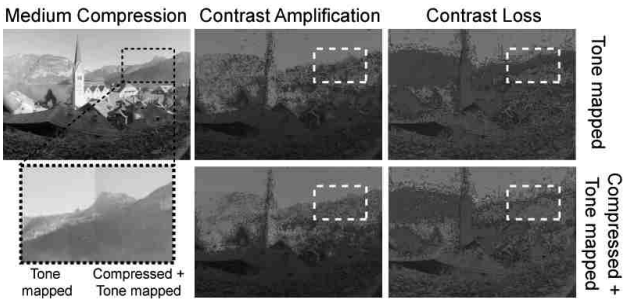


Figure 17: *Contrast amplification and loss predicted with respect to the reference HDR sequence for the compressed (at medium quality) and then tone mapped sequence using Drago's operator. Note the slightly increased contrast amplification and loss in the tone mapped version of the compressed HDR video. As shown in Figure 15, 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.*

6.3 Rendering

Our metric can be used to compare different rendering approaches. Figure 18 shows the visible differences of a dynamic scene walkthrough rendered with indirect lighting using reflective shadow maps [Dachsbacher and Stamminger 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.

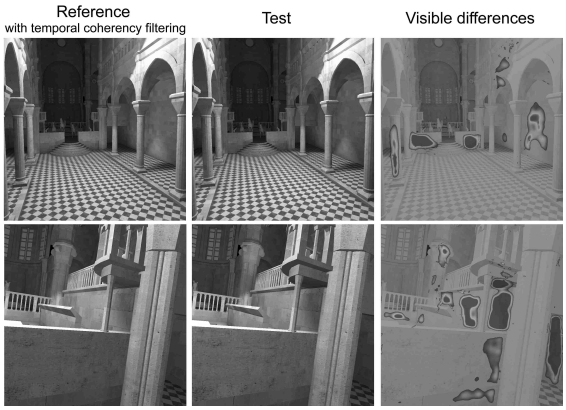


Figure 18: *Visible differences between rendering techniques. Even though the rendered frames are visually indistinguishable when viewed side-by-side, the test method produces significantly visible flickering artifacts, which is not the case for the reference method with temporal coherency filtering. 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 (better visible in the supplementary video).*

To complement the previous scene with mostly temporal distortions, we show another example with artifacts of spatiotemporal nature (Figure 19). 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. 2009] (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.

6.4 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 20 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 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

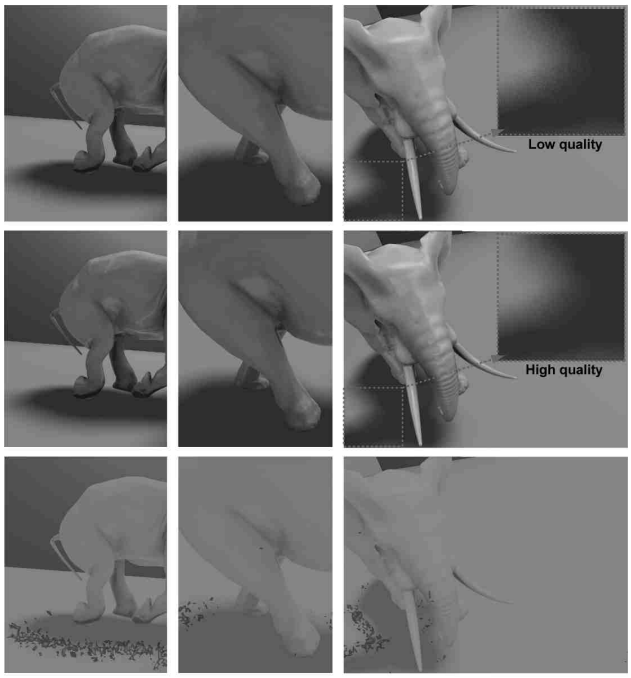


Figure 19: *Visible differences (bottom row) between the high (top row) and low quality (middle row) renderings are focused mostly near shadow boundaries.*

of this scene utilizing volume rendering that gives an overview of the distortion data (Figure 21). Note that the perception of frame freezes and drops has further aspects (e.g. judder) that are not accounted for by our method.

7 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 (~ 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 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 4 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 vi-

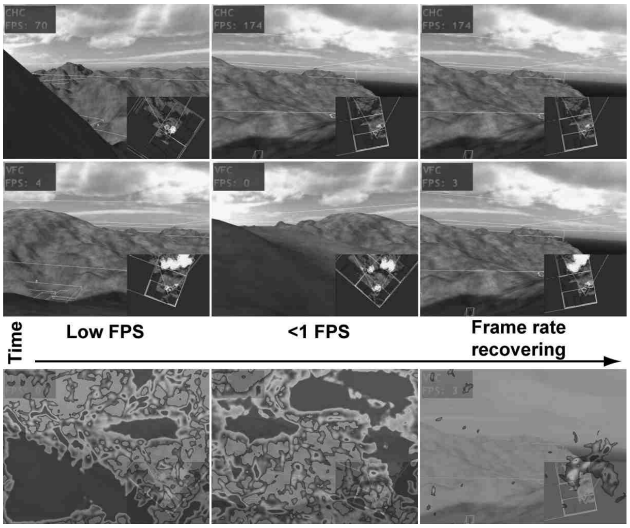


Figure 20: *The effect of dropped frames to perceived quality. One should note, however, that our method does not compensate for camera movements and assumes frames are perfectly aligned with each other.*

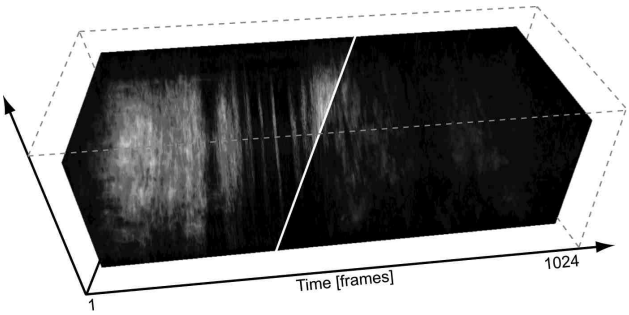


Figure 21: *An alternative 3D visualization. The left slice shows a volume rendering of the entire visible differences data. The right slice shows only the differences with detection probability above* 75% *where the locations of the missing frames along the time axis are better visible.*

sual 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 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.

8 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.

Acknowledgements

Thanks to Robert Herzog for generating the rendered sequences, to Oliver Mattausch for view frustum culling sequences, and to Rafal Mantiuk for providing us with his HDR compression codes and helping us running it. Thanks to Jens Kerber for his help with volumetric visualizations, and to Makoto Okabe for editing and Glenn Lawyer for dubbing the supplemental video. Thanks to all the stuff members and students at MPI Informatik who participated in our experiments. Pisa HDR image and RNL HDR video courtesy of Paul Debevec.

References

AYDIN, T. O., MANTIUK, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2008. Dynamic range independent image quality assessment. In *Proc. of ACM SIGGRAPH*, vol. 27(3). Article 69.

BAVOIL, L., SAINZ, M., AND DIMITROV, R. 2008. Image-space horizon-based ambient occlusion. In *SIGGRAPH '08: ACM SIGGRAPH 2008 talks*, ACM, New York, NY, USA, 1–1.

BITTNER, J., WIMMER, M., PIRINGER, H., AND PURGATHOFER, W. 2004. Coherent hierarchical culling: Hardware occlusion queries made useful. *Computer Graphics Forum 23*, 3 (Sept.), 615–624. Proceedings EUROGRAPHICS 2004.

BOLIN, M., AND MEYER, G. 1998. A perceptually based adaptive sampling algorithm. In *Proc. of Siggraph'98*, 299–310.

DACHSBACHER, C., AND STAMMINGER, M. 2005. Reflective shadow maps. In *I3D '05: Proceedings of the 2005 symposium on Interactive 3D graphics and games*, ACM, New York, NY, USA, 203–231.

DALY, S. 1993. The Visible Differences Predictor: An algorithm for the assessment of image fidelity. In *Digital Images and Human Vision*, MIT Press, A. B. Watson, Ed., 179–206.

DALY, S. J. 1998. Engineering observations from spatiovelocity and spatiotemporal visual models. SPIE, B. E. Rogowitz and T. N. Pappas, Eds., vol. 3299, 180–191.

DRAGO, F., MYSZKOWSKI, K., ANNEN, T., AND N.CHIBA. 2003. Adaptive logarithmic mapping for displaying high contrast scenes. *Computer Graphics Forum 22*, 3.

FATTAL, R., LISCHINSKI, D., AND WERMAN, M. 2002. Gradient domain high dynamic range compression. In *SIGGRAPH*

'02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques, ACM Press, 249–256.

FERWERDA, J., AND PELLACINI, F. 2003. Functional difference predictors (fdps): measuring meaningful image differences. In *Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2, 1388 – 1392 Vol.2.

FREDERICKSEN, R. E., H. R. F. 1998. Estimating multiple temporal mechanisms in human vision. In *Vision Research*, vol. 38, 1023–1040.

FREEMAN, W. T., AND ADELSON, E. H. 1991. The design and use of steerable filters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 13*, 9, 891–906.

HERZOG, R., EISEMANN, E., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2010. Spatio-temporal upsampling on the GPU. In *I3D '10: Proceedings of the 2010 symposium on Interactive 3D graphics and games*, ACM, New York, NY, USA, 91–98.

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

LINDH, P., AND VAN DEN BRANDEN LAMBRECHT, C. 1996. Efficient spatio-temporal decomposition for perceptual processing of video sequences. In *Proceedings of International Conference on Image Processing ICIP'96*, IEEE, vol. 3 of *Proc. of IEEE*, 331–334.

LUBIN, J. 1995. *Vision Models for Target Detection and Recognition*. World Scientific, ch. A Visual Discrimination Model for Imaging System Design and Evaluation, 245–283.

LUKIN, A. 2009. Improved visible differences predictor using a complex cortex transform. *GraphiCon*, 145–150.

MANTIUK, R., KRAWCZYK, G., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2004. Perception-motivated high dynamic range video encoding. *ACM Trans. Graph. 23*, 3, 733–741.

MANTIUK, R., DALY, S., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2005. Predicting visible differences in high dynamic range images - model and its calibration. In *Human Vision and Electronic Imaging X*, vol. 5666 of *SPIE Proceedings Series*, 204–214.

MASRY, M. A., AND HEMAMI, S. S. 2004. A metric for continuous quality evaluation of compressed video with severe distortions. *Signal Processing: Image Communication 19*, 2, 133 – 146.

MYSZKOWSKI, K., ROKITA, P., AND TAWARA, T. 2000. Perception-based fast rendering and antialiasing of walkthrough sequences. *IEEE Transactions on Visualization and Computer Graphics 6*, 4, 360–379.

MYSZKOWSKI, K., TAWARA, T., AKAMINE, H., AND SEIDEL, H.-P. 2001. Perception-guided global illumination solution for animation rendering. In *SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, ACM, New York, NY, USA, 221–230.

PATTANAIK, S. N., TUMBLIN, J. E., YEE, H., AND GREENBERG, D. P. 2000. Time-dependent visual adaptation for fast realistic image display. In *Proc. of ACM SIGGRAPH 2000*, 47–54.

REEVES, W. T., SALESIN, D. H., AND COOK, R. L. 1987. Rendering antialiased shadows with depth maps. In *SIGGRAPH '87: Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, ACM, New York, NY, USA, 283–291.

RITSCHEL, T., GROSCH, T., AND SEIDEL, H.-P. 2009. Approximating dynamic global illumination in image space. In *I3D '09: Proceedings of the 2009 symposium on Interactive 3D graphics and games*, ACM, New York, NY, USA, 75–82.

RUSHMEIER, H., WARD, G., PIATKO, C., SANDERS, P., AND RUST, B. 1995. Comparing real and synthetic images: some ideas about metrics. In *Rendering Techniques '95*, Springer, P. Hanrahan and W. Purgathofer, Eds., 82–91.

SAMPAT, M. P., WANG, Z., GUPTA, S., BOVIK, A. C., AND MARKEY, M. K. 2009. Complex wavelet structural similarity: A new image similarity index. *Image Processing, IEEE Transactions on 18*, 11 (Nov.), 2385–2401.

SCHWARZ, M., AND STAMMINGER, M. 2009. On predicting visual popping in dynamic scenes. In *APGV '09: Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization*, ACM, New York, NY, USA, 93–100.

SESHADRINATHAN, K., AND BOVIK, A. 2007. A structural similarity metric for video based on motion models. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1, I–869–I–872.

SESHADRINATHAN, K., AND BOVIK, A. C. 2010. Motion tuned spatio-temporal quality assessment of natural videos. *Image Processing, IEEE Transactions on 19*, 2 (Feb.), 335– 350.

VAN DEN BRANDEN LAMBRECHT, C., AND VERSCHEURE, O. 1996. Perceptual Quality Measure using a Spatio-Temporal Model of the Human Visual System. In *IS&T/SPIE*.

VAN DEN BRANDEN LAMBRECHT, C., COSTANTINI, D., SICURANZA, G., AND KUNT, M. 1999. Quality assessment of motion rendition in video coding. *Circuits and Systems for Video Technology, IEEE Transactions on 9*, 5 (Aug), 766–782.

WANDELL, B. A. 1995. *Foundations of Vision*. Sinauer Associates, Inc.

WANG, Z., AND SIMONCELLI, E. 2005. Translation insensitive image similarity in complex wavelet domain. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 2, 573–576.

WATSON, A. B., AND MALO, J. 2002. Video quality measures based on the standard spatial observer. In *ICIP (3)*, 41–44.

WATSON, A. B., HU, J., AND III, J. F. M. 2001. DVQ: A digital video quality metric based on human vision. *Journal of Electronic Imaging 10*, 20–29.

WATSON, A. B. 1986. Temporal sensitivity. In *Handbook of Perception and Human Performance*, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds. John Wiley and Sons, New York, 6–1–6–43.

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

WINKLER, S. 1999. A perceptual distortion metric for digital color video. In *Proceedings of the SPIE Conference on Human Vision and Electronic Imaging*, IEEE, vol. 3644 of *Controlling Chaos and Bifurcations in Engineering Systems*, 175–184.

WINKLER, S. 2005. *Digital Video Quality: Vision Models and Metrics*. Wiley.

YEE, H., PATTANAIK, S., AND GREENBERG, D. P. 2001. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Trans. Graph. 20*, 1, 39–65.

Appendix G

On Evaluation of Video Quality Metrics: an HDR Dataset for Computer Graphics Applications

M. Čadík, T. O. Aydın, K. Myszkowski, and H.-P. Seidel. On evaluation of video quality metrics: an HDR dataset for computer graphics applications. *Human Vision and Electronic Imaging XVI*, Vol. 7865, No. 1, 2011.

On Evaluation of Video Quality Metrics: an HDR Dataset for Computer Graphics Applications

Martin Čadík, Tunç O. Aydın, Karol Myszkowski, Hans-Peter Seidel

MPI Informatik

ABSTRACT

In this paper we propose a new dataset for evaluation of image/video quality metrics with emphasis on applications in computer graphics. The proposed dataset includes LDR-LDR, HDR-HDR, and HDR-LDR reference-test video pairs with various types of distortions. We also present an example evaluation of recent image and video quality metrics that were applied in the field of computer graphics. In this evaluation all video sequences were shown on an HDR display, and subjects were asked to mark the regions where they saw differences between test and reference videos. As a result, we capture not only the magnitude of distortions, but also their spatial distribution. This has two advantages: on one hand the local quality information is valuable for computer graphics applications, on the other hand the subjectively obtained distortion maps are easily comparable to the maps predicted by quality metrics.

1. INTRODUCTION

Experimental evaluation of computer graphics (CG) techniques is necessary to validate their impact on perceived quality of resulting images. So far such image quality evaluation in CG is mostly performed informally without referring to well-established subjective and objective methods, which are commonly used in other fields, such as in image compression. In particular, CG field could benefit greatly from objective quality metrics due to the simplicity of their use and low costs involved. This however, requires extensive perceptual validation of such image and video quality metrics, which should be sensitive to image artifacts and distortions specific in CG. Another important aspect of such validation is high dynamic range (HDR) of images that are often generated by HDR rendering pipelines, which are today common in computer games (utilizing GPU or specialized consoles) and computer-aided design systems (in particular dealing with realistic image synthesis).

To make the validation (or calibration) of image or video quality metric possible, one needs to design a set of input stimuli (i.e. a *dataset*) and perform a user study which results in a set of subjective (mean) opinion scores. Subjective studies are very laborious and may be stimuli dependent, thus the community benefits from publicly available, standardized data sets. Therefore, a few datasets were published in the past.[1–5]

Unfortunately, none of the existing datasets is suitable for evaluation of video quality metrics in computer graphics field, where the images and videos often exhibit high dynamic range of luminance values and specific artifacts (see Figures 1, 2). Existing datasets are limited in dynamic range of the input stimuli (only low-dynamic (LDR) range videos), in the distortions they cover (mostly compression-related artifacts), and in the extent of subjective responses (usually the numerical rating of the quality of the stimulus). Few authors employed a concept of *image distortion maps*[6–8] in evaluation of image quality metrics, but this has not been done for temporal distortions in videos so far. To overcome the above limitations, we propose and make publicly available* a new dataset for evaluation of image/video quality metrics with emphasis on applications in computer graphics. Several aspects were influential while designing the dataset: (i) in addition to the assessmet of the quality of LDR videos, the assessment of high-dynamic range videos, as well as comparing HDR videos with LDR videos and vice versa, and (ii) the outcome of the subjective experiment 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. Furthermore, we show an example evaluation of recent image and video quality metrics that were

Further author information: e-mail: {mcadik, tunc, karol, hpseidel}@mpi-inf.mpg.de
Max Planck Institut Informatik, Saarbrücken

*http://www.mpi-inf.mpg.de/resources/hdr/quality

applied in the field of computer graphics. The goal of this evaluation was to examine the correlation between the objective quality predictions computed by the video quality metrics, and the subjective responses obtained by the experimental procedure. It is known[9] that applications of image/video quality metrics into the field of computer graphics are still far from maturity, we believe however, that the published dataset helps in validation and improvement of existing, and the design of future metrics for computer graphics and other applications.

To that end the proposed dataset and the subjective study have the following unique features 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 were not asked to assess only an overall quality of the video, but to mark the regions where they saw differences between test and reference videos, resulting in distortion maps similar to the metric outcome.

In the remainder of this paper we describe the proposed dataset for evaluation of video quality metrics, the experimental setup and procedure (Section 2), present an example evaluation of video quality metrics using the dataset (Section 3) and discuss the results based on the correlation between the outcome of the subjective data and corresponding predictions of state-of-the art video quality metrics.

Figure 1. An example of typical artifacts in rendered images and video sequences: an indoor scene rendered using progressive photon mapping algorithm.[10] Left: non-converged solution (2 iterations) exhibits *low-frequency noise*. Right: fully converged solution.

2. DATASET FOR VIDEO METRIC EVALUATION

The proposed dataset consists of 9 reference-test video pairs (1 LDR-LDR, 2 HDR-LDR, and 6 HDR-HDR), they are listed in Table 1. The video stimuli were generated by imposing temporally varying visual artifacts to HDR scenes (Figure 3), 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. In sequences #1-#4, and #9 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 (referred as "high stddev") and 5 (referred as "low stddev") 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 (referred as "low magnitude"), and clearly visible in the other (referred as "high magnitude"). In sequences #5 and #6, the HDR compression[13]

Figure 2. An example of artifacts due to the tone mapping of HDR images and HDR video sequences. Left: global tone mapping technique by Pattanaik et al.[11] preserves overall image contrast, but results in severe *loss of details*. Right: gradient-based technique of Fattal et al.[12] is able to reproduce virtually all the image details, at the cost of an overall contrast and *contrast reversals* (i.e. halo artifacts).

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[11,12] to each input HDR video frame were used in the dynamic range independent comparisons (sequences #7 and #8).

All test videos consist of 60 frames, and should be 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 17 samples taken from the displayable luminance range. The sample points were then fitted to a 3rd 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.

#	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 - low magnitude, low stddev

Table 1. List of the experimental stimuli. Refer to text for details.

The participants of the experimental study were 16 subjects of age 23 to 50. 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 4). 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

Cars Lamp Desk Tree Cafe Tower

Figure 3. The video test set is generated from 6 calibrated HDR scenes (tone mapped for presentation purpose[14]). 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.

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.

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, see Figure 6 (first column). The descriptive statistics of these maps are summarized in Table 2 (first column). Figure 5 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.

Figure 4. The experiment was performed through a graphical user interface 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.

3. EXAMPLE EVALUATION

To illustrate the utilization of the proposed dataset, we show an evaluation of four state-of-the-art image/video quality metrics: DRIVQM,[15] PDM,[16] HDRVDP,[8] and DRIVDP.[17] For each of the evaluated metrics the predictions for each stimulus were calculated, averaged over the whole 60 frames, and downsampled to the same

Stimulus #	Subjective Response	DRIVQM	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 2. Descriptive statistics of distortion maps (depicted in Figure 6) for each input stimulus. Abbreaviations used: min=minimal value, max=maximal value, avg=average value, std=standard deviation, of the distortion map averaged over all subjects/metric responses for a particular stimulus (1-9).

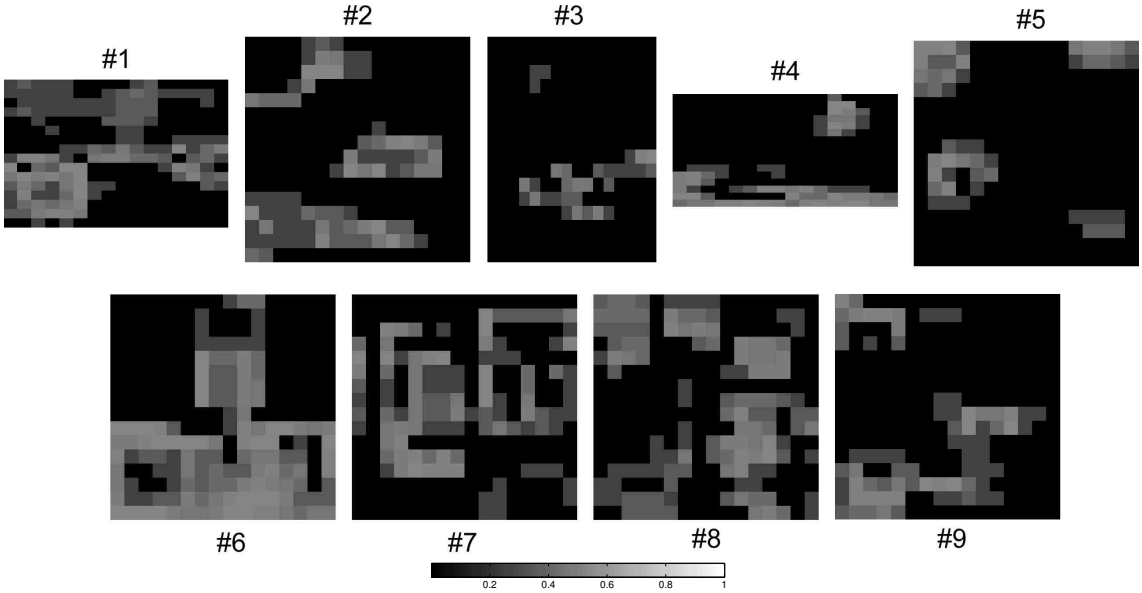


Figure 5. Maps showing the standard deviations over subjects for each stimulus. The numbers refer to the first column of Table 1.

resolution as the mean subjective response. In all the tests in this paper we used a frequency domain implementation of DRIVQM metric, precisely following the reference publication. We found that using this implementation was still feasible for the frame sizes and durations of the video sequences in our dataset. For larger sequences we also implemented an alternative version of the metric, where the frequency domain channel decomposition is replaced by a Steerable Pyramid based spatial decomposition (described in Appendix A). A web service based on this implementation is publicly available at `http://drivqm.mpi-inf.mpg.de`. The HDRVDP and DRIVDP metrics are designed for image quality evaluation, thus the video stimuli were evaluated for each frame separately. For each video pair, we computed the 2D correlation between the mean subjective response and the metric prediction (Table 3) and used the results to evaluate the performance of the metrics. The Figure 6 shows the mean subjective distortion maps along with the corresponding metrics predictions for visual inspection. The descriptive statistics of these maps are summarized in Table 2.

For the purposes of generating the maps in Figure 6, 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:

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

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.

3.1 Discussion

As DRIVQM is the only evaluated metric that was designed specifically for the purposes of dynamic range-independent video quality assessment of computer graphics sequences, it is not surprising that it overcomes the other metrics in most cases. 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. The worst prediction of DRIVQM is (#8, 0.733).

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 3, DRIVQM significantly outperforms others in most cases. The significant difference in average correlations over the entire test set (last row of Table 3) shows that overall DRIVQMs predictions are clearly more accurate than others. The corresponding distortion maps predicted by PDM, HDRVDP and DRIVDP are shown in Figure 6 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 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.

4. CONCLUSION

The main goal of this work was to develop a dataset of video sequences accompanied by the corresponding subjective data evaluating their quality in a local manner. Such locality is the key in computer graphics applications, where local image distortions should be detected and if possible corrected in rendering. Another important aspect of such dataset is dynamic range of frames, where pairs of HDR, LDR and mixed LDR-HDR reference and test videos, which are calibrated in terms of pixel luminance, are considered. We propose also a novel subjective testing setup that involves an HDR display, which is suitable for reproducing luminance levels in the videos, as well as interactive marking of local image regions where distortions are visible.

The dataset proved to be useful in calibrating and validating the DRIVQM,[15] which has been developed specifically for computer graphics applications. The dataset is publicly available (`http://www.mpi-inf.mpg.de/resources/hdr/quality`) and our hope is that it can be used in other validation tasks.

Stimulus #	DRIVQM	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 3. Correlations of subjective responses with predictions of DRIVQM, 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.

APPENDIX A. EFFICIENT IMPLEMENATATION OF DRIVQM

As we admit in the discussion section of the previous publication,[15] executing DRIVQM becomes infeasible for long video sequences due to the long processing time. Moreover, the 64-frames window size practically means that for frame sizes larger than VGA the memory consumption is prohibitively large for an average desktop computer. As suggested by the authors of the metric, we replace the frequency domain Cortex Transform by the Steerable Pyramid[18] with 6 levels (where each differ by one octave) and 6 orientations, and use Winkler's[19] 9-tap approximation of the transient and sustained temporal mechanisms. The base band of the Steerable Pyramid, analogous to Cortex Transform, does not have any orientations. For the purpose of accounting for spatial phase uncertainty, the Hilbert transforms of the steerable filters provided by Freeman and Adelson's formulation are used, whereas temporal phase uncertainty is ignored. All other components of the DRIVQM are left intact, including the extended Contrast Sensitivity Function. The updated implementation is more efficient in memory consumption and running time, without any significant deviation in results from the original implementation. A publicly available web service that uses our updated implementation can be found at `http://drivqm.mpi-inf.mpg.de`. The web service requires the users to upload the frames of their test and reference video pair of either the same dynamic range, or different dynamic ranges. Users are allowed to change metric parameters, such as the pixels per visual degree, frames per second and adaptation luminance. Once the setup is complete, the metric is run on the uploaded video pair. Upon completion, the input video pair is immediately deleted, and the user is provided with a link to the contrast difference, contrast loss, and contrast amplification maps.

ACKNOWLEDGMENTS

The authors would like to thank all the staff members and students at MPI Informatik who kindly participated in the subjective experiments.

REFERENCES

[1] Carney, T., Klein, S. A., Tyler, C. W., Silverstein, A. D., Beutter, B., Levi, D., Watson, A. B., Reeves, A. J., Norcia, A. M., Chen, C.-C., Makous, W., and Eckstein, M. P., "The development of an image/threshold database for designing and testing human vision models," in [*Proc. of Human Vision, Visual Processing, and Digital Display IX*], SPIE, Bellingham, WA, 3644 (1999).

[2] The Video Quality Experts Group, "Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment, Phase I." `http://www.its.bldrdoc.gov/vqeg/projects/frtv` (2000).

[3] Sheikh, H. R. and Bovik, A. C., "LIVE image quality assessment database." `http://live.ece.utexas.edu/research/quality/subjective.htm` (2003).

[4] Seshadrinathan, K., Soundararajan, R., Bovik, A. C., and Cormack, L. K., "LIVE video quality database." `http://live.ece.utexas.edu/research/quality/live_video.html` (November 2009).

[5] Seshadrinathan, K., Soundararajan, R., Bovik, A. C., and Cormack, L. K., "A subjective study to evaluate video quality assessment algorithms," *Human Vision and Electronic Imaging XV* **7527**(1), 75270H, SPIE (2010).

[6] Zhang, X., Setiawan, E., and Wandell, B., "Image distortion maps," in [*Fifth Color Imaging Conference: Color Science, Systems and Applications*], 120–125 (1997).

[7] Zhang, X. and Wandell, B. A., "Color image fidelity metrics evaluated using image distortion maps," *Signal Processing* **70**(3), 201 – 214 (1998).

[8] Mantiuk, R., Daly, S., Myszkowski, K., and Seidel, H.-P., "Predicting visible differences in high dynamic range images - model and its calibration," in [*Human Vision and Electronic Imaging X*], *SPIE Proceedings Series* **5666**, 204–214 (2005).

[9] Rogowitz, B. E. and Rushmeier, H. E., "Are image quality metrics adequate to evaluate the quality of geometric objects?," in [*Proc. of Human Vision and Electronic Imaging VI*], 340–348, SPIE (2001).

[10] Hachisuka, T., Ogaki, S., and Jensen, H. W., "Progressive photon mapping," in [*ACM Transactions on Graphics (Proc. of SIGGRAPH Asia'08)*], **27**(5), 1–8, ACM, New York, NY, USA (2008).

[11] Pattanaik, S. N., Tumblin, J. E., Yee, H., and Greenberg, D. P., "Time-dependent visual adaptation for fast realistic image display," in [*ACM Transactions on Graphics (Proc. of ACM SIGGRAPH'00)*], 47–54, ACM Press (2000).

[12] Fattal, R., Lischinski, D., and Werman, M., "Gradient domain high dynamic range compression," in [*ACM Transactions on Graphics (Proc. of ACM SIGGRAPH'02)*], 249–256, ACM Press (2002).

[13] Mantiuk, R., Krawczyk, G., Myszkowski, K., and Seidel, H.-P., "Perception-motivated high dynamic range video encoding," in [*ACM Transactions on Graphics (Proc. of SIGGRAPH'04)*], **23**(3), 733–741, ACM (2004).

[14] Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J., "Photographic tone reproduction for digital images," in [*ACM Transactions on Graphics (Proc. of SIGGRAPH'02)*], 267–276, ACM Press (2002).

[15] Aydın, T. O., Čadík, M., Myszkowski, K., and Seidel, H.-P., "Video quality assessment for computer graphics applications," in [*ACM Transactions on Graphics (Proc. of SIGGRAPH Asia'10)*], 1–10, ACM, Seoul, Korea (2010).

[16] Winkler, S., [*Digital Video Quality: Vision Models and Metrics*], Wiley (2005).

[17] Aydın, T. O., Mantiuk, R., Myszkowski, K., and Seidel, H.-P., "Dynamic range independent image quality assessment," in [*ACM Transactions on Graphics (Proc. of ACM SIGGRAPH'08)*], **27(3)** (2008). Article 69.

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

[19] Winkler, S., "A perceptual distortion metric for digital color video," in [*Proc. of Human Vision and Electronic Imaging, SPIE*], *Controlling Chaos and Bifurcations in Engineering Systems* **3644**, 175–184, IEEE (1999).

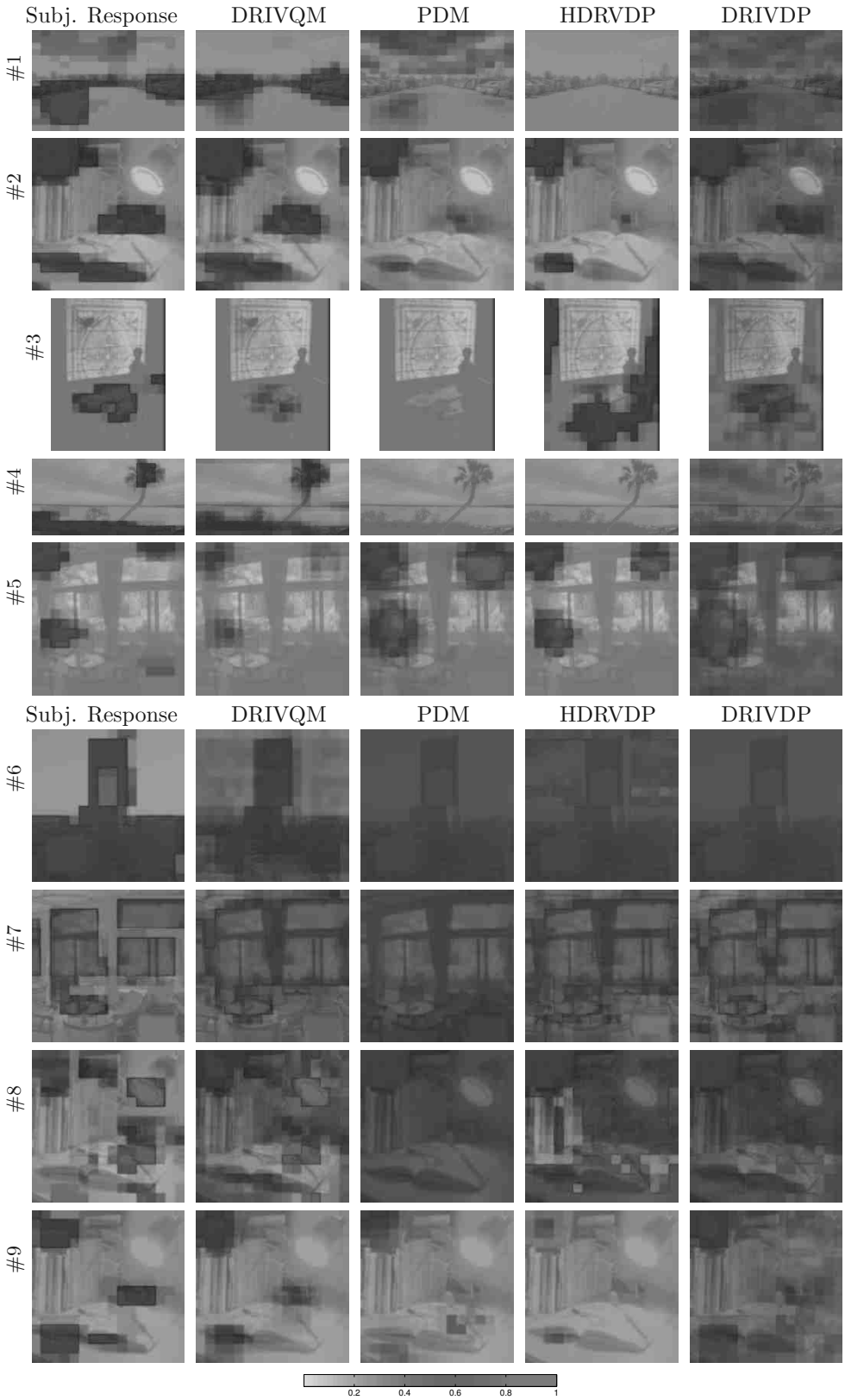


Figure 6. Mean subjective response distortion maps and corresponding mean metric predictions pairs.

Appendix H

New Measurements Reveal Weaknesses of Image Quality Metrics in Evaluating Graphics Artifacts

M. Čadík, R. Herzog, R. Mantiuk, K. Myszkowski, and H.-P. Seidel. New Measurements Reveal Weaknesses of Image Quality Metrics in Evaluating Graphics Artifacts. *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia)*, Vol. 31, No. 6, pp. 147:1–147:10, 2012.
IF=3.725

New Measurements Reveal Weaknesses of Image Quality Metrics in Evaluating Graphics Artifacts

Martin Čadík* Robert Herzog* Rafał Mantiuk◇ Karol Myszkowski* Hans-Peter Seidel*
*MPI Informatik Saarbrücken, Germany ◇Bangor University, United Kingdom

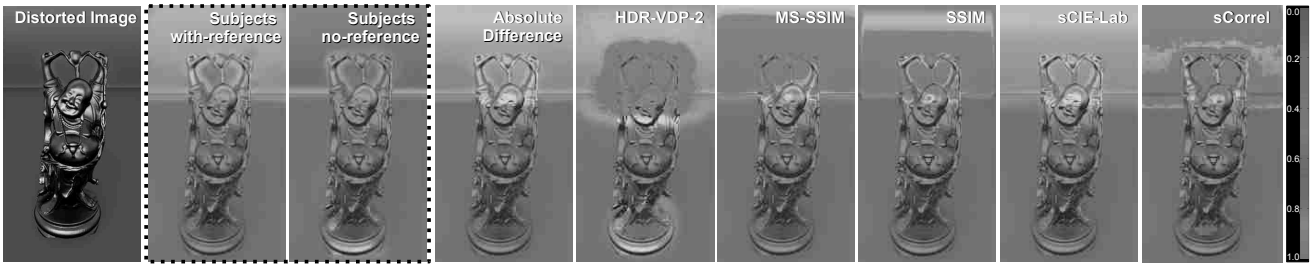


Figure 1: *State-of-the-art image quality metrics often fail in the prediction of the human-perceived distortions in complex images. Here, we show the predicted detection probabilities (color-coded) for gradient-based tone mapping artifacts [Fattal et al. 2002] in a synthetic image.*

Abstract

Reliable detection of global illumination and rendering artifacts in the form of localized distortion maps is important for many graphics applications. Although many quality metrics have been developed for this task, they are often tuned for compression/transmission artifacts and have not been evaluated in the context of synthetic CG-images. In this work, we run two experiments where observers use a brush-painting interface to directly mark image regions with noticeable/objectionable distortions in the presence/absence of a high-quality reference image, respectively. The collected data shows a relatively high correlation between the with-reference and no-reference observer markings. Also, our demanding per-pixel image-quality datasets reveal weaknesses of both simple (PSNR, MSE, sCIE-Lab) and advanced (SSIM, MS-SSIM, HDR-VDP-2) quality metrics. The most problematic are excessive sensitivity to brightness and contrast changes, the calibration for near visibility-threshold distortions, lack of discrimination between plausible/implausible illumination, and poor spatial localization of distortions for multi-scale metrics. We believe that our datasets have further potential in improving existing quality metrics, but also in analyzing the saliency of rendering distortions, and investigating visual equivalence given our with- and no-reference data.

CR Categories: I.3.0 [Computer Graphics]: General;

Keywords: Image quality metrics (IQM), perceptual experiments, global illumination, noticeable and objectionable distortions

Links: DL PDF WEB DATA

*e-mail: mcadik@mpi-inf.mpg.de, the complete dataset is available at: `http://www.mpii.de/resources/hdr/iqm-evaluation/`

1 Introduction

Rendering techniques, in particular global illumination, are prone to image artifacts, which might arise due to specific scene configurations, imbalanced scene complexity that might lead to a locally varying convergence-rate of the solution, and numerous simplifications in the rendering algorithms themselves. With the proliferation of 3D rendering services, where the user may often arbitrarily interact with the content, the role of automatic rendering-quality control gains in importance. Even in well-established industries such as gaming a massive approach to automatic quality testing is desirable. In practice, objective image quality metrics (IQM) that are successful in lossy image compression and transmission applications [Wang and Bovik 2006] are predominantly used in graphics, including advanced attempts of their adaptation to actively steer rendering [Rushmeier et al. 1995; Bolin and Meyer 1998; Ramasubramanian et al. 1999]. Such objective IQM are trained to predict a single value of mean opinion score (MOS) for image blockiness, noise, blur, or ringing distortions. However, their performance for other distortion types as well as their spatial localization within an image has not been systematically validated so far.

The goal of this work is to generate a new rendering-oriented dataset with localized distortion maps and use it for the evaluation of existing IQM. For this purpose we prepare a set of images with distortions that are typical for popular global illumination and rendering techniques as well as the corresponding distortion-free reference images. Table 1 presents a summary of our stimuli. In two separate experiments (Sec. 3) we ask the observers to locally mark *noticeable* and *objectionable* distortions where the reference image is either shown or hidden, respectively. We demonstrate that the observers can reliably perform both tasks, yielding high coefficients of agreement (Sec. 4.1). In general, our results show a high correlation between the observer marking for the *with-reference* and *no-reference* datasets, but we also indicate the most common sources of discrepancies in such marking (Sec. 4.2).

We use the with-reference dataset to evaluate the performance of state-of-the-art *full-reference* (FR) IQM in detecting and localizing rendering distortions (Sec. 5). We show that even advanced IQM fail for some common computer graphics artifacts (e.g., Fig. 1). Our data shows that in general no IQM performs better than any other, even including the simple absolute difference (AD), which is equivalent to the peak signal-to-noise ratio (PSNR) or mean-

square-error (MSE) given our non-parametric metric performance measures. Moreover, our analysis reveals some interesting weaknesses of FR IQM, including the lack of robustness to brightness and contrast change, the inability to distinguish between plausible and implausible illumination patterns, and poor localization of distortions due to multi-scale processing.

2 Related work

In this section we briefly characterize general purpose full reference (FR) IQM which are central for our comparison against the subjective data. Also, we review major other developments in the evaluation of IQM performance. For more in depth discussion of the image quality problem we refer the reader to the recent textbooks [Wang and Bovik 2006; Wu and Rao 2005], and survey papers [Lin and Kuo 2011; Pedersen and Hardeberg 2011].

2.1 Image quality metrics (IQM)

Full reference IQM can be categorized into different groups based on the principles behind their construction [Wang and Bovik 2006; Pedersen and Hardeberg 2011].

Mathematically-based metrics directly measure the difference of pixel intensity. The root mean square error (RMSE) and peak signal-to-noise-ratio (PSNR) are the most prominent examples of metrics belonging to this category.

HVS-based metrics model early human vision characteristics such as luminance adaptation, contrast sensitivity, visual masking, and visual channels. The most prominent examples of such metrics include the Visible Differences Predictor (VDP) [Daly 1993] and Visual Discrimination Model (VDM) [Lubin 1995]. VDP has also been used in the evaluation of rendered image quality [Rushmeier et al. 1995]. Recently, extensions of VDP have been proposed to handle high dynamic range (HDR) images [Mantiuk et al. 2005; Mantiuk et al. 2011].

Structure-based metrics detect structural changes in the image by means of a spatially localized measure of correlation in pixel values. The Structural Similarity Index Metric (SSIM) is based on this principle. In addition, it is sensitive to the differences in the mean intensity and contrast [Wang and Bovik 2006, Ch. 3.2].

Other metrics combine the strengths of different metric categories. For example, in sCIE-Lab [Zhang and Wandell 1998] spatial color sensitivity is added to a standard color-difference measure in the perceptually-uniform CIE-Lab color-space. In the Visual Signal-to-Noise Ratio (VSNR) metric [Chandler and Hemami 2007] at first an HVS-model is applied to eliminate distortions below the visibility threshold and then a simple mathematically-based metric is used. Other modern metrics, such as the Visual Information Fidelity (VIF) index [Wang and Bovik 2006, Ch. 3.3], rely on natural-scene statistics and employ an information-theoretic approach to measure the amount of information that is shared between two images.

2.2 Evaluation of image quality metrics

The comparison of IQM performance against data collected in experiments with human subjects is required to evaluate metric prediction accuracy and robustness for different types of visual distortions. Standardized procedures for subjective image- and video-quality evaluation have been developed by the International Telecommunication Union [ITU-T-P.910 2008; ITU-R-BT.500-11 2002]. They rely on subjectively collected *mean opinion score (MOS)* data, which is compared against a single number derived from the error pooling over pixels. While such a procedure works well for estimating the overall *magnitude of distortions*, information on different distortion types, their possible interactions and spatial distribution is not captured. In computer graphics applications the prediction of *local distortion detectability* by a human observer is essential, and in this work we favor *image distortion maps*, which capture such spatial information.

Mean opinion score (MOS) data. A number of databases of images with different distortion types and MOS subjective quality scores is publicly available where LIVE [Sheikh et al. 2006] and Tampere Image Database [Ponomarenko et al. 2009] are the most prominent examples featuring both significant variety of distortions and large number of stimuli, which have been judged by many subjects (30–200). Lin and Kuo [2011] present a more complete summary of such databases with detailed characterization of supported distortion types, which arise mostly in image compression and transmission. Distortions covered by those databases that are more relevant for graphics applications include blur, mean intensity shifts, contrast changes, and various types of noise.

Image distortion maps. The spatial aspect of distortion detectability has been addressed in calibration and performance evaluation for HDR image [Mantiuk et al. 2005] and video [Čadík et al. 2011] quality metrics. Thereby, the screen is divided into discrete blocks of about 30×30 pixels and the subjects mark blocks with noticeable distortions. Similar to our work, Zhang and Wandell [1998] used a brush-painting interface for freely marking reproduction artifacts due to half-toning or JPEG compression given the reference image. The marked errors produced by 24 subjects have been pooled for each distorted image and as a result image distortion maps with the probability of error detection have been obtained. In our experiments we enable pixel-precise distortion marking, which we then average in downsampled images that are used in our analysis. This improves the quality of the data compared to the heuristic-driven pixel rejection used in [Zhang and Wandell 1998]. Unlike that study, we focus exclusively on rendering-related artifacts, and we consider both the with- and no-reference experiment scenarios.

In this work we extend our dataset [Herzog et al. 2012], which consists of 10 stimuli exhibiting mostly supra-threshold distortions for 3 selected distortion types, with 27 new stimuli (refer to Table 1 and the supplementary material for a more detailed summary of both datasets). The new stimuli exhibit sub-threshold, near-threshold, and supra-threshold distortions, which are often present in a single image. In comparison to that previous work, the new dataset reduces the subject learning effect by mixing different types of distortions within a single image, restricting their appearance to randomly selected parts of an image and increasing the number of distortion types to 12. This also let us test the metrics in more challenging scenarios, where the distortions are non-uniformly distributed across an image. Moreover, while the previous dataset contained mostly well visible distortions, the new images contain also low amplitude distortions, which are near the visibility threshold. The new dataset reinforces the quality and robustness of the subjective data, which is achieved by stabilizing the distance to the screen using a chin-rest and involving a large number of observers (35).

3 Localized image distortion experiment

The goal of the study is to mark areas in the images that contain *noticeable* distortions and those that contain *objectionable* distortions. The former will let us test how well the IQM predict visibility, while the latter can tell how robust the metrics are to image modifications that are not perceived as distortions. In addition, the analysis of the experimental data alone, for both visible and detectable thresholds, can reveal which image differences are seen as disturbing and which are most likely ignored or interpreted as a part of the original

Scene	Distortion Type	Mask	Method (Ref.)	Tonem.	Settings Artifact (Ref.)
#1 Apartment	VPL Clamp.	no	IGI (LC)	[Rein.]	$0.1 \cdot 10^6$ vpls ($2 \cdot 10^6$ vpls)
#2 CG Figures	Struc-Noise	no	IGI (PT)	[Drag.]	10^6 vpls (97K spp)
#3 Disney	Struc-Noise	no	IGI (PT)	[Drag.]	10^6 vpls (43K spp)
#4 Kitchen	Struc-Noise	no	IGI (PT)	[Drag.]	10^6 vpls (380K spp)
#5 Red Kitchen	Struc-Noise	no	IGI (PT)	[Drag.]	10^6 vpls (200K spp)
#6 Sponza Above T.	Alias. (Shadow)	no	GL (GL)	[Rein.]	Shadowmap-pcf 1024^2 (4096^2)
#7 Sponza Arches	Alias. (Shadow)	no	GL (GL)	[Rein.]	Shadowmap-pcf 1024^2 (4096^2)
#8 Sponza Atrium	Alias. (Shadow)	no	GL (GL)	[Rein.]	Shadowmap 1024^2 (4096^2)
#9 Sponza Tree Shad.	Alias. (Shadow)	no	GL (GL)	[Rein.]	Shadowmap 1024^2 (4096^2)
#10 Sponza Trees	VPL Clamp.	no	IGI (LC)	[Rein.]	$60 \cdot 10^3$ vpls ($2 \cdot 10^6$ vpls)
#11 Apartment II	Struc-Noise	yes	RC,RC (PPM)	[Rein.]	RC+PM phot.: $0.5 \cdot 10^6$ ($3 \cdot 10^9$)
#12 Atrium	Struc-Noise	yes	PPM (PPM)	[Rein.]	$5 \cdot 10^9$ ($20 \cdot 10^9$) photons
#13 Bathroom	Noise	yes	PPM,PPM (PPM)	[Rein.]	custom renderer
#14 Buddha	Tonemap (Halo/Bright.)	yes	[Fat.'02] (γ=3.0)	–	PBRT / pfstools
#15 Chairs	High/Med-freq Noise	yes	MCRT (MCRT)	γ=2.2	backward RT [ITBT/Inspirer]
#16 City-d	Alias. (Downsampl.)	yes	NN (–)	γ=2.2	PBRT / Matlab
#17 City-u	Upsampl. (Lanczos)	yes	NN,Lanczos (–)	γ=2.2	PBRT / Matlab
#18 Cornell	Alias./Struc-Noise	yes	RC (RC)	γ=1.8	PBRT 1 spp (128 spp)
#19 Dragons	Noise	no	RC (RC)	γ=2.2	PBRT 16 spp (128 spp)
#20 Hall	Brightness	no	MCRT (MCRT)	γ=2.2	backward RT [ITBT/Inspirer]
#21 Icido	Struc-Noise	yes	RC (PPM)	[Rein.]	RC+LC vpls: $0.5 \cdot 10^6$ ($3 \cdot 10^9$)
#22 Kitchen II	Struc-Noise/Bright.	yes	RC (PPM)	[Drag.]	RC+PM phot.: $0.5 \cdot 10^6$ ($3 \cdot 10^9$)
#23 Livingroom	Noise	yes	PPM,PPM (PPM)	[Rein.]	custom renderer
#24 MPII	Tonemap. (Grad.)	yes	[Man.'06] (γ=4.5)	–	PBRT / pfstools
#25 Plants-d	Alias. (Downsampl.)	yes	NN (–)	γ=2.2	PBRT / Matlab
#26 Plants-u	Upsampl. (Lanczos)	yes	NN,Lanczos (–)	γ=2.2	PBRT / Matlab
#27 Room Teapot	Struc-Noise	yes	RC (RC)	γ=2.2	PBRT
#28 Sala	Struc-Noise	no	RC (PPM)	[Drag.]	RC+PM phot.: $0.5 \cdot 10^6$ ($5 \cdot 10^9$)
#29 Sanmiguel	Aliasing/Bright	yes	RC,RC (RC)	γ=2.2	PBRT 1 spp (16 spp)
#30 Sanmiguel cam3	Light leaking	yes	PM (RC)	γ=2.2	PBRT
#31 Sanmiguel cam4	Alias./Struc-N./Bright.	yes	RC,RC (RC)	γ=2.2	PBRT 1 spp (16 spp)
#32 Sibenik	VPL Clamp.	no	RC (PPM)	[Rein.]	RC+LC vpls: $0.5 \cdot 10^6$ ($2 \cdot 10^9$)
#33 Sponza	Light leaking	no	PM (RC)	γ=1.8	PBRT
#34 TT	Alias./Noise/Struc-N.	yes	RC,RC (RC)	γ=2.2	PBRT 1 spp (16 spp)
#35 Villa cam1	Noise/Struc-Noise	yes	RC,RC (PM)	[Man.]	PBRT
#36 Villa cam2	Alias./Struc-N./Bright.	yes	RC (PM)	[Man.]	PBRT
#37 Villa cam3	Struc-Noise	yes	RC (PM)	[Man.]	PBRT

Table 1: *Our dataset, from left to right: the scene identifier, distortion type(s), if manually blended by a mask, the rendering method (reference algorithm and settings in parenthesis), tone mapping, and the relevant rendering parameters (if known) used to generate our image dataset (e.g., Fig. 2). The tone mapping operators Fat., Rein., Drag., Man., Man.'06 correspond to [Fattal et al. 2002], global version of [Reinhard et al. 2002], [Drago et al. 2003], [Mantiuk et al. 2008], [Mantiuk et al. 2006], respectively. GL stands for an OpenGL based deferred-renderer using shadow maps with percentage closer filtering (PCF). IGI is an instant global illumination renderer, which supports glossy virtual point lights (VPLs). RC stands for irradiance or radiance caching either in combination with photon-maps (RC+PM) [Křivánek et al. 2005] or lightcuts (RC+LC) [Herzog et al. 2009]. The reference solutions are computed either by pathtracing (PT) or bidirectional pathtracing (Bi-PT) with a constant number of samples per pixel (spp), the lightcuts algorithm (LC) [Walter et al. 2005] with* 1% *error threshold, or progressive photon mapping (PPM) [Hachisuka et al. 2008]. Some images were blended with artifacts of two different strengths or types, which is indicated by the comma-separated method.*

scene. In the following section we describe the design, procedure, and results of the perceptual experiment that we conducted to gather subjective labeling of artifacts in rendered images.

3.1 Stimuli

Table 1 summarizes the rendering algorithms and the distortions that were introduced to the images. Stimuli #1 – #10 come from our previous *EG'12 dataset* [Herzog et al. 2012], while in this work we performed a similar but more extensive experiment (stimuli #11 – #37) in a more rigorous setup. The key differences between the datasets are outlined in the Section 1, and they are further discussed in the supplementary material.

All scenes were rendered into high-dynamic-range images and tone mapped for display as indicated in Table 1. Each scene was rendered using a high- and low-quality setting. In some cases a few distortions of different character were introduced by varying different rendering parameters. Finally, the high quality image was in some cases manually blended (column *Mask* in Table 1) with the corresponding low quality image to reveal the distortions in randomized regions. This additional level of randomness was necessary, as many distortions appeared consistently either in low-illuminated parts of the scene or near the edges. Without blending, the observers were likely to learn the typical locations for a particular artifact and mark them regardless whether the artifact was noticeable/objectionable or not. Some test scenes were blended with more than one distorted image to contain distortions of very different character (in those cases more than one *Method* appears in Table 1). This was meant to test whether a metric can handle a mixture of heterogeneous distortions and account for their impact on image quality.

We now briefly summarize the distortions we have encountered in various rendering algorithms, which are also listed in Table 1. We restrict ourselves to typical global illumination (GI) related artifacts and do not cover banding, tessellation, shadow bias or other more specific artifacts that mostly arise in real-time rendering. For more details about the nature of the individual rendering-specific artifacts we refer the interested reader to our supplementary material. Furthermore, for the later analysis and readability we manually clustered the numerous distortion types into one of six distortion categories which share a similar subjective appearance.

High-frequency noise is probably the most common error encountered in photo-realistically rendered images, which arises as a by-product of all random sample-based integration techniques (e.g., path-tracing, progressive photon mapping [Hachisuka et al. 2008]).
Structured noise represents the class of distortions with correlated pixel errors, which exhibit both noise *and* bias. These are for example interpolation and caching artifacts commonly encountered in approximate GI algorithms such as photon mapping [Jensen 2001], instant radiosity [Keller 1997] as well as the popular (ir-)radiance caching algorithm [Ward et al. 1988; Křivánek et al. 2005].
VPL clamping and light leaking: approximate GI algorithms systematically introduce local errors, often even intentionally, in order to hide the more visually disturbing artifacts (noise). *VPL clamping* in instant radiosity and *light leaking* in photon mapping and irradiance caching fall into this category.
Brightness: another distortion we have noticed is a consistent change in brightness in large regions of an image. Reasons for this can be of systematic nature (e.g., wrong normalization, incorrect material usage) or approximative nature (e.g., only one-bounce indirect light, no caustics, only diffuse VPLs are computed).
Aliasing is the result of insufficient super-sampling or missing prefiltering during rendering. Our examples comprise aliasing in synthetic images including shadow maps, which we partially generated by downsampling the reference image followed by upsampling to the original resolution using the nearest neighbors approach.
Tone mapping can introduce disturbing artifacts, in particular if local gradient-based tone mapping operators (TMO) are applied. Therefore, we included examples of two gradient TMOs into our test set: typical halo artifacts appear in the *buddha* (#14) scene (refer to Fig. 1) [Fattal et al. 2002], and characteristic gradient "leaking" is demonstrated in the *mpii* (#24) scene [Mantiuk et al. 2006].

3.2 Participants and apparatus

A total of 35 observers (11 females and 24 males; age 19 to 52 years) took part in our experiments, and 21 of them completed the no-reference followed by with-reference sessions. The first group of 17 observers consisted of computer graphics students and researchers (denoted as *Experts* in the further analysis), while 18 observers were naïve to the field of computer graphics (denoted as *Non-experts*). All observers had normal or corrected-to-normal vision, and they were naïve as to the purpose of the experiment.

The evaluated images were displayed on two characterized and calibrated displays: 1) LCD Barco Coronis MDCC 3120 DL display (10-bit, 21-inch, 2048 $\times$ 1536 pixels), and 2) NEC MultiSync PA241W display (10-bit, 24-inch, 1920 $\times$ 1200 pixels). The calibration was performed using the X-Rite i1 Display Pro colorimeter (to D65, 120 $\mathrm{cd/m^2}$, colorimetric characterization by means of measured ICC profiles). The experimentation room was neutrally painted, darkened (measured light level: 2 lux), and the observers sat: 1) 71 cm from the Barco display, and 2) 92 cm from the NEC display, which corresponds to 60 pixels per visual degree. The observing distance was enforced by using a chin-rest.

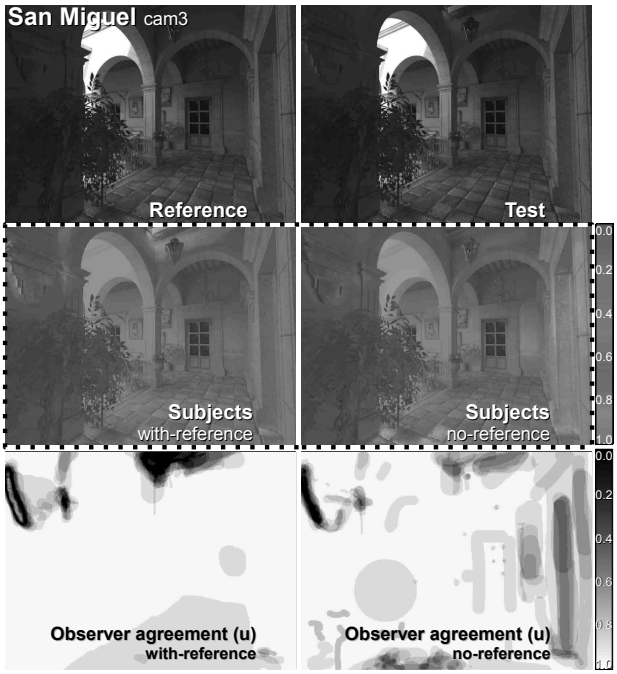


Figure 2: *Example reference and distorted images from our test set along with the mean observer data and maps of Kendall's* u *for both experiments. (Please refer to supplementary material for all the images.)*

3.3 Experimental procedure

We performed two experiments: in the first (*no-reference*) experiment, the observers saw only a distorted image exhibiting rendering artifacts, while in the second (*with-reference*) experiment the distorted image was presented next to the high-quality reference-image. In each experiment the sequence of images was randomized. We asked the observers to freely mark the image regions where they could see artifacts using a custom brush-paint interface. The brush size could be reduced up to per-pixel resolution by the user.

Each observer was introduced to the problem before the experiment as follows. In the no-reference experiment, the observers were instructed to label all the areas in the image with objectionable distortions. In the with-reference experiment, the observers were asked to mark those regions in the distorted image, where they could notice any differences with respect to the reference image. Each experiment took on average 30 minutes per observer. Note that the subjective distortion maps for images #1 – #10 were taken from our previous dataset [Herzog et al. 2012].

4 Analysis of subjective data

In this section we show that the data indicates high agreement between observers, giving evidence that the experimental method is reliable. Then, we analyze differences between the with-reference and no-reference experiments.

4.1 Inter-observer agreement

The experimental task of marking distortions seems challenging, especially in the no-reference setup, so the variations between observers are expected to be high. If the task is deemed to be impossible, we can expect to see little agreement in the distortion maps produced by individual observers. To test the inter-observer agreement, we compute *Kendall's coefficient of agreement* (u) per pixel [Salkind 2007]. The coefficient u ranges from $u = -1/(o-1)$, which indicates no agreement between o observers, to $u = 1$ indicating that all observers responded the same. An example of such a map of coefficients for the *sanmiguel_cam3* (#30) scene is shown in Fig. 2. The complete set of per-scene coefficients can be found in the supplementary materials.

To get an overall indicator of agreement, an average coefficient $\overline{u}$, is computed for each scene. Such overall coefficient is skewed toward very high values because most pixels did not contain any distortion and were consistently left unmarked by all observers. Therefore, we also compute a more conservative measure $\overline{u}_{mask}$, which is equal to the average u of only those pixels that were marked as distorted by at least 5% of the observers.

The values of $\overline{u}$ and $\overline{u}_{mask}$ averaged across the scenes were 0.78 and 0.41 for the with-reference experiment, and 0.77 and 0.49 for the no-reference experiment. These values are relatively high as compared to the values typically reported in such experiments. For example, Ledda et al. [2005] reported u between 0.05 and 0.43 for the task of pairwise comparison of tone mapping operators. This let us believe that the observers can reliably perform the distortion marking task even without much experience or knowledge of the underlying distortions.

4.2 With-reference and no-reference experiments

The main motivation for two experimental designs was to study the relationship between *noticeable* (the with-reference experiment) and *objectionable* distortions (the no-reference experiment). Fig. 3 shows the correlation of the probabilities of marking distortions for both experiment designs. The Spearman correlation values are very high: 0.88 for EG'12 and 0.85 for the new dataset, though these values can be biased by a larger size of unmarked regions. Such strong correlation is a further evidence that the task is well defined and, even in the no-reference experiment, the observers perform consistently and detect most distortions they would detect in the with-reference experiment. The regression line for our dataset in Fig. 3 indicates that fewer observers are marking the same distortions in the no-reference experiment.

To get further insight, we analyze differences in individual images. To find the regions that were marked systematically different between both with- and no-reference experiments, we perform the non-parametric *Kruskal-Wallis* test between the results of both experiments [Salkind 2007]. The test is run separately for each pixel, resulting in the map of p-values as visualized in Fig. 4. Note that although $p < 0.05$ should indicate that two pixels were marked statistically significantly different in the two experiments, this is only the case if each pixel is considered as an independent measurement. Given the high spatial consistency of the markings, per-pixels measurements are unlikely to be independent. However, such

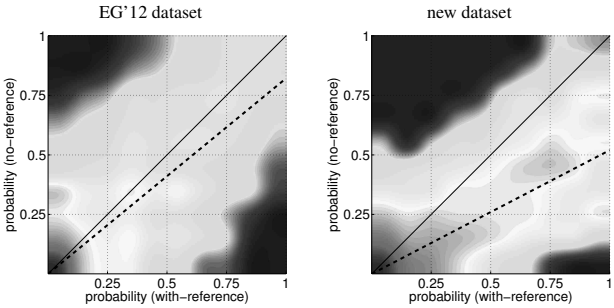


Figure 3: *The relation between the probability of marking a region in the with- and no-reference experiments, plotted separately for each dataset. Similar plots for individual scenes can be found in the supplementary materials. The dashed line shows a linear least-squares regression. The color map was generated from the logarithm of the joint probabilities. The results of with- and no-reference experiments are strongly correlated with fewer observers marking the same regions in the no-reference experiment.*

p-measure is a good indicator of relevant differences in the lack of a suitable statistical test for our dense pixel-based measurements.

The comparison of with- and no-reference results in Fig. 4 shows that the observers sometimes marked regions in the no-reference experiment which were left unmarked in the with-reference experiment. The *buddha* (#14) scene for example exhibits aliasing on the pedestal of the statue (marked in red in Fig. 4 (left)), which was not marked in the with-reference experiment because it was also present in the reference image. However, there were only few such cases in the entire dataset, which were all due to the imperfections of the reference image. In the majority of the cases the observers missed more differences when not seeing the reference image. For example, the brightness change in the background of the buddha statue caused by tone mapping (shown as green in Fig. 4 (left)) was seldom marked in the no-reference experiment.

The number of differences between both experiments indicates that both tasks are different. But at the same time, the high correlation

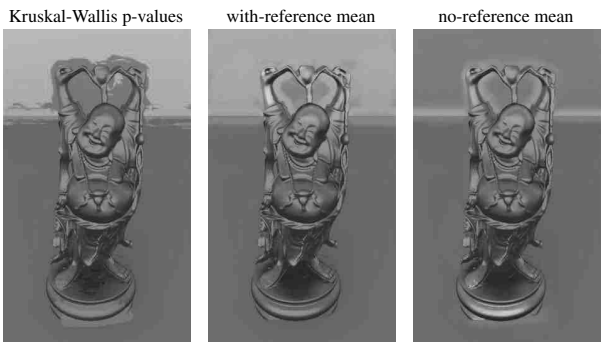


Figure 4: *Differences between with- and no-reference results for the* buddha *scene. The left image shows in green the pixels that were missed in the no-reference experiments (false negatives) and in red those that were marked despite the lack of a difference between the test and reference image (false positives). Only those areas are marked for which the p-values from the Kruskal-Wallis test is less than 0.05. Observers missed in the no-reference experiment the smooth gradient and brightness changes due to tone mapping, but marked aliasing that was also present in the reference image.*

values show that many artifacts are salient enough to be spotted in both with- and no-reference conditions. Note that both experimental designs are still *less conservative* than a typical detection measurement, which involves some form of temporal presentation of co-located test and reference images, for example by sequentially showing the test and reference images in the same screen location. Please refer to the project webpage for such presentation of the differences.

We performed a similar analysis to compare the differences between the expert and non-expert observers. However, we found only a few isolated cases in all scenes where the experts spotted more distortions, such as darkening of corners due to VPL clamping. More extensive discussions of these differences can be found in the supplementary materials.

5 Evaluation of quality metrics

In this section we investigate the performance of existing IQM in detecting distortions marked by the subjects in our experiments. At first, we justify our metric selection and briefly characterize each metric's strength. Then, we present statistical tools that we used for their performance analysis and discuss the outcome.

5.1 Image quality metric selection

Numerous IQM evaluations clearly show that it is impossible to indicate a single metric that performs steadily well for all tested stimuli [Lin and Kuo 2011; Larson and Chandler 2010]. The most problematic cases include images with spatially varying artifacts of different magnitude, as well as mixed distortion types and less common distortions [Lin and Kuo 2011]. Our dataset represents well all such difficult cases. Our choice of metrics in this study is based on the observation that metrics involving perceptual or statistical modeling perform significantly better than PSNR [Wang and Bovik 2006; Lin and Kuo 2011]. Nevertheless, because of its popularity for image quality evaluation in computer graphics, we also consider a simple *absolute difference (AD)* metric that is directly related to the commonly used RMSE and PSNR. We use absolute rather than squared differences because our statistical analysis is robust to any monotonic transformations, such as the quadratic power function.

Another popular choice in graphics is CIE-Lab, but here due to even more favorable conformance with image distortion maps [Zhang and Wandell 1998] we select its spatial extension *sCIE-Lab*. HVS-based metrics are represented by *HDR-VDP-2* [Mantiuk et al. 2011], which provides much improved predictions with respect to its predecessors HDR-VDP [Mantiuk et al. 2005] and VDP [Daly 1993]. Also, we investigate the *SSIM* that is often reported as the most reliable metric [Larson and Chandler 2010], as well as its multi-scale version *MS-SSIM* [Wang et al. 2003], which accounts for structural and contrast changes at different scales to compensate for the variations of image resolution and viewing conditions. MS-SSIM is reported as the best-performer in many IQM comparison studies [Sheikh et al. 2006; Ponomarenko et al. 2009]. Finally, we include as a metric the *Spearman rank-order correlation (sCorrel)* computed over local 8×8-pixel blocks, which can be regarded as a subset of the SSIM functionality, to better understand the importance of eliminated contrast and lightness factors.

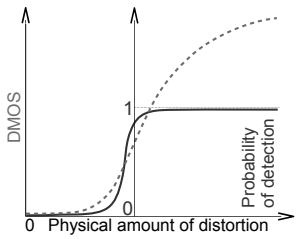
Our metric selection is also representative with respect to computational complexity. AD and sCIE-Lab are attractive due to their mathematical simplicity. On the other hand, HDR-VDP-2 is the most complex but has been shown to successfully predict near-threshold distortions. The medium complexity SSIM has been demonstrated to meaningfully estimate the magnitude of supra-threshold distortions, while its sensitivity to near threshold distor-

tions seems to be more problematic due to the lack of explicit HVS modeling. The sCIE-Lab prediction also conforms to the distortion magnitude and its sensitivity to spatial color patterns is based on the HVS-model. MS-SSIM seems to bridge the gap between SSIM and HDR-VDP-2 by emphasizing on the structural differences while processing at multi-scale.

To account for the differences in viewing distance between our two datasets, the parameters of the metrics that respect a viewing distance (HDR-VDP-2 and sCIELab) were adjusted accordingly, and for the other metrics images were resampled to match the angular resolution of 30 pixels-per-visual-degree.

5.2 Statistical measures of metric performance

It is important to recognize that, in contrast to other image quality experiments, our measurements do not capture the perceived magnitude of distortion. For that reason we need to use different measures for the metric performance. Most image quality assessment experiments measure a single scalar differential-/mean-opinion-score (DMOS/MOS) per test image, shown as the dashed red line in the plot on the right. The non-parametric correlation between a metric and the MOS values is considered as a measure of the metric's performance [Sheikh et al. 2006; Ponomarenko et al. 2009]. Unfortunately, there is no method to measure MOS efficiently for each location in an image. Our experiments capture how likely an average observer will notice *local* distortions, shown as the continuous blue line. It is correlated with MOS in the limited range where the psychometric function (blue line) does not saturate. If this probability of detection is equal or close to either 0 or 1, we have no information about the perceived magnitude of a distortion.



However, our data is well suited to benchmark the metrics ability to spot problematic regions in terms of binary classification: marking the pixels that contain noticeable or objectionable distortions. The performance of such classification is usually analyzed using the receiver-operator-characteristic (ROC) [Baldi et al. 2000]. ROC captures the relation between the size of regions that contain distortions and were correctly marked by a IQM (true positives), and the regions that do not contain distortions but were still marked (false positives). ROC captures the relation of these two quantities for a varying classification threshold. The metric that produces a larger area under the ROC curve (AUC) is assumed to perform better. To simplify considerations it is convenient to assume that a certain percentage of observers need to mark the distortion to consider it noticeable. In Fig. 5 (top-left) we present the results for regions marked by 50% or more observers, but the supplementary materials also include the data for the $\geq$25% and $\geq$75% criteria.

However, AUC values alone may give a wrong impression of the actual metric performance because usually only a small portion of the pixels in the images of our experiments showed distortions. Thus, the reference classification data is strongly unbalanced. For that reason, in addition to ROC, we also plot *Matthews correlation coefficient* [Baldi et al. 2000], which is robust to unbalanced classification data. The coefficient indicates correlation of classification data in the range from -1 to 1, where +1 represents a perfect prediction, 0 no better than random prediction, and -1 indicates total disagreement between prediction and observation.

5.3 Metric performance comparison

The key question is whether any of the IQM performs significantly better than the others in terms of detecting noticeable or objectionable graphics artifacts. The overall metric performance for both datasets and the two experimental designs is summarized in Fig. 5. Such summary, however, requires careful interpretation before any winning or loosing metric can be indicated.

Generalization of ranking. Although the ranking in Fig. 5 is a good summary of metric performance for a particular dataset, care must be taken when extrapolating any conclusions outside our measured data. To test robustness of our ranking to randomization of images, we computed the distribution of AUC by bootstrapping the set of images used for each experiment. The procedure involved computing AUC values 500 times, each time for a random set of images selected from the original set, so that the number of images was the same as in the original dataset and some of them could appear more than once (randomization with repetition) [Howell 2007, ch.18]. The computed 500 AUC values resulted in the distribution, which allowed for statistical testing. After applying Bonferroni's adjustment to compensate for multiple comparisons [Howell 2007, p.377], we found *no statistically significant differences between any pair of the metrics* in the EG'12 dataset, and *only one significant difference* between the metrics on the extreme ranking positions in the new no-reference dataset. This means that neither dataset provides conclusive evidence that any of the metrics is better than the others in a general case, and we cannot generalize the presented rankings to the entire population of images and distortions. The main reason for this is that the individual metric performance differs significantly from image to image depending on the nature of the underlying distortions. Therefore, *no* IQM is robust enough to perform significantly better for the distortions contained in our dataset.

It is important to note that our method of statistical testing differs from the methods used in other IQM comparison studies, such as [Sheikh et al. 2006] and [Ponomarenko et al. 2009]. The statistical testing employed in these studies was meant to prevent false hypothesis only due to the variance in subjective responses. The results of those statistical tests show that the ranking of the metrics is very likely to be the same for a different group of observers while assuming that the *same* set of images and distortions is used. Our testing is much more demanding as it requires the metric to perform better for any set of images (taken from the original population) in order to be considered better in the statistically sense.

Overall metric performance. Due to the unbalanced ratio of the marked and unmarked regions, we refer to *Matthews correlation coefficient* instead of the AUC values to assess the overall performance of the metrics. The average values of the Matthews coefficient for all scenes as shown in Fig. 5 are low, ranging from 0.2 to 0.35 for the EG'12 dataset, and between 0.25 and 0.45 for the new dataset. These values are much lower than Spearman's rank order correlation of 0.953 reported for the LIVE database [Sheikh et al. 2006] and 0.853 reported for the TID2008 database [Ponomarenko et al. 2009] for the best metric (MS-SSIM). However, it must be flagged that Spearman's correlation, although also scaled from -1 to 1, is different to Matthews coefficient, as discussed in Section 5.2. The low correlation values indicate that classifying distortions in the distortion maps is a much more difficult task than correlating a single value per image with the MOS. It also means that our dataset is a more demanding and accurate test for IQM since it can point out the areas where the metric's performance could be improved. Fig. 6 summarizes the Matthews correlation coefficients between the metric predictions and subjective responses in the with-reference experiment. As can be seen the highest correlation is achieved for the high-frequency-noise distortions, while for high-contrast structured

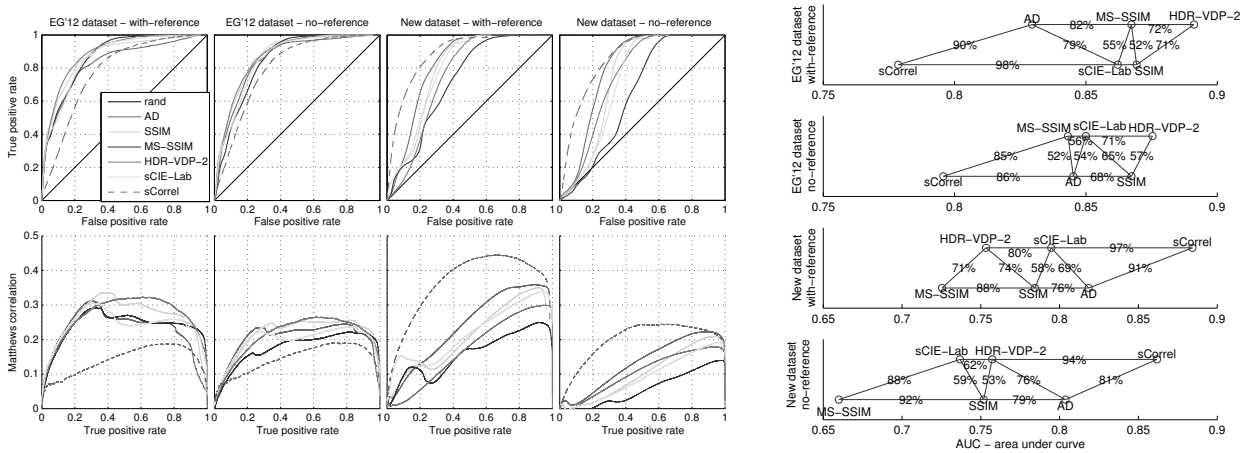


Figure 5: *The performance of quality metrics shown as ROC plots (top-left), Matthews correlation (bottom-left) and ranked according to the area-under-curve (AUC) (right) (the higher the AUC, the better the classification into distorted and undistorted regions). The percentages indicate how frequently the metric on the right results in higher AUC when the image set is randomized using a bootstrapping procedure.*

noise with only localized appearance (e.g., in the *kitchen* (#4) and *red kitchen* (#5) scenes) the correlation drops abruptly. Images with mixed distortions seem to be problematic as well.

5.4 Analysis of image quality metric failures

The ranking plots in Fig. 5 reveal different performance of the metrics for both datasets. HDR-VDP-2 performed the best for the EG'12 dataset, but was the second to the last in the new dataset. Surprisingly, the simple non-parametric correlation metric sCorrel performed the best for the new dataset, but at the same time it was the worst metric for the EG'12 dataset. This unexpected result cannot be easily explained by looking at the aggregated results and requires investigating individual images. In the following we summarize our analysis of individual images and reveal the most pronounced cases of metric failure.

Brightness and contrast change is a very common artifact of many rendering algorithms, as discussed in Section 3.1, and also the cause of failure of most advanced IQM. The best example of that is the *sala* (#28) scene shown in Fig. 7. The brightness differs significantly between the test and reference images for all surfaces, but the observers marked only the floor and in a lesser extent the walls, both affected by low-frequency noise. The noise was more visible on the floor than on the walls because the floor lacked texture and thus did not mask the noise. One metric that excelled in this task was sCorrel, with Matthews correlation exceeding 0.6. This is because non-parametric correlation is also invariant to non-linear transformations of pixel values, including low-frequency brightness changes. The second best performing metric, sCIELab, contains a band-pass model of the CSF, which attenuates low-frequency variations and thus makes this metric more robust to brightness changes. Although HDR-VDP-2 also includes a band-pass CSF model, it is far too sensitive to contrast changes to disregard numerous suprathreshold pixel modifications. Even MS-SSIM, which partially relies on the measure of correlation, did not perform much better than a random guess for this image. This shows that invariance to brightness and contrast changes must be an essential feature of any IQM that needs to reflect the observers' performance in the side-by-side comparison or non-reference tasks.

Visibility of low-contrast differences. For several scenes the test images have been computed using instant global illumination (IGI) while the reference images have been generated by path tracing, which often features certain amount of stochastic per-pixel noise. One example of such an image pair is the *disney* (#3) scene shown in Fig. 8. While the stochastic noise in a well-converged image is usually invisible, and thus remains unmarked in subjective experiments, it clearly affects the absolute pixel values and image struc-

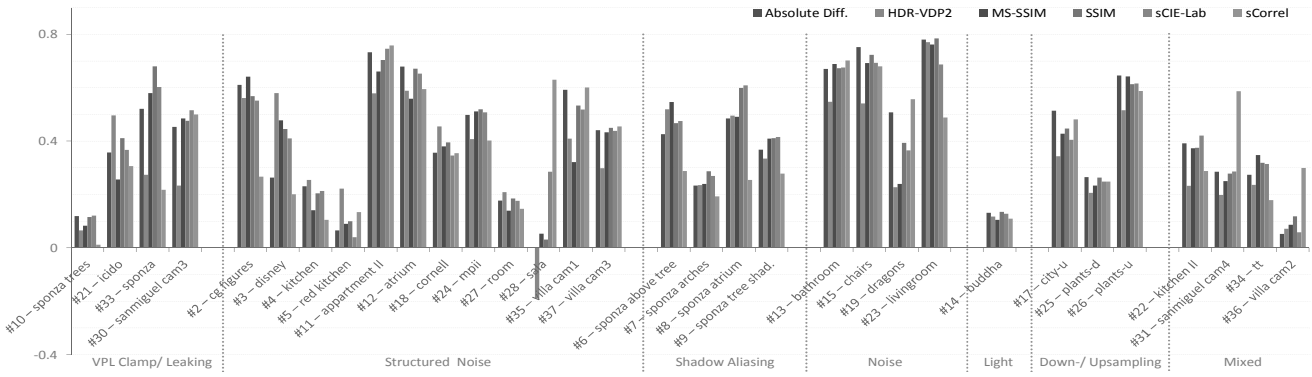


Figure 6: *Matthews correlation coefficient for predictions of HDR-VDP-2, SSIM, MS-SSIM, sCIE-Lab, sCorrel, and Absolute Difference with respect to subjective responses (with-reference experiment). Results are grouped according to the type of artifact as indicated at the bottom.*

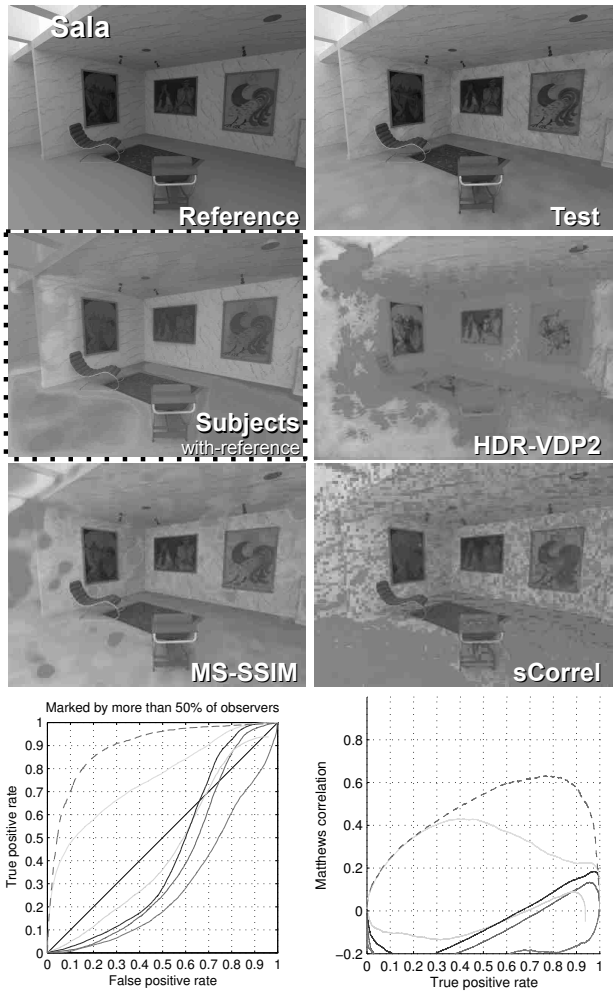


Figure 7: *Scene* sala *(top), distortion maps for selected metrics (2nd and 3rd rows), ROC and correlation plots (bottom). Most metrics are sensitive to brightness changes, which often remain unnoticed by observers.* sCorrel *is the only metric robust to these artifacts. Refer to the legend in Fig. 5 to check which lines correspond to which metrics in the plots.*

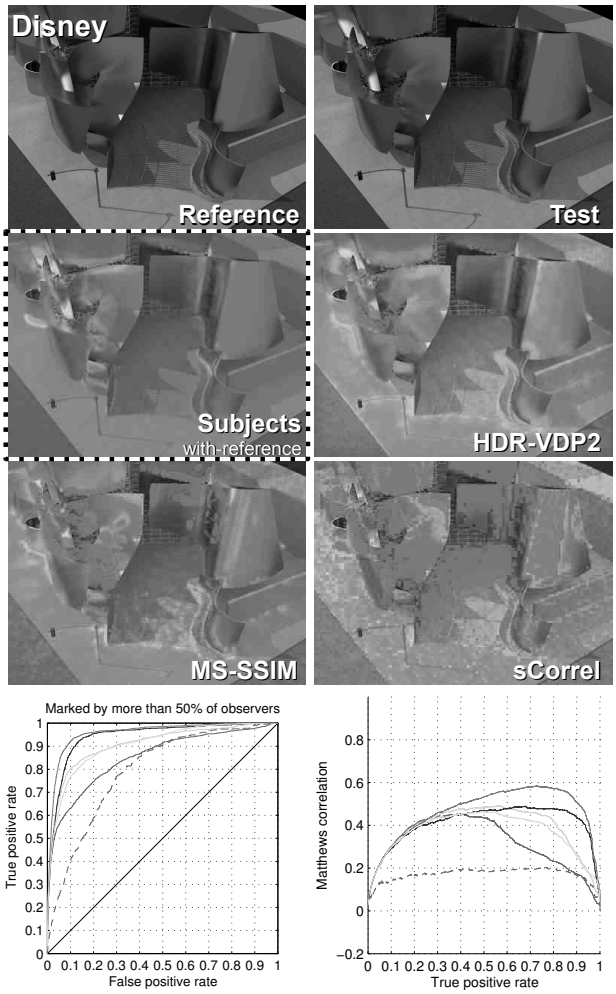


Figure 8: *Scene* disney*: simple metrics, such as sCorrel and AD, fail to distinguish between visible and invisible amount of noise resulting in worse performance.*

ture. Both AD and sCorrel metrics are sensitive to such differences, so they report distortions regardless of their visibility. What makes sCorrel insensitive to global brightness changes, makes it also insensitive to the amplitude of the noise, which prevents this metric from finding a reliable visibility threshold. For that reason both metrics poorly correlate with subjective data, as seen in the plot of Fig. 8. The metrics specifically tuned for near threshold signal detection, such as HDR-VDP-2, performed much better in this task. This stresses the importance of proper visual system modeling, which improves the metric's accuracy for the near-threshold distortions.

Plausibility of shading. A similar kind of distortion can be seen differently depending whether it leads to plausible or implausible shading. For example, two scenes shown in Fig. 9 contain VPL clamping and photon leaking distortions, respectively, near the corners. In the case of the *sponza* (#33) scene photon leaking results in brightening of dark corners. This was marked as distortion by most observers because bright patches are unlikely to be found in dark corners. However, the VPL clamping in the *sibenik* (#32) scene resulted in the opposite effect, the corners were darkened. Such distortion was marked by much fewer observers because darkening could have resulted from self-shadowing and in fact appeared realistic in the given context. All metrics failed to distinguish between these two cases. This suggests that robust IQM may require a higher-level analysis of scene and illumination that could distinguish between plausible and implausible patterns of illumination. This is difficult to achieve if images are the only source of information, but could be possible if information about the 3D scene and its shading were available [Herzog et al. 2012].

Spatial accuracy of the prediction map. Many sophisticated metrics perform often worse than the AD because they are unable to precisely localize distortions. This is well visible in the *dragons* (#19) scene shown in Fig. 10. The distortion maps for MS-SSIM show visible differences that widely disperse from the edges of the dragon figures into the background regions that do not contain any physical difference. This problem affects mostly multi-scale metrics, such as MS-SSIM and HDR-VDP, but SSIM is also affected because of its 8×8 sliding window approach, which limits the effective accuracy of the distortion map. This observation suggests that the metrics should employ techniques that respect object boundaries and thus can produce more accurate distortion maps.

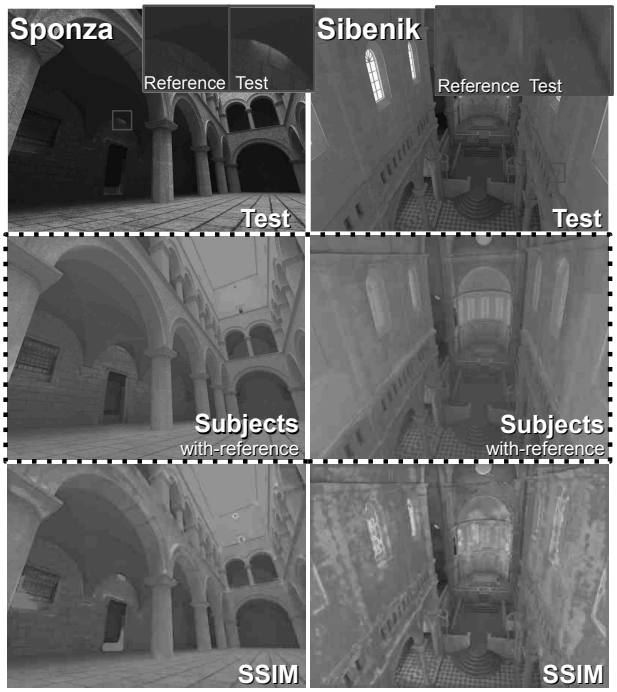


Figure 9: *Photon leaking and VPL clamping artifacts in scenes* sponza *and* sibenik *result in either brightening or darkening of corners. However, darkening is subjectively acceptable, whereas brightening leads to objectionable artifacts.*

6 Conclusions and future work

In this work we propose rendering-oriented datasets for image quality evaluation, which provide detailed distortion maps along with the probability of their detection by human observers. We show that *objectionable distortions* marked by the observers that did not see the reference image are strongly correlated in terms of their spatial location with the distortions marked in the presence of the reference image. This may suggest that by further improvement of *full-reference* IQM, we can achieve quality predictions similar to *no-reference* human judgments, which should be an easier task than the development of a *no-reference* IQM that directly mimics the human perception. *Full-reference* perceptual experiments, on the other hand, may potentially be approximated by a no-reference experiment if a reference image is not available.

For existing full-reference IQM our datasets turned out to be very demanding, and our analysis of metric failures suggests directions for improvement. The relatively good performance of the simplistic non-parametric correlation measure (sCorrel) clearly indicates its importance. Although SSIM and MS-SSIM also incorporate a correlation factor their performance is strongly influenced by their excessive sensitivity to brightness and contrast changes. Clearly, near-threshold contrast accuracy is important to disregard all non-noticeable distortions. At the same time proper spatial distortion localization is required, which is the problem for all multi-scale approaches, in particular, in the proximity of high contrast distortions. In general, the performance of state-of-the-art IQM in graphics applications is not very consistent, and one should not be too reliant on them. In particular the IQM originating in the image/video compression community may not be the most suitable for graphics applications where the artifacts are often very distinct.

We believe that all those insights are essential towards improving existing metrics or developing new ones, which we relegate as future work. Upon the public release our datasets should be useful to train such future metrics and compare their performance. However, for a systematic and quantitative study of metric failures further experiments are required.

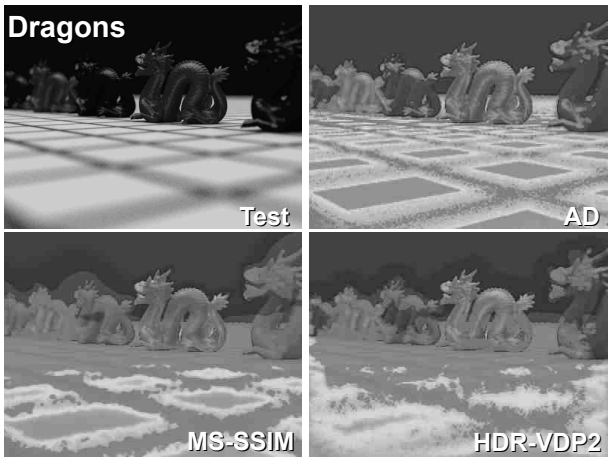


Figure 10: Dragons *scene contains artifacts on the dragon figures but not in the black background. Multi-scale IQM, such as MS-SSIM and HDR-VDP-2, mark much larger regions due to the differences detected at lower spatial frequencies. Pixel-based AD can better localize distortions in this case.*

Our datasets provide the probability of noticing distortions, which could offer interesting insights on the saliency of artifacts in rendering. Such artifact saliency could be investigated in the context of comparing a pair of images, searching for distortions within a single image, as well as task-free image inspection. Similarly to the concept of visual equivalence [Ramanarayanan et al. 2007], objectionable distortions dictate less conservative requirements on image quality, thus enabling further computational savings when used as the measure of desirable quality.

Our published datasets could also be interesting for the broader vision science community, as the complex stimuli presented in our experiments differ significantly from the usual "laboratory" ones and enable inspection of higher-level vision tasks. However, more experiments (based on photo-realistic images) are clearly needed as well as a further study of cognitive factors in the quality assessment, such as inattentional blindness or task fatigue. To this end, a speculative question raised by our results is whether it is beneficial and promising at all to model the early human vision processes (bottom-up modeling) or whether we should concentrate on data-driven approaches that are statistically trained on subjective results (top-down modeling). The bottom-up approach may result in worse than expected predictive power for complex images, while the top-down approach is prone to over-training as image quality databases will offer only very limited sample from the huge population of all potential images and distortions. This study is a step towards combining both approaches that enables training and testing the metrics of any complexity on the per-pixel basis.

Acknowledgements

We thank the creators of the test scenes used in the experiments, in particular to T. Davidovič for VPL renderings (#2, #3, #4, #5), A. Voloboy for MCRT renderings (#15, #20), I. Georgiev for bidirectional pathtracing results (#13, #23), and to the observers at Bangor University and MPII who participated in our experiments. This work was partly supported by the EPSRC research grant

EP/I006575/1 and COST Action IC1005 (STSM-100312-015783).

References

BALDI, P., BRUNAK, S., CHAUVIN, Y., ANDERSEN, C. A. F., AND NIELSEN, H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics 16*, 5 (May), 412–424.

BOLIN, M., AND MEYER, G. 1998. A perceptually based adaptive sampling algorithm. In *Proc. of SIGGRAPH*, 299–310.

ČADÍK, M., AYDIN, T. O., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2011. On evaluation of video quality metrics: an HDR dataset for computer graphics applications. In *SPIE HVEI XVI*.

CHANDLER, D., AND HEMAMI, S. 2007. VSNR: a wavelet-based visual signal-to-noise ratio for natural images. *IEEE Trans. on Image Processing 16*, 9, 2284–2298.

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

DRAGO, F., MYSZKOWSKI, K., ANNEN, T., AND N.CHIBA. 2003. Adaptive logarithmic mapping for displaying high contrast scenes. *Proc. of Eurographics 22*, 3, 419–426.

FATTAL, R., LISCHINSKI, D., AND WERMAN, M. 2002. Gradient domain high dynamic range compression. In *Proc. of SIGGRAPH*, 249–256.

HACHISUKA, T., OGAKI, S., AND JENSEN, H. W. 2008. Progressive photon mapping. In *Proc. of SIGGRAPH Asia*, 130:1–130:8.

HERZOG, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2009. Anisotropic radiance-cache splatting for efficiently computing high-quality GI with lightcuts. *Proc. of Eurographics*, 259–268.

HERZOG, R., ČADÍK, M., AYDIN, T. O., KIM, K. I., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2012. NoRM: no-reference image quality metric for realistic image synthesis. *Computer Graphics Forum 31*, 2, 545–554.

HOWELL, D. C. 2007. *Statistical Methods for Psychology*, 6th edition ed. Thomas Wadsworth.

ITU-R-BT.500-11, 2002. Methodology for the subjective assessment of the quality of television pictures.

ITU-T-P.910. 2008. Subjective audiovisual quality assessment methods for multimedia applications. Tech. rep.

JENSEN, H. W. 2001. *Realistic Image Synthesis Using Photon Mapping*. AK, Peters.

KELLER, A. 1997. Instant radiosity. In *Proc. of SIGGRAPH*, 49–56.

KŘIVÁNEK, J., GAUTRON, P., PATTANAIK, S., AND BOUATOUCH, K. 2005. Radiance caching for efficient global illumination computation. *IEEE TVCG 11*, 5, 550–561.

LARSON, E. C., AND CHANDLER, D. M. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electron. Imaging 19*, 1, 011006:1–21.

LEDDA, P., CHALMERS, A., TROSCIANKO, T., AND SEETZEN, H. 2005. Evaluation of tone mapping operators using a high dynamic range display. *Proc. of SIGGRAPH 24*, 3, 640–648.

LIN, W., AND KUO, C.-C. J. 2011. Perceptual visual quality metrics: A survey. *JVCIR*, 297–312.

LUBIN, J. 1995. *Vision Models for Target Detection and Recognition*. World Scientific, ch. A Visual Discrimination Model for Imaging System Design and Evaluation, 245–283.

MANTIUK, R., DALY, S., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2005. Predicting visible differences in high dynamic range images - model and its calibration. In *SPIE HVEI X*.

MANTIUK, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2006. A perceptual framework for contrast processing of high dynamic range images. *ACM Trans. on Applied Perception 3*, 3, 286–308.

MANTIUK, R., DALY, S., AND KEROFSKY, L. 2008. Display adaptive tone mapping. In *Proc. of SIGGRAPH*, vol. 27(3), #68.

MANTIUK, R., KIM, K. J., REMPEL, A. G., AND HEIDRICH, W. 2011. HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. In *Proc. of SIGGRAPH*, #40.

PEDERSEN, M., AND HARDEBERG, JON, Y. 2011. Full-Reference Image Quality Metrics: Classification and Evaluation. *Foundations and Trends in Computer Graphics and Vision 7*, 1, 1–80.

PONOMARENKO, N., LUKIN, V., ZELENSKY, A., EGIAZARIAN, K., CARLI, M., AND BATTISTI, F. 2009. TID2008 - A database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics 10*, 30–45.

RAMANARAYANAN, G., FERWERDA, J., WALTER, B., AND BALA, K. 2007. Visual equivalence: towards a new standard for image fidelity. In *Proc. of SIGGRAPH*, #76.

RAMASUBRAMANIAN, M., PATTANAIK, S., AND GREENBERG, D. 1999. A perceptually based physical error metric for realistic image synthesis. In *Proc. of SIGGRAPH*, 73–82.

REINHARD, E., STARK, M. M., SHIRLEY, P., AND FERWERDA, J. A. 2002. Photographic tone reproduction for digital images. In *Proc. of SIGGRAPH*, 267–276.

RUSHMEIER, H., WARD, G., PIATKO, C., SANDERS, P., AND RUST, B. 1995. Comparing real and synthetic images: some ideas about metrics. In *Rendering Techniques '95*, 82–91.

SALKIND, N., Ed. 2007. *Encyclopedia of measurement and statistics*. A Sage reference publication. SAGE, Thousand Oaks.

SHEIKH, H., SABIR, M., AND BOVIK, A. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. on Image Processing 15*, 11, 3440–3451.

WALTER, B., FERNANDEZ, S., ARBREE, A., BALA, K., DONIKIAN, M., AND GREENBERG, D. 2005. Lightcuts: A scalable approach to illumination. *Proc. of SIGGRAPH*, 1098–1107.

WANG, Z., AND BOVIK, A. C. 2006. *Modern Image Quality Assessment*. Morgan & Claypool Publishers.

WANG, Z., SIMONCELLI, E. P., AND BOVIK, A. C. 2003. Multiscale structural similarity for image quality assessment. In *Proc. IEEE Asilomar Conf. on Signals, Systems & Comp.*, 1398–1402.

WARD, G. J., RUBINSTEIN, F. M., AND CLEAR, R. D. 1988. A ray tracing solution for diffuse interreflection. In *Proc. of SIGGRAPH*, 85–92.

WU, H., AND RAO, K. 2005. *Digital Video Image Quality and Perceptual Coding*. CRC Press.

ZHANG, X., AND WANDELL, B. A. 1998. Color image fidelity metrics evaluated using image distortion maps. *Signal Proc. 70*, 3, 201 – 214.

Appendix I

NoRM: No-Reference Image Quality Metric for Realistic Image Synthesis

R. Herzog, M. Čadík, T. O. Aydın, K. I. Kim, K. Myszkowski, and H.-P. Seidel. NoRM: No-Reference Image Quality Metric for Realistic Image Synthesis. *Computer Graphics Forum*, Vol. 31, No. 2, pp. 545–554, 2012.
IF=1.595

EUROGRAPHICS 2012 / P. Cignoni, T. Ertl
(Guest Editors)

Volume 31 (2012), Number 2

NoRM: No-Reference Image Quality Metric for Realistic Image Synthesis

Robert Herzog[1] and Martin Čadík[1] and Tunç O. Aydın[1,2] and Kwang In Kim[1] and Karol Myszkowski[1] and Hans-P. Seidel[1]

[1] MPI Informatik Saarbrücken, Germany, [2] Disney Research Zurich, Switzerland, http://www.mpi-inf.mpg.de/resources/hdr/norm/

Abstract

Synthetically generating images and video frames of complex 3D scenes using some photo-realistic rendering software is often prone to artifacts and requires expert knowledge to tune the parameters. The manual work required for detecting and preventing artifacts can be automated through objective quality evaluation of synthetic images. Most practical objective quality assessment methods of natural images rely on a ground-truth reference, which is often not available in rendering applications. While general purpose no-reference image quality assessment is a difficult problem, we show in a subjective study that the performance of a dedicated no-reference metric as presented in this paper can match the state-of-the-art metrics that do require a reference. This level of predictive power is achieved exploiting information about the underlying synthetic scene (e.g., 3D surfaces, textures) instead of merely considering color, and training our learning framework with typical rendering artifacts. We show that our method successfully detects various non-trivial types of artifacts such as noise and clamping bias due to insufficient virtual point light sources, and shadow map discretization artifacts. We also briefly discuss an inpainting method for automatic correction of detected artifacts.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Image Quality Assessment

1. Introduction

While photo-realistic rendering methods are getting more advanced over time, various rendering artifacts still appear as a problem in the results. These artifacts can be reduced or completely avoided by fine-tuning the rendering algorithm's parameters through trial and error. But this manual process is often time-consuming and requires some level of understanding about the inner machinery of the rendering method in consideration. Analogous to the field of objective image quality assessment where one can use computational *quality metrics* that predict the subjective quality evaluation, the objective quality assessment of synthetic images is highly beneficial because it eliminates the tedious manual labor required otherwise. Additionally, such a metric enables automatic detection and elimination of rendered images of unacceptable quality. To that end we propose an objective *image quality metric for realistic image synthesis* based on a machine learning system trained with various types of rendering artifacts.

Building a quality metric for synthetic images has additional challenges over a metric for natural images. The metrics for natural images are often *full-reference*, namely they rely on a non-distorted copy of the image for evaluating the distorted (test) image. Unlike in applications like compression and watermarking, in rendering such a reference image is often not available in practice and a metric for synthetic images should detect and predict the strength of rendering artifacts based solely on the test image. Although humans often detect distortions just as well without a reference, in contrast non-reference image quality metrics are usually inferior in performance to full-reference metrics [WR05]. Thus, the absence of a reference image is a significant constraint in metric design.

The central idea of this paper is to leverage 3D scene information to compensate for the lack of a reference image while detecting rendering artifacts. Any scene specific per pixel data data beyond color, such as depth, texture and material, is difficult, if at all possible, to obtain reliably in natu-

ral images. This is not the case for rendered scenes, and we show that taking full advantage of this additional information enables non-reference quality assessment of synthetic images with a prediction performance comparable to full-reference metrics.

- a fully automatic metric that detects rendering artifacts without a reference,
- a learning framework for common rendering artifacts that also guides our artifact removal,
- a human visual system-inspired model that predicts the perceived strength of rendering artifacts,
- a dataset of photo-realistic rendering artifacts including subjective artifact probability detection maps.

In a subjective study, we show that the performance of our metric matches the state-of-the-art in full-reference metrics. Our metric could be employed in rendering farms, as well as in controlling the rendering quality in client-server or cloud computing settings. One could also use it as a diagnostic tool for rendering quality, or in an optimization framework to find optimal parameters for a rendering method.

2. Related Work

In this section we review previous work on *non-reference* (NR) image/video quality assessment. First, we discuss the NR metrics for imaging applications, and then, we present rendering-specific solutions. For a detailed discussion of the *full-reference* (FR) and *reduced reference* (RR) quality metrics we refer the reader to the recent textbooks [Win05, WB06, WR05]. FR metrics tailored for computer graphics and HDR imaging applications are summarized in [RWD*10, Ch. 10] and [MKRH11].

NR metrics in imaging applications The key difficulty in developing NR metrics is the absence of a non-distorted reference image or some features representing it. Common approaches to compensate for this are (1) modeling distortion-specific characteristics, (2) using natural scene statistics, and (3) employing learning based classification methods.

Distortion-specific NR methods capitalize on the knowledge of artifact type and its unique characteristics [WR05, Ch. 3]. Examples include metrics for detecting blockiness due to lossy JPEG and MPEG compression and ringing at strong contrast edges [WB06], blurriness due to high frequency coefficients suppression [CCB11, LH11], banding (false contouring) at low gradient regions due to the excessive quantization [DF04]. There are some attempts of building more general NR quality metrics, which evaluate a combined contribution of individually estimated image features such as sharpness, contrast, noise, clipping, ringing, and blocking artifacts [WR05, Ch. 10]. The contribution of all features including their simple interactions is summed up with weights derived through fitting to subjective data.

Natural scene statistics [Sim05] derived from artifact-free images can be helpful in detecting artifacts. Sheikh et al. show that noise, blurriness, and quantization can be identified as deviations from these statistics [SBC05].

Image features extracted from distorted and non-distorted images are used for training machine learning techniques such as support vector machines (SVM) or neural networks. Moorthy and Bovik [MB10] use generalized Gaussian distribution (GGD) to parametrize wavelet subband coefficients and create 18-D feature vector (3 scales $\times$ 3 orientations $\times$ 2 GGD parameters), which is used to train an SVM classifier based on perceptually calibrated distortion examples from the LIVE IQA database. The classifier discriminates between five types of mostly compression-related distortions and estimates their magnitude. Saad et al. [SBC10] train a statistical model to detect distortions in DCT-based contrast and structure features.

Discussion: Our technique differs from previous work in three ways: (1) we use depth buffer and albedo information in addition to color, (2) our output is a distortion *map* rather than a scalar value, and thus, we show spatial distribution of distortions, and (3) our work specializes in rendering artifacts rather than compression/transmission related artifacts.

Rendering-specific NR metrics Some metrics in this category rely on *predicted* reference images. In image-based rendering the mis-registration error of pixels with respect to the ground truth reference image is a good measure of visual quality [KSGH09]. In 3DTV applications the lack of ground truth can be compensated by reprojecting (warping) images from different cameras to the mid-point view [KSGH09]. Also, when temporal frame replication is performed for reducing the rendering cost or display hold-type blur, similar reprojection in temporal domain is feasible [MRT99]. In contrast to these, our method is purely NR in that we do not need to predict a reference. This is also the case for the recent work of Berger et al. [BLL*10] where specialized ghosting detector explicitly works on an interpolated image.

Other work in computer graphics literature includes a model of the elevation of contrast discrimination threshold due to visual masking, which can be predicted based on the texture pattern only [RPG99, WPG02]. An estimator of bias, which mostly leads to blurred shading details, has been proposed within the progressive photon mapping framework [HJJ10]. This estimator relies strongly on intrinsic renderer information such as derivatives of estimated lighting function, which becomes feasible only for density estimation methods with smooth kernel functions. Our data-driven approach aims for using less rendering specific and easier to acquire data. Stokes et al. [SFWG04] introduced a perceptual NR metric, which predicts the contribution of the indirect illumination components towards perceived image quality. While the metric cannot detect local artifacts, similar to our metric it considers per pixel reflectance information.

Ramanarayanan et al. [RFWB07] proposed metrics that

Figure 1: *Example images with artifacts used for our no-reference quality metric. Insets show magnified artifacts regions, letters indicate the type of artifact (C: VPL clampling, G: glossy VPL noise, S: shadow map aliasing). The numbered images correspond to the testset used in the user study.*

utilize per object reflectance and surface bumpiness information for training SVM classifier on subjective data. Their method measures the overall *visual equivalence* instead of identifying problematic image regions. Křivánek et al. [KFB10] investigated visual equivalence for instant radiosity (virtual point light) algorithms and proposed a number of useful rendering heuristics, which were difficult to formalize into a ready to use computational model.

3. Overview

In this work, we are interested in automatically detecting rendering artifacts, which are typical for global illumination solutions (Fig. 1), and that we briefly describe in Section 4. We achieve this via a machine learning approach (see Section 5) based on the discrimination of rendering-specific features (Section 5.2) trained on our generated database of synthetic image pairs (Section 5.1). The whole system is depicted in Fig. 2. Note that we do not intend to classify an image as a whole but rather predict the locations of artifacts in an image. Optionally, we can "clean" the image making effective use of inpainting techniques (Section 7) based on the same set of training image pairs and obtain a "pseudo-reference" image, which is then used to perceptually normalize the distortion map for the visibility of artifacts (Section 8). We present our results in Section 9 and demonstrate in a user study (Section 10) that our method is competitive with state-of-the-art reference methods (VDPs). Finally, we conclude with future prospects in Section 11.

4. Rendering-specific Artifacts

Photo-realistic rendering is still very time-consuming and rendering a high-resolution, globally-illuminated image may take several minutes to hours becoming even more critical in the case of an animation. Therefore, many rendering algorithms trade quality (bias) for speed and often leave it to the user to find the right parameters, eventually resulting in algorithm-specific artifacts, which are hard to control, i.e., the generated image might look fine partially but exhibits strong degradations in small areas.

In our experiments we focused on artifacts inherent to popular rendering algorithms, which comprise *Instant Radiosity* [Kel97] with glossy virtual point lights (VPLs), *Lightcuts* [WFA*05], and OpenGL rasterization using *PCF shadow maps* [RSC87], which produce VPL-based artifacts (i.e., low-frequency noise), clamping bias (darkened corners), and shadow map aliasing (jaggy shadow boundaries), respectively. Examples of images showing these artifacts are given in Fig. 1. We exclude stochastic noise (pixel-variance), e.g., anti-aliasing, path-tracing, from our study, which is much easier to handle and well-studied in the rendering community [BM98, RPG99, TJ97, KA91]. We also do not discuss temporal artifacts, which are beyond the scope of this work [YPG01].

5. Learning Rendering Errors from Examples

Computing the perceived image errors along with the final pixel colors during the rendering process can be very helpful for example to adapt the rendering. However, this is only feasible for easily analyzable errors in very specific algorithms, which often boils down to storing and evaluating lower order statistics (e.g., variance in path-tracing). In general, estimating the visual error without a reference is a difficult and ill-posed problem, which may easily become more demanding than the rendering process itself. Another issue is that we may not always have access to the renderer's source code or that we simply have not enough understanding of the underlying problem and in particular how to quantify the visible error. This could be because the rendering problem is hard to analyze or there are many hidden factors that have a large impact on the final, perceived rendering quality, like for example the shape of the geometry, local or global lighting distribution, scene material, rendering parameters, or even visual masking effects. All these thoughts have led to our data-driven non-reference quality metric (NoRM).

5.1. Image Data Collection

The problem of understanding and classifying rendering artifacts in general is too complex to be tackled analytically and we have chosen a data-driven approach that relies on machine learning. Since the space of artifacts even for one specific type is high-dimensional, we need many images with "right" and "wrong" examples to train a classifier initially. In general, while generating "clean" reference images may be time-consuming, producing various kind of artifacts in the rendered images is often trivial. Hence, we generated a database of rendered images with positive and negative example-images for each type of artifact (see some examples in Fig. 1). In contrast to image datasets used in computer vision tasks, our database comprises:

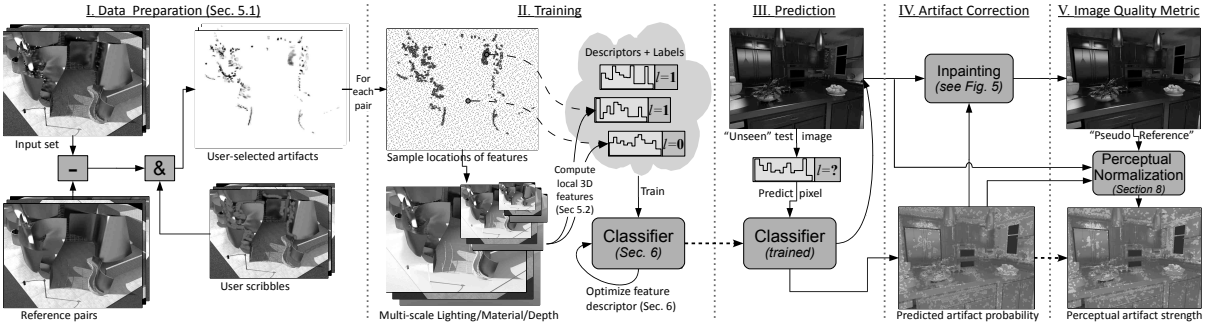


Figure 2: *Overview of the whole NoRM pipeline. Labels are semi-automatically extracted by differencing and thresholding the image with its reference and then masking this residual with a coarse user mask. For training the classifier, these labels are uniformly sampled with equal number of positive/negative samples at which we compute our multi-scale 3D features. The resulting high-dimensional descriptors are fed to the classifier (SVM). After optimizing parameters and feature dimension reduction via cross-validation, the classifier can predict artifacts in a new image. For artifact-prone pixels we inpaint reference patches from the same training image pairs to generate a pseudo-reference that is finally used in our perceptual normalization.*

- color image with reference,
- depth buffer,
- diffuse material buffer (textures),

which we refer to as a *frame* (see Fig. 3 for an example).

The reason why we restrict ourselves to this data – although we could in principle extract more – is that this data is relatively easy to dump and requires only little modification, if at all, of the rendering software. Specifically, these buffers are commonly stored in a *deferred renderer*. † Given a frame we generate other useful data, which we need for computing feature descriptors: screen-space normals from linearized depth, and approximate lighting (irradiance) using the color and material buffer.

In order to focus on artifacts which are above the threshold of visibility (and also on one specific type of artifacts) during the learning stage, initially, a coarse mask is manually painted over the tone mapped image. In the masked regions we compute the error between the pixels in the reference and the artifact-image via differencing. This way the user only needs to provide a rough mask in which we label those pixels for which the error really exists, see Fig. 2. Finally, we avoid that a few pixels are *not* assigned artifact labels (e.g., due to zero-crossings in the error signal) although neighboring pixels would indicate so. Therefore, we perform an additional dilation plus erosion (morphological closing) on the labels with a disc of pixel radius 2.

5.2. Features for Classification

Finding good descriptors or a combination of descriptors that discriminate the feature space well is crucial for any machine learning approach. We experimented with various standard techniques to classify and discriminate artifacts from the remaining "correct" pixels. Those feature descriptors comprised local histograms of color and depth, HoG (histogram of oriented gradients), multi-scale Hessian, and frequency domain descriptors based on discrete cosine transform (DCT). We applied these descriptors to compute features for our depth, color and material buffer pixels. It turned out that none of those techniques was discriminative enough to give satisfying results and we had to dig more into the rendering process itself exploiting scene and rendering-specific knowledge, which we will describe further.

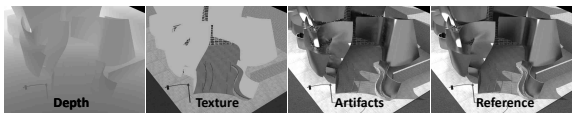


Figure 3: *Example of the data used as input for training.*

5.2.1. Texture Removal

Instead of computing features in the material buffer and increasing the dimension of the feature space, we partially remove the correlation of pixel color and texture to obtain the approximate lighting (compare the left image in Fig. 4 with Fig. 3). We only restrict ourselves to diffuse textures since these usually convey the most information about material structure in a synthetic scene. For diffuse surfaces in a scene this provides us with information about irradiance instead of pixel radiance, which is locally low-dimensional. Since the color buffer is given as HDR image, we simply divide the color pixels by the corresponding linearized material pixels. Care has to be taken with material values clamped to zero where the original lighting information in one or more color channels is essentially lost. For such rare cases we diffuse the lighting in the clamped color channels from spatially

† The depth buffer is always present in a rasterizer and the material buffer could be obtained by rendering the scene shot with only ambient lighting or using a simple "*eye-light* shader".

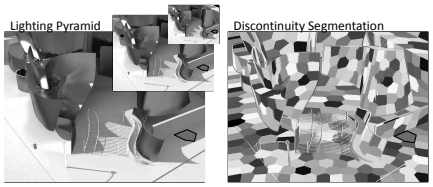


Figure 4: *Computing statistics of lighting in a local, contiguous neighborhood (red patch) at different scales.*

neighboring pixels, that are not affected by clamping. Entirely black material pixels are considered as light sources or specular surfaces and the corresponding color pixels are not altered. These heuristics worked well for our image database but are certainly not always satisfactory when dealing with complex glossy materials possibly consisting of several layered textures. For such a scenario the user can still provide the lighting image instead of relying on an ill-posed deconvolution of BRDF and lighting.

5.2.2. Screen-space ambient occlusion

Screen-space ambient occlusion (SSAO) has been developed for the GPU to efficiently compute an approximate scalar ambient-occlusion term $s_{ao}(x)$ solely based on the depth buffer. Essentially, ambient-occlusion computes the solid-angle covered by the non-occluded environment (far field) in the visible hemisphere of directions. Although SSAO is a crude approximation in screen-space, it can deliver good results for pixels where the surrounding occluders are all visible in the depth buffer. Ambient occlusion is highly correlated with the harmonic mean distance to the surrounding surfaces, which is often taken as an upper bound for the irradiance gradient of the indirect lighting [WRC88]. Since many lighting artifacts are due to large indirect gradients, the complement, $1 - s_{ao}(x)$, is also a good indicator for potential artifacts.

5.2.3. Rectified Tiles – Descriptors in Texture-space

In contrast to computer vision approaches, the presence of exact depth per pixel allows us to "unfold" a local image region from the surface captured by the depth buffer and transform it to its canonical view. This way, we are able to preserve depth discontinuities and perspective when computing local feature descriptors (e.g., histograms) and essentially reduce the dimensionality of the classification problem since we can operate in 2D texture space[‡]. For computing the local texture parametrization of the decals we use *discrete exponential maps* [SGW06] computed over the depth buffer, which is based on Dijkstra's graph-distance algorithm. An example of the computed decal parametrization is shown in Fig. 5. The so computed texture parametrization gives us the mapping from 2D texture space to projected 2D image space but we need the inverse mapping. Instead of directly "unwrapping" the surface colors via a splatting approach, we first compute the inverse texture mapping (the displacement field) via *splatting* and then use this (smooth) vector field for *gathering* the surface colors [Sze10]. Since splatting may lead to holes for overly stretched pixels, we fill the "deformation vector field" using a push-pull approach [GGSC96]. This two-stage approach better preserves high-frequencies in the colors and introduces only a small amount of blur due to (bilinear) resampling when gathering the color via the inverse texture mapping. To this end, we use the computed parametrization for computing histograms of oriented gradients (HoG) directly in texture space but also for the inpainting described in Section 7.

[‡] For 30.000 randomly extracted 16×16 pixel blocks from the glossy VPL images we run PCA on rectified, non-rectified blocks and captured 99.5% of the variance in 12 (10 for shadow map aliasing), 16 basis vectors out of 256, respectively.

5.2.4. Joint-Bilateral Filtering

To detect high-frequency artifacts in the image we perform frequency analysis. To eliminate the influence of edges and discontinuities in the depth buffer we blur the image with a joint-bilateral filter with weights steered by the depth and surface normal differences of the pixels under the filter footprint. The Gaussian variance of the depth and normal filter is automatically estimated from the 80-th and 91-th percentile of the depth and normal angle histogram, respectively. Next, we compute the residual as the filtered lighting subtracted from the original lighting image. For each feature sample we perform a local discrete cosine transform (8×8 DCT) in the residual image within a weighted Gaussian window (Gabor filter) at 2 different scales in a pyramid.

5.2.5. Local Statistics

Artifact image regions have different color distributions than the reference counterpart and we compute the first four central moments (mean, variance, skewness, kurtosis) in a local window of 16×16 pixels in 3 different scales. Similar to the joint-bilateral filtering, we only compute the statistics in a window over pixels that correspond to a contiguous surface in the depth buffer. In order to do so, we segment the depth buffer in piecewise continuous image segments via k-means clustering of pixels with respect to depth and surface normal (see Fig. 4).

6. Classification and Feature Optimization

We have proposed and tested several standard as well as specialized features described above. However, many of those features are not useful for our task or might be redundant. Certainly, using too many features, it is likely that the model overfits the training data and that we cannot provide enough examples to train the classifier efficiently. Hence, we have to select a subset from our feature pool such that the combined

feature is the most discriminative with respect to our artifact type.

Before the optimization, we linearly rescale the extracted feature vectors such that the 5th percentile of all data points maps to 0 and the 95th percentile to 1. This way, we estimate the feature bounds only from the "inlier" training data, i.e., we only account for samples within a standard deviation $\sigma \approx 1.5$ if we assume data samples are Gaussian distributed. Then, similar to [LSAR10], we use a greedy approach to select the "best" feature subset. The idea is to select the feature, one at a time, that minimizes the cross-validation error measure (*BER* see below) computed over the training set and add this feature to the current best feature set, which is initially empty. This procedure is continued till the cross-validation error of the classifier is increased when adding more features (i.e., increasing the feature dimension). The resulting features after this optimization starting with the entire feature pool are listed in Tab. 1. We use 10-fold cross-validation over the feature descriptors (computed from subimages) and split the randomly permuted training features into 90% training and 10% validation data and perform 5 iterations (i.e., evaluate the feature performance on half of the data set).

For the classifier we use a support vector machine (SVM) with a radial basis function (RBF) kernel. The two main parameters of the SVM, the regularization parameter C, and the RBF kernel width γ, are automatically computed by again minimizing the cross-validation error in a hierarchical manner (coarse-to-fine grid search). The best parameters for each type of artifact can be found in Table 1.

When testing the classifier on new (unseen) images, which may contain much fewer artifacts than non-artifact pixels, the classification may result in high recognition rates (>90%) even when every pixel is classified as "non-artifact". To get a more sensible error measure, we chose the balanced error rate (BER): $BER = \frac{1}{2}\left(\frac{|\{i \mid l(i)\in\Omega_+ \wedge p(i)\neq l(i)\}|}{|\Omega_+|} + \frac{|\{i \mid l(i)\in\Omega_- \wedge p(i)\neq l(i)\}|}{|\Omega_-|}\right)$, where $l(i)$ is the correct label, $p(i)$ is the predicted label of sample i and Ω_+, Ω_- is the set of positive, and negative labeled samples.

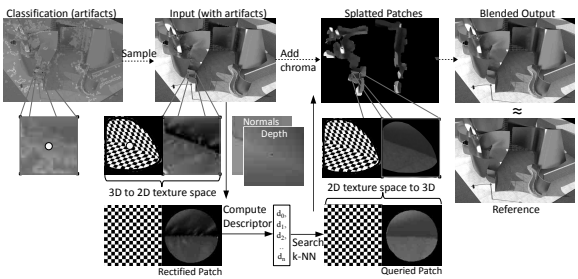


Figure 5: *Outline of the proposed inpainting algorithm (without the material decorrelation) illustrated for one patch including texture parametrization computed as described in Section 5.2.3.*

7. Artifact Correction via Inpainting

Certainly, detecting artifacts is appealing, but we would also like to remove the artifacts to obtain a higher quality, acceptable image, which we can utilize as a "pseudo-reference", see Section 8. In cases where the artifacts are minor and cover only a small fraction of the image, this is possible. We already described how to compute the likelihood of pixels to be prone to artifacts of certain types using SVM classification. An obvious approach to artifact elimination is to perform regression and learn the error function of the artifact training images. However, our tests using support vector regression were unsatisfactory, perhaps because the error function is often too noisy. Hence, we chose an approach based on context-sensitive inpainting.

First of all, during the correction phase we only touch those pixels that are classified as artifacts with a certain minimum strength. The main idea is to inpaint tiny images seamlessly into the detected artifact regions that match the local configuration of this region. Again, we exploit the additional information in the depth and material buffers to facilitate the inpainting process. First, we only inpaint rectified images that live in texture space, which we glue onto the contiguous surface as described in Section 5.2.3. Second, we remove the textures from the image before the inpainting process (see Section 5.2.1). Nevertheless, the inpainting procedure must still be able to preserve high-frequency edges (e.g., caustics, shadows) and must also hide the transition at the inpainting boundary. The later is achieved via linear blending of the splatting result with the original image, where the blending weights are computed from the binary artifact labels (red pixels in Fig. 5), which we blur with a Gaussian ($\sigma = 3$) after dilating them by a quarter of the patch size (i.e., 4 pixels). We also experimented with Poisson image blending [PGB03] but it produced sometimes unrealistically looking color bleedings.

Now, we need to find artifact-free image blocks to be painted into the local artifact region. Our inpainting operates in LDR *YCbCr* color space and we only inpaint tone mapped luminance (Y) while chroma ($CbCr$) is copied from a filtered version of the artifact image. We use a joint-bilateral filter as described in Section 5.2.4. For each artifact pixel a local image block (16×16 pixel) is extracted and rectified, which is then used to construct an index to query a database for the k-nearest neighbors (k-nn). This database is initially generated from our training image pairs and contains tens of thousands of rectified reference lighting patches together with the artifact descriptor index. As a descriptor we use the downsampled luminance (8×8) of the rectified artifact patch multiplied with a Gaussian envelope to penalize off-center pixels. In order to detect also large scale patterns, we use a multi-scale search and extract image blocks from the first l levels ($l = 2$) of a Gaussian pyramid. Therefore, the k retrieved reference patches are first upsampled (bicubic) to the corresponding scale of the search descriptor, then cropped

to our patch resolution, and blended according to their k-nn distance norm (L_1) before being warped to image space using the computed texture parametrization. Finally, after all pixels are sampled, the material is added back and the image is blended with the original image as described previously. The main algorithm steps are illustrated in Fig. 5.

8. Perceptual Normalization of Image Contrast

While a binary metric that detects the presence or absence of common rendering artifacts is useful, for most practical purposes it is also desirable to predict the perceived strength of these artifacts. The prediction of the perceived strength of artifacts in no-reference metrics involves additional challenges due to the absence of a reference image (I_{ref}). At a conceptual level, full-reference metrics often assume that the evaluated test image I_{tst} is simply I_{ref} plus some distortions D, and thus, D can be obtained by $I_{ref} - I_{dst}$. Without I_{ref}, obtaining D from I_{dst} is not trivial. To that end, we take advantage of the observation that the rendering artifacts we consider are of medium to high frequency, and approximate I_{ref} via inpainting I_{tst} (Section 7).

Given a rendered image, we employ a multiscale luminance contrast perception model [MDK08] to compute the hypothetical supra-threshold HVS response. The outcome of this computation is perceptually linearized local contrast of the input image. To do so, we first compute a 6-level Laplacian pyramid of image luminance L. Then, a Wilson's transducer [Wil80] function T is applied at each pyramid level L_k. The transducer function operates on HVS-referred values which take human spatial contrast and adaptation luminance sensitivity into account. The luminance adaptation map is approximated by the low-pass residue of the Laplacian pyramid.

The process above is repeated separately for I_{ref} and I_{tst}: given the luminance differences L_k and HVS sensitivities S, the transducers non-linearity models the contrast self-masking properties of the visual system at each pyramid level k. The differences of HVS responses scaled in Just Noticeable Difference (JND) units are then combined using a Minkowski summation with exponent 2. Formulae and implementation details are summarized in the supplementary material.

9. Results

We have tested our method on a set of 24 images generated from several 3D scenes (subset shown in Fig. 1), composed of 6 images containing glossy VPL artifacts, 12 images with shadow map artifacts, and 6 with VPL clamping artifacts. These images were rendered with different software: a GPU-based deferred renderer, an instant radiosity (VPL) renderer, a pathtracer and lightcuts [WFA*05] implementation, for producing the shadow map artifacts, glossy

Figure 6: *Additional depth and material information improves the artifact detection significantly (4th image) compared to pure image-based classification (3rd image).*

VPL noise and clamping bias, and reference images, respectively. Each image has a corresponding depthbuffer, diffuse material buffer and a reference image. This small number of images may seem too low compared with pure image classification. However, remember that we train the classifier only on small multi-resolution subimages, which are also of low-dimension due to our decorrelation with geometry and material.

For training each artifact we extracted approximately 15.000 randomly sampled subimages (50% positive and 50% negative samples) from all images excluding the one for present testing. The most discriminative features we have found (which are also shown in Table 1) are: SSAO (Section 5.2.2), rectified depth histograms of oriented gradients (HoG) (see Section 5.2.3), rectified light HoG (i.e., color without textures as described in Section 5.2.1), multi-scale light, depth, and material statistics (i.e., variance, skewness, kurtosis, Section 5.2.5), and frequency analysis of the difference of bilateral-filtered images (bilateral DCT) (Section 5.2.4). SSAO is very effective in detecting clamping bias but only in combination with other features since isolated, it always predicts clamping even in reference images. The most important feature overall is the (rectified) HoG for color, which also slightly outperformed the bilateral DCT feature. In general, having more information behind the pixels clearly improves classification as shown in Fig. 6. We tested our descriptors on SVMs and approximate k-nearest neighbor (k-nn) classifiers (with 5 k-nn). The difference is quite diverse. For the shadow artifacts SVM clearly outperformed k-nn (approx. 10% smaller error) whereas for relatively fuzzy artifacts, clamping and VPL noise, both methods performed similar. Therefore, in our results we only provide results for SVM.

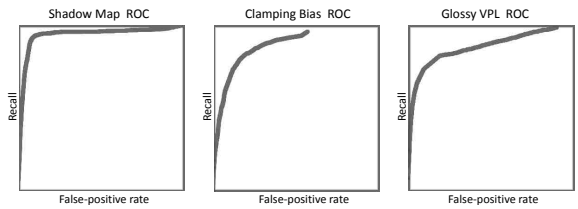


Figure 7: *Mean ROC curves for the shadow map (left), VPL clamping (center), and glossy VPL artifacts (right).*

A visualization of our detected artifacts versus ground truth user annotations is shown in Fig. 8. Further, numerical results and statistics can be found in Table 1 and in

the average *receiver operating charateristics* (ROC) curves (Fig. 7) for the different artifact classes. The classification works best for the shadow mapping artifacts. This is not surprising as shadow aliasing has usually a distinctive regular structure and high contrast, whereas the VPL clamping bias and glossy noise is difficult to address locally and without the global scene knowledge might be mistaken as shadow or highlights, respectively. Moreover, the initial user labeling is very subjective and any mistake (wrongly marked or missing artifact label) confuses the classifier rendering the problem much more complex and noisy, which also shows the downside of a data-driven approach. However, we highlight (in Sections 8,10) how to transform the initial noisy classification into a perceptualized output in form of a distortion map, which is comparable with reference-based VDPs. Besides, we should stress that some of our training examples (clamping bias) exhibit only subtle artifacts, which were even confused by human subjects in our user study.

Class	Features	SVC (C,γ)	Img. #	Acc. [%]	1-BER [%]
Shadow	Light-HoG-$16\times4\times4$, Light Bilateral DCT, Depth (Skew)	$31.1, 0.036$	#5	95.7	89.5
		$31.1, 0.036$	#6	96.2	84.7
		$31.1, 0.036$	#7	91.2	69.2
		$31.1, 0.036$	#8	86.6	90.4
Clamping	SSAO, Depth-HoG-$16\times2\times2$, Light-HoG-$16\times3\times3$, Light (Skew), Mat. (Var, Kurt)	$10.3, 0.027$	#9	92.0	74.2
		$10.3, 0.027$	#10	91.6	58.8
VPL noise	SSAO, Depth-HoG-$16\times2\times2$, Light-HoG-$16\times3\times3$, Light: (Var, Kurt), Depth: (Var, Skew), Mat.: (Var, Skew)	$9.2, 0.02$	#1	91.2	65.6
		$9.2, 0.02$	#2	85.0	68.2
		$9.2, 0.02$	#3	95.0	89.9
		$9.2, 0.02$	#4	75.8	71.6

Table 1: *The classification accuracy (Acc) and balanced error rate (BER) for different artifacts together with classification parameters (SVC) and the optimized feature set for each artifact type. The 3 dimensions of the HoG feature define the angular, spatial-X, and spatial-Y resolution of the histogram, respectively. The statistics over local light, depth, material (Mat.) regions are Variance (Var), Skewness (Skew), Kurtosis (Kurt). Corresponding predictions are shown in Fig. 8.*

The inpainting procedure works well for diffuse surfaces and even better for textured surfaces, which mask small inpainting errors (see Fig. 8, 2 last rows). On glossy surfaces with smooth low-frequency gradients and color bleedings inpainting seams may become visible but the overall quality is still improved. In particular, the shadow map artifacts are easy to cure and are perceptually hard to distinguish from the reference. However, there are also a few challenges. First, there is a tradeoff between patch size and reconstruction quality. If the patches are too small the artifact structure might be overlooked (e.g., for the shadow map aliasing we need a larger window to recognize the "jaggy" structure of the edge), whereas too large patches quickly increase the search space (curse of dimensionality) and produce overly blurred results. Besides, the larger the variety of image patches in the database, the better is the resulting inpainting quality. Currently, we extract in total around 50.000 16×16 patch pairs from the first 2 pyramid levels of the reference and artifact images. The patches also compress well and any dimension reduction (e.g., via PCA) would further speed up the inpainting and reduce the memory footprint considerably. Such improvements we leave as future work.

In general, the results of the inpainting procedure are subjectively better than the original distorted images, but they are not perfect and may still exhibit perceivable differences to artifact-free images. However, the main purpose of the inpainting step is to generate a pseudo-reference that makes perceptual normalization possible resulting in clearly improved quality of the distortion maps, as one may see in Fig. 8 (row 8) as well as in correlation values (Table 2).

10. User Study

We performed a subjective user study to validate the prediction performance of our metric. To our best knowledge this was the first attempt to subjectively label locations of visual artifacts caused by rendering techniques both in with- and without- the reference setups. Furthermore, we performed the comparison of existing full-reference metrics, which were not validated for the detection of rendering artifacts before. In this section, we summarize the obtained results, please refer to the supplementary material for a detailed discussion of the user study.

In the experiment, we displayed the set of 10 rendered test images (see Fig. 1) on a calibrated monitor to a group of 20 observers (15M/5F, aged 21–38, all of whom had normal or corrected vision). The observers were asked to mark the perceived artifact regions using a custom scribbling application. We performed two experiments: *with-the-reference*, where an image exhibiting rendering artifacts was presented along with the reference image; and *without-the-reference*, where subjects saw only the distorted image.

The marked regions for each trial were stored as distortion maps, which were then averaged over all subjects to find the mean subjective response. Next, the metric prediction for the corresponding stimulus was computed. Besides our proposed metric, we involved two full-reference metrics in the evaluation. Results of the experiment are visually summarized in Fig. 8. For the numerical analysis, we computed the 2D correlation between the mean subjective response and the metric prediction (for each test image and each experiment separately), as shown in Table 2.

Interestingly, the subjective distortion maps show apparent agreement between the artifact perception experiments in the presence and absence of the reference, which is corroborated by high correlation values (second column in Table 2). The exceptions are images 9 and 10, where the perceptual strength of clamping bias artifacts is rather low. The subjects are seemingly able to mark strong artifacts quite accurately without seeing the reference, while for perceptually weak artifacts, the reference is needed.

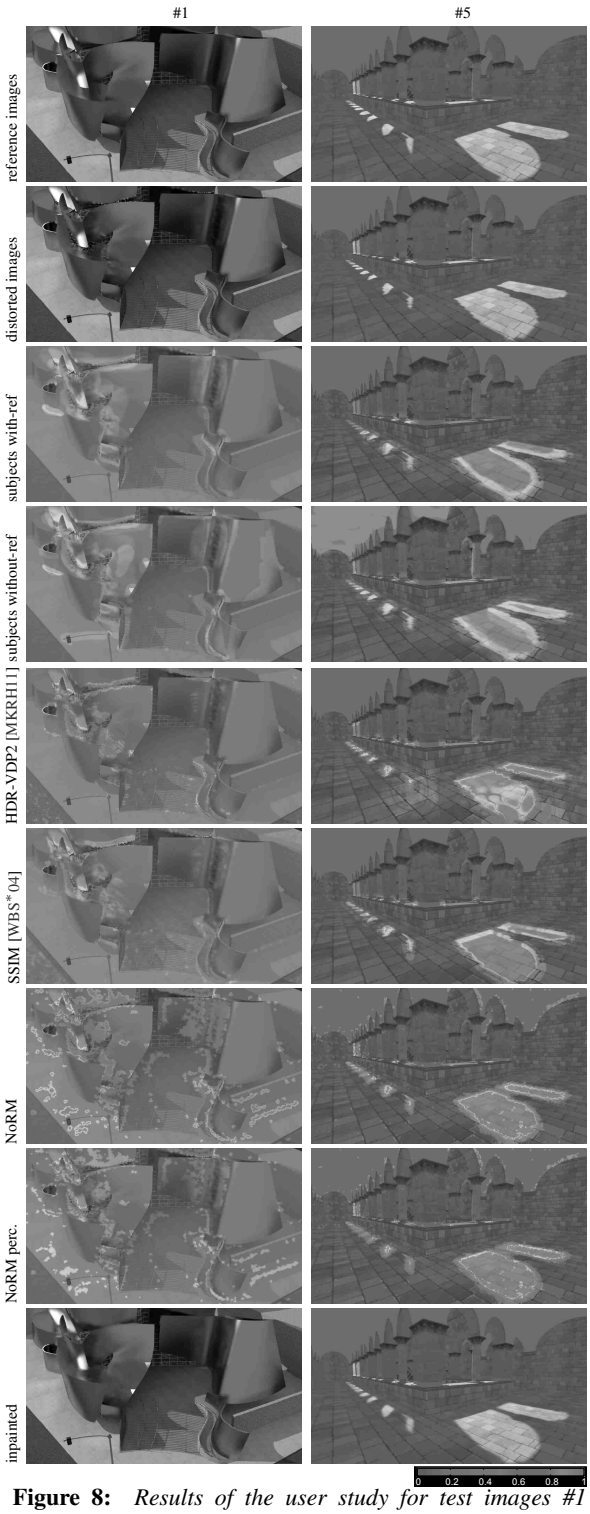


Figure 8: *Results of the user study for test images #1 and #5: average subjective artifact strenghts, and the comparison to predictions of current state-of-the art full reference metrics as well as the proposed no-reference technique. (Please refer to the supplementary material for all the images.)*

Distortion maps produced by classifier (NoRM) are binary, meaning the presence or absence of an artifact. These distortion maps sometimes tend to show too many locations, which may be correct, but the artifact severity is in reality obviously not uniform. However, thanks to the inpainting procedure, we are able to perform the perceptual normalization step (Section 8), which makes the strength of detected artifacts substantially closer to average subjective distortion maps. The prediction after the perceptual normalization (NoRMperc.) is a continuous supra-threshold distortion map calibrated in JND (just noticeable differences) units.

We compared the predictions of the proposed no-reference metric NoRM, with the state-of-the-art full-reference metrics HDR-VDP2 [MKRH11] and SSIM [WBS*04]. Neither HDR-VDP2 nor SSIM were designed or calibrated to predict the strength of rendering artifacts, but the distortion maps they produce are quite plausible. According to average correlations to the subjective ground truth distortion maps, SSIM slightly outperforms HDR-VDP2 (0.56 vs 0.535). The result of our metric (0.534) is qualitatively quite similar, making it competitive with current full-reference metrics in the targeted application. Finally, the perceptual normalization step makes predictions of NoRMperc. even closer to the experimental ground truth, resulting in the highest average correlation (0.586).

Image #	subj. no-ref.	HDR-VDP2	SSIM	NoRM	NoRM perc.
1	0.903	**0.725**	0.674	0.628	0.662
2	0.908	0.579	0.538	0.558	**0.590**
3	0.828	**0.778**	0.643	0.682	0.727
4	0.913	**0.495**	0.469	0.298	0.436
5	0.769	0.542	0.602	0.677	**0.748**
6	0.772	0.669	0.742	0.638	**0.767**
7	0.857	0.390	0.374	0.383	**0.479**
8	0.805	0.618	**0.692**	0.607	0.657
9	0.510	**0.418**	0.231	0.416	0.320
10	0.186	0.134	**0.637**	0.450	0.470
Average	0.745	0.535	0.560	0.534	**0.586**

Table 2: *Correlations of subjective responses in* with-the-reference *experiment with subjective responses in* no-reference experiment *and with the predictions of HDR-VDP2, SSIM, NoRM and NoRM after the perceptual normalization. The last row shows the average correlations over the test set. The best correlations (excluding the no-reference subjective experiment) for each stimulus are printed in bold.*

11. Conclusions and Future Work

In this paper, we proposed a novel learning based *no-reference* image quality metric for computer-generated images, which, as shown in our user study is competitive in performance with state-of-the-art visual difference predictors that *do* require a reference. Our work enables detecting and partially removing rendering artifacts. An important result of this work is that the depth and partial material information used in conjunction with color data drastically improves the classification, and even the inpainting procedure (see Fig. 6). We also present the first comparative subjective study of quality metrics on synthetic images.

Exploring further sources of information as well as classification techniques is a natural future direction. Also, a more challenging problem is quality assessment of the images with multiple types of artifacts. In the future we would like to investigate the classification of combined artifacts using a multi-class classifier and imposing a smoothness prior on the classified labels which could be facilitated by adopting e.g., Markov random fields using *belief propagation* in order to spatially smooth the labels while incorporating the correlations between different artifacts.

Acknowledgements

Many thanks to Tomáš Davidovič for VPL renderings, Mario Fritz, Chuong Nguyen, and Rafał Mantiuk for fruitful discussions, and to anonymous reviewers for suggestions on readability improvements. This work was partly supported by COST Action IC1005.

References

[BLL*10] BERGER K., LIPSKI C., LINZ C., SELLENT A., MAGNOR M.: A ghosting artifact detector for interpolated image quality assessment. In *Proc. IEEE International Symposium on Consumer Electronics (ISCE)* (2010), pp. 52–57.

[BM98] BOLIN M., MEYER G.: A perceptually based adaptive sampling algorithm. In *Proc. SIGGRAPH* (1998), pp. 299–309.

[CCB11] CHEN C., CHEN W., BLOOM J. A.: A universal reference-free blurriness measure. In *SPIE vol. 7867* (2011).

[DF04] DALY S., FENG X.: Decontouring: Prevention and removal of false contour artifacts. In *Proc. of Human Vision and Electronic Imaging IX* (2004), SPIE, vol. 5292, pp. 130–149.

[GGSC96] GORTLER S. J., GRZESZCZUK R., SZELISKI R., COHEN M. F.: The lumigraph. In *Proc. of SIGGRAPH* (1996), ACM, pp. 43–54.

[HJJ10] HACHISUKA T., JAROSZ W., JENSEN H. W.: A progressive error estimation framework for photon density estimation. In *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia)* (2010), pp. 144:1–144:12.

[KA91] KIRK D., ARVO J.: Unbiased sampling techniques for image synthesis. In *Proc. of SIGGRAPH* (1991), pp. 153–156.

[Kel97] KELLER A.: Instant radiosity. In *Proceedings of SIGGRAPH* (1997), pp. 49–56.

[KFB10] KŘIVÁNEK J., FERWERDA J. A., BALA K.: Effects of global illumination approximations on material appearance. In *ACM Transactions on Graphics (Proc. of SIGGRAPH)* (2010), pp. 112:1–112:10.

[KSGH09] KILNER J., STARCK J., GUILLEMAUT J., HILTON A.: Objective quality assessment in free-viewpoint video production. *Signal Proc.: Image Comm. 24*, 1-2 (2009), 3–16.

[LH11] LIU H., HEYNDERICKX I.: Issues in the design of a no-reference metric for perceived blur. In *SPIE vol. 7867* (2011).

[LSAR10] LIU C., SHARAN L., ADELSON E., ROSENHOLTZ R.: Exploring features in a bayesian framework for material recognition. In *23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2010), pp. 239–246.

[MB10] MOORTHY A., BOVIK A.: A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters 17*, 5 (2010), 513 –516.

[MDK08] MANTIUK R., DALY S., KEROFSKY L.: Display adaptive tone mapping. In *ACM Transactions on Graphics (Proc. of SIGGRAPH)* (2008), vol. 27(3), pp. 68:1–68:10.

[MKRH11] MANTIUK R., KIM K. J., REMPEL A. G., HEIDRICH W.: HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on Graphics (Proc. of SIGGRAPH)* (2011), 40:1–40:14.

[MRT99] MYSZKOWSKI K., ROKITA P., TAWARA T.: Perceptually-informed accelerated rendering of high quality walkthrough sequences. In *EGSR* (1999), pp. 5–18.

[PGB03] PÉREZ P., GANGNET M., BLAKE A.: Poisson image editing. *ACM Trans. on Grap. (Proc. of SIGGRAPH)* (2003), 313–318.

[RFWB07] RAMANARAYANAN G., FERWERDA J., WALTER B., BALA K.: Visual equivalence: towards a new standard for image fidelity. In *ACM Trans. on Grap. (Proc. of SIGGRAPH)* (2007), pp. 76:1–76:11.

[RPG99] RAMASUBRAMANIAN M., PATTANAIK S. N., GREENBERG D. P.: A perceptually based physical error metric for realistic image synthesis. In *Proc. SIGGRAPH* (1999), pp. 73–82.

[RSC87] REEVES W. T., SALESIN D. H., COOK R. L.: Rendering antialiased shadows with depth maps. In *Proc. of SIGGRAPH* (1987), pp. 283–291.

[RWD*10] REINHARD E., WARD G., DEBEVEC P., PATTANAIK S., HEIDRICH W., MYSZKOWSKI K.: *High Dynamic Range Imaging*. Morgan Kaufmann Publishers, 2nd edition, 2010.

[SBC05] SHEIKH H., BOVIK A., CORMACK L.: No-reference quality assessment using natural scene statistics: JPEG2000. *IEEE Trans. on Image Processing 14*, 11 (2005), 1918 –1927.

[SBC10] SAAD M., BOVIK A., CHARRIER C.: A DCT statistics-based blind image quality index. *IEEE Signal Processing Letters 17*, 6 (2010), 583 –586.

[SFWG04] STOKES W. A., FERWERDA J. A., WALTER B., GREENBERG D. P.: Perceptual illumination components: a new approach to efficient, high quality global illumination rendering. In *Proc. of ACM SIGGRAPH* (2004), pp. 742–749.

[SGW06] SCHMIDT R., GRIMM C., WYVILL B.: Interactive decal compositing with discrete exponential maps. *ACM Transactions on Graphics (Proc. of SIGGRAPH) 25*, 3 (2006), 605–613.

[Sim05] SIMONCELLI E. P.: Statistical modeling of photographic images. In *Handbook of Image and Video Processing* (2005), Bovik A. C., (Ed.), Academic Press, Inc., pp. 431–441.

[Sze10] SZELISKI R.: *Computer Vision: Algorithms and Applications*. Springer, 2010.

[TJ97] TAMSTORF R., JENSEN H. W.: Adaptive sampling and bias estimation in path tracing. In *EGSR* (1997), pp. 285–295.

[WB06] WANG Z., BOVIK A. C.: *Modern Image Quality Assessment*. Morgan & Claypool Publishers, 2006.

[WBS*04] WANG Z., BOVIK A. C., SHEIKH H. R., MEMBER S., SIMONCELLI E. P., MEMBER S.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing 13* (2004), 600–612.

[WFA*05] WALTER B., FERNANDEZ S., ARBREE A., BALA K., DONIKIAN M., GREENBERG D.: Lightcuts: A scalable approach to illumination. *ACM Transactions on Graphics (Proc. of SIGGRAPH)* (2005), 1098–1107.

[Wil80] WILSON H.: A transducer function for threshold and suprathreshold human vision. *Biological Cybernetics 38* (1980), 171–178.

[Win05] WINKLER S.: *Digital Video Quality: Vision Models and Metrics*. Wiley, 2005.

[WPG02] WALTER B., PATTANAIK S. N., GREENBERG D. P.: Using perceptual texture masking for efficient image synthesis. *Computer Graphics Forum 21*, 3 (2002), 393–399.

[WR05] WU H., RAO K.: *Digital Video Image Quality and Perceptual Coding*. CRC Press, 2005.

[WRC88] WARD G. J., RUBINSTEIN F. M., CLEAR R. D.: A ray tracing solution for diffuse interreflection. In *Proceedings of ACM SIGGRAPH* (1988), pp. 85–92.

[YPG01] YEE H., PATTANAIK S., GREENBERG D. P.: Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Transactions on Graphics 20* (2001), 39–65.

Appendix J

Learning to Predict Localized Distortions in Rendered Images

M. Čadík, R. Herzog, R. Mantiuk, R. Mantiuk, K. Myszkowski, and H. Seidel. Learning to Predict Localized Distortions in Rendered Images. *Computer Graphics Forum*, Vol. 32, No. 7, pp. 401–410, 2013.
IF=1.595

Pacific Graphics 2013
B. Levy, X. Tong, and K. Yin
(Guest Editors)

Volume 32 (2013), Number 7

Learning to Predict Localized Distortions in Rendered Images

Martin Čadík[*◇] Robert Herzog[*] Rafał Mantiuk[○] Radosław Mantiuk[▷] Karol Myszkowski[*] Hans-Peter Seidel[*]
[*]MPI Informatik Saarbrücken, Germany [○]Bangor University, United Kingdom
[◇]Brno University of Technology, Czech Republic [▷]West Pomeranian University of Technology, Poland

Abstract

In this work, we present an analysis of feature descriptors for objective image quality assessment. We explore a large space of possible features including components of existing image quality metrics as well as many traditional computer vision and statistical features. Additionally, we propose new features motivated by human perception and we analyze visual saliency maps acquired using an eye tracker in our user experiments. The discriminative power of the features is assessed by means of a machine learning framework revealing the importance of each feature for image quality assessment task. Furthermore, we propose a new data-driven full-reference image quality metric which outperforms current state-of-the-art metrics. The metric was trained on subjective ground truth data combining two publicly available datasets. For the sake of completeness we create a new testing synthetic dataset including experimentally measured subjective distortion maps. Finally, using the same machine-learning framework we optimize the parameters of popular existing metrics.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Image Quality Assessment

1. Introduction

Image quality evaluation [WB06, PH11] is one of the fundamental tasks in imaging pipelines, in which the role of synthesized images continuously increases. Modern rendering tools differ significantly in terms of the employed algorithms, e.g., global illumination techniques, which are prone to a great variability of visual artifacts [MKRH11]. Typically such artifacts are of local nature, and their visual appearance differs from more uniformly distributed image blockiness, noise, or blur that arise in compression and broadcasting applications. Existing objective image quality metrics (IQM) are specialized in predicting the level of annoyance caused by such globally present artifacts, and conform well with a single quality value, which is derived in mean opinion score (MOS) experiments with human observers [SSB06]. While some of the objective IQMs such as structural similarity index (SSIM) [WB06, Ch. 3], Sarnoff visual discrimination model (VDM) [Lub95], or the high-dynamic range visual difference predictor (HDR-VDP) [MKRH11] can locally predict perceived differences, they are not always reliable in rendering [ČHM*12]. Clearly, a need arises for novel metrics that can locally predict the visibility of numerous rendering artifacts, which are simultaneously present in a single image.

Many traditional IQMs can be modeled with a generic two-stage processing: (1) extraction of carefully designed features from the image, and (2) pooling of those features to correlate the aggregated value with subjective experiment data. At the feature extraction stage typically multi-resolution filtering with optional perceptual scaling is performed (VDM, HDR-VDP), or alternatively local pixel statistics are computed (SSIM). At the pooling stage the Minkowski summation of feature differences with respect to the reference solution (VDM, HDR-VDP), or the product of feature differences with optionally controlled non-linearity of each component (SSIM) are considered. However, such a limited feature set might not be sufficient to correctly predict the multitude of rendering-specific distortion types, especially given the variety of image content and nonuniformly distributed, mixed distortion types in a single image. Another limiting factor is the rigid form of the pooling models, which prevents the adaptation to local scene configurations and artifact constellations.

In this work, we propose a novel *data-driven full-reference metric*, which outperforms existing metrics in the

* e-mail: mcadik@mpi-inf.mpg.de, project webpage: http://www.mpii.de/resources/hdr/metric/

prediction of visible rendering artifacts. First, we systematically analyze the features used in IQMs, and then introduce a great variety of other features originating from the fields of computer vision [TM08] and natural scene statistics (NSS) [SBC05]. Additionally, we propose a few custom features including saliency data captured with an eye tracker (Section 3). We select the best suited features based on their discriminative power with respect to the rendering artifacts (Section 4). Our feature selection ensures that any distortion type we investigate is covered by a sufficiently large subset of supporting features. Instead of the feature pooling used in IQMs, we refer to machine learning solutions (Section 5), which learn an optimal mapping from the selected feature descriptors to a local quality map with respect to the perceptually measured ground-truth data [HČA*12] and [ČHM*12] (jointly referred in this paper as the LOCCG dataset for LOCalized Computer Graphics artefacts). This way our metric implicitly encapsulates highly non-linear behavior of the human visual system (HVS) that was learned from the perceptual data. To evaluate its generalization performance we also test our metric on an independent synthetic dataset, which we designed as a comprehensible tool that is suitable for evaluating other local quality metrics as well. At last, we use the same methodology to improve the performance of SSIM and HDR-VDP in rendering applications, by carefully tuning the weights associated with the features at the pooling stage (Section 6).

2. Previous Work

In this section we focus on quality metrics, which employ machine learning tools. While the metric proposed in this paper belongs to the category of *full-reference* (FR) metrics as it requires a non-distorted copy of the test image, in our discussion we refer also to *non-reference* (NR) and *reduced-reference* (RR) metrics, where data-driven approach is more common. For a more general discussion and applications of quality metrics we refer the reader to [WB06,PH11], more graphics oriented insights concerning FR metrics can be found in [MKRH11, ČHM*12].

The utility of machine learning methods in image quality evaluation has mostly been investigated for NR metrics. Typically it is assumed that the distortion type is known in advance, and then based on the correlation of its amount with human perception the image quality prediction is reported. The blind image quality index (BIQI) [MB10] introduces a distortion-type classifier to estimate the probability of distortions that are supported by the metric, and then a distortion-specific IQM is deployed to measure its amount. NSS features are employed, whose correlation with subjective quality measure for each distortion is known, and an SVM classifier is used for the quality prediction. NSS features expressed as statistics of local DCT coefficients are used in BLIINDS [SBC10], which can handle multiple distortions as well. Overall, the performance similar to the FR PSNR metric (peak signal-to-noise ratio) is reported for the LIVE dataset [SWCB06], but both BIQI and BLIINDS have trouble for JPEG and Fast Fading (FF) noise distortions. Better results have been reported in [LBW11] when instead of NSS-based features, the more perceptually relevant features: phase congruency, local information (entropy), and gradients are used. Better performance than BIQI and BLIINDS is also reported for the learning-based blind image quality measure (LBIQ) [TJK11] where complementary properties of features stem from NSS, texture and blur/noise statistics.

In RR IQM that are used in digital broadcasting a challenge is to select a representative set of features, which are extracted from an undistorted signal and transmitted along with the possibly distorted image. Redi et al. [RGHZ10] identify the *color correlograms* as suitable feature descriptors for this purpose, which enable the analysis of alterations in the color distribution as a result of distortions.

Machine learning in FR IQM remains mostly an uninvestigated area. Narwaria and Lin [NL10] propose an FR metric based on support vector regression (SVR), which uses singular vectors computed by a singular value decomposition (SVD) as features that are sensitive for structural changes in the image. Remarkably, the proposed metric shows good robustness to untrained distortions and overall outperforms SSIM.

All discussed IQMs have successfully been tested with the LIVE database [SWCB06] (and two other similar databases [NL10]), where a single value with the quality score (MOS) is available for each image. Such testing strategy precludes any conclusions concerning the accuracy of artifact localization and its visibility in the distortion map, which is the goal of this work. While the number of images available in LIVE approaches one thousand, the diversity of distortions is limited to five major classes with the emphasis on compression distortions, noise, and blur, which structurally differ significantly from rendering artifacts. For each stimulus only one distortion is present, which makes the metric performance evaluation for distortion superposition less reliable.

Machine learning solutions have been used in the context of rendered image quality assessment. Ramanarayanan et al. [RFWB07] employed an SVM classifier to predict *visual equivalence* between a pair of images with blurred or warped environment maps that are used to illuminate the scene, but problematic regions in the image cannot be identified. Herzog et al. [HČA*12] proposed a NR metric (NoRM), which is trained independently for three different rendering distortion types. The metric can produce a distortion map, and the lack of reference image is partially compensated by exploiting internal rendering data such as per pixel texture and depth. In this work we focus on solutions that are based merely on images, and can simultaneously handle more than one artifact. We utilize perceptual data derived in [HČA*12] (a part of the LOCCG dataset) to train our FR metric and we compare its performance with respect to NoRM.

3. Features for Image Quality Assessment

Many FR (i.e. the undistorted reference image needs to be available) IQMs have been developed that claim to predict localized image distortions as observed by a human [WB06]. It was shown that there exists no clear winner and each metric has its pros and cons for image distortions measured on synthetic ground-truth datasets [ČHM*12]. For better understanding of which parts of the varying metrics are important for predicting image distortions, we decompose those metrics into their individual features and analyze their strength by means of a data-driven learning framework. Moreover, we introduce new complementary features commonly used in information theory and computer vision [TM08]. Finally, we acquire saliency maps using an eye tracker and include these as a feature into our framework for the analysis of the importance of visual attention. We implemented 32 features of various kinds and origins spanning 233 dimensions which, in our opinion, is an exhaustive set (see Table 1).

3.1. Features of Traditional Image Quality Metrics

In our analysis we include features inspired by popular IQMs, including absolute difference (*ad*), SSIM [WB06, Ch. 3], HDR-VDP-2 [MKRH11], and sCIE-Lab [ZW97]. For those metric features that are only computed at a single scale (e.g., SSIM, *ad*), we additionally include their multi-scale variants. This is achieved by decomposing the feature maps into Gaussian or Laplacian pyramids (without subsampling). Despite its simplicity, *ad* (or PSNR and MSE) are still frequently used quality predictors. In contrast, SSIM measures differences using texture statistics (mean and variance) rather than pixel values. It is computed as a product of three terms:

$$SSIM(\mathbf{x},\mathbf{y}) = [lum(\mathbf{x},\mathbf{y})]^{\alpha} \cdot [con(\mathbf{x},\mathbf{y})]^{\beta} \cdot [struc(\mathbf{x},\mathbf{y})]^{\gamma}, \quad (1)$$

which are a luminance term *lum*, a contrast term *con*, and a structure term *struc* (see Fig. 8) computed for a block of pixels denoted by $\mathbf{x}$ and $\mathbf{y}$. We include each SSIM term as a separate feature: *ssim lum*, *ssim con* and *ssim struct*. Similarly, we include all frequency bands of HDR-VDP-2 differences and their logarithms (more details in Section 6.2), and denote them as *hdrvdp band* and *hdrvdp band log*.

We also introduce a few variations of the SSIM contrast components, which we found to be well correlated with subjective data. The standard contrast component is expressed as: $con(\mathbf{x},\mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2+\sigma_y^2+C_2}$, where σ_x and σ_y are the per-block variances in the test and reference images, and C_2 is a positive constant preventing division by zero. The product in the nominator introduces a strong non-linear behavior; the increase of contrast (variance) and decrease have different effect on the value of the component. Marginally better results can be achieved if the contrast difference is expressed as: $con_{bal}(\mathbf{x},\mathbf{y}) = \frac{(\sigma_x-\sigma_y)^2}{\sqrt{\sigma_x^2+\sigma_y^2+\varepsilon}}$, where ε is a small constant (0.0001). We denote this feature as *ssim con bal*. The denominators in these expressions are effectively responsible for contrast masking, which reduces sensitivity to contrast changes with increasing magnitude of the contrast. Such masking can be determined by the image of higher contrast (test or reference): $con_{balmax}(\mathbf{x},\mathbf{y}) = \frac{(\sigma_x-\sigma_y)}{\max(\sigma_x,\sigma_y)+\varepsilon}$. We denote this feature as *ssim con bal max*. Finally, we observed that individual distortions are more noticeable when isolated, rather than uniformly distributed over an image. This effect can be captured by the inhibited contrast feature (*ssim con inhibit*): $con_{inhibit}(\mathbf{x},\mathbf{y}) = \frac{con(\mathbf{x},\mathbf{y})}{\overline{con}(\mathbf{x},\mathbf{y})}$, where $\overline{con}(\mathbf{x},\mathbf{y})$ is the mean value of the contrast term in the image.

	Feature Name	Dim.	Multi scale	Import. multi-dim. (greedy)	Import. multi-dim. (stacking)	Import. scalar (dec. trees)	Import. scalar (AUC)
1	ad [Sec.3.1]	11	✓				
2	bow [Sec.3.2]	32			1.0	1.0	
3	dense-sift diff [BZM07]	1		0.72047			0.86216
4	diff [Sec.3.3]	11	✓		0.48596	0.66906	
5	diff mask [Sec.3.3]	1		0.19609			0.85772
6	global stats [Sec.3.3]	5					
7	grad dist [Sec.3.3]	1					
8	grad dist 2 [Sec.3.3]	1			0.32785	0.66382	0.85919
9	Harris corners [HS88]	12	✓			0.76699	
10	hdrvdp band [MKRH11]	6	✓			0.68933	0.85035
11	hdrvdp band log	6	✓				
12	hog9 [DT05]	62			0.46443		
13	hog9 diff [Sec.3.2]	1			0.32178	0.67821	
14	hog4 diff [Sec.3.2]	1					
15	location prior [Sec.3.4]	2					
16	lum ref [Sec.3.3]	11	✓	0.58963			
17	lum test [Sec.3.3]	11	✓	0.21429			
18	mask entropy I [Sec.3.3]	1		0.40419	0.52820	0.99389	0.86358
19	mask entropy II [Sec.3.3]	5	✓	1.0		0.67035	0.86676
20	patch frequency [Sec.3.4]	1			0.41590		
21	phase congruency [Kov99]	10	✓	0.19712			
22	phow diff [BZM07]	1					
23	plausibility [Sec.3.4]	1			0.32051		
24	sCorrel [Sec.3.3]	1		0.18956			0.8496
25	spyr dist [Sec.3.3]	1				0.85793	
26	ssim con [WBSS04]	11	✓				0.8496
27	ssim con inhibit [Sec.3.1]	1			0.44840		0.84517
28	ssim con bal [Sec.3.1]	1					
29	ssim con bal max [Sec.3.1]	1					
30	ssim lum [WBSS04]	11	✓	0.58791			
31	ssim struc [WBSS04]	11	✓	0.18681	0.53080	0.65608	0.86484
32	vis attention [Sec.3.5]	1					
Metric performance (AUC)				0.880	0.897	0.916	0.892

Table 1: *Left to right: implemented features, their dimensionality, scale selection, estimated normalized importance for best joint features, and one-dimensional sub-features (only the best sub-feature importance is reported for scalar selection methods), see Section 4. The importance of the selected features is color-coded (from blue to green to red). For each set we show the performance (area under the ROC curve for the LOCCG dataset) of a data-driven metric utilizing only the selected features (Section 5). Notice that only* ten best *features in each column are reported for clarity.*

3.2. Computer Vision Features

Much research on features comes from the field of computer vision. Therefore, we analyze popular features from com-

puter vision in the mutual spirit *"what's good for computer vision may also help human vision"* and vice versa. In particular, we consider the following features for image quality assessment: bag-of-visual-words (*bow*) [FFP05], histogram-of-oriented-gradients with 9 orientation bins (*hog9*) [DT05], the Euclidean distance between *hog9* (coarse version *hog4*), dense-SIFT [BZM07], pyramid-histogram-of-visual-words [BZM07] computed for test and reference images denoted as *hog9 diff* (*hog4 diff*), *dense-sift diff*, *phow diff*, respectively, *Harris corners* [HS88], and *phase congruency* [Kov99].

Bag-of-visual-words (*bow*) is perhaps the most commonly used feature in computer vision with a whole field of research devoted to it. Briefly, the typical *bow* feature extraction pipeline consists of two steps: first, the computation of a dictionary of visual words and second, encoding an image with a histogram by pooling the individual dictionary responses on the image. The strength (and weakness) of *bow* is that it ignores the location of sub-image parts making it invariant to global image constellation and thus requiring less training data in supervised learning.

We compute the *bow* feature on the error-residual image, i.e., difference between test and reference image. To generate the dictionary we use a set of artifact-reference image-pairs and randomly extract normalized pixel patches of size $n_p \times n_p$ pixels ($n_p = 8$) from all residual images. Then, we run *k-means clustering* on the patches using the L2-distance metric to generate $k = 200$ clusters from which we then extract a smaller dictionary ($k_d = 32$) by iteratively removing the cluster with the highest linear correlation. The remaining clusters form the visual words of the dictionary. To encode a new image-pair using our dictionary, we first compute the correlation of the error-residual image with each visual word and for each pixel we store the index of the visual word with the maximum response, which is pooled to build a histogram of k_d bins. In contrast to the traditional *bow* we do not compute one histogram for the entire image but a histogram for each pixel by pooling the responses in a local window ($4 \times n_p$ pixels) weighted by a Gaussian with $\sigma = n_p$.

3.3. Statistical Features

As shown in [ČHM*12] and [WBSS04] simple statistics may be powerful features for visual perception. We include both local and global statistics for an image. As local statistics we compute non-parametric *Spearman correlation* per 6×6 pixel block (*sCorrel*), parametric correlation is captured by the SSIM structure term (*ssim struc*), the gradient magnitude distance (*grad dist*) between test and reference image, the sum of squared distances between test and reference image decomposed in a steerable pyramid (*spyr dist*), and visual masking computed by a measure of entropy (*mask entropy I*), which is computed per 3×3 pixel block as the ratio of the entropy in the residual-image block $\mathbf{x} - \mathbf{y}$ to the entropy in the reference-image block $\mathbf{y}$:

$$H_{mask}(\mathbf{x},\mathbf{y}) = \frac{\sum_{i,j} p(x_{ij} - y_{ij}) \log_2 p(x_{ij} - y_{ij})}{\sum_{i,j} p(y_{ij}) \log_2 p(y_{ij})}, \quad (2)$$

where $p(x_{ij} - y_{ij})$, $p(y_{ij})$ is the probability of the value of pixel (i, j) in the normalized residual-, reference-image block, respectively. We also include a multi-scale version of this feature with larger window size (5×5) denoted as *mask entropy II*. For completeness we also add the luminance of the pixel in the test image (*lum test*) and reference image (*lum ref*), as well as the signed difference (*diff*) at varying image scales as individual scalar features.

In order to see whether global image distortions influence the perception of local artifacts, we add global distortion statistics to our analysis that is computed over the entire image. Specifically, we compute the mean, variance, kurtosis, skewness, and entropy of the distortions in the entire image, which are grouped into one feature class denoted as *global stats* in Table 1.

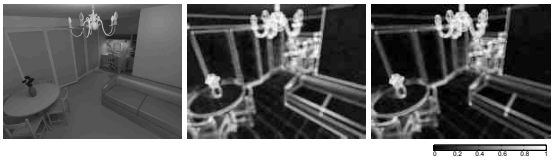


Figure 1: *Plausibility (middle) and the patch frequency feature (right) for the apartment image in LOCCG dataset. Note how repeating structures in the image (e.g., edges and texture on the floor) receive high values.*

3.4. High-level Visual Features

The features described so far are "memory-less" and only of local nature meaning that the information content is restricted to a small image region around the sample point. However, the perception of image distortions is largely dependent on the higher-level human vision following Gestalt laws and learned scene understanding. While a simulation of higher-level human vision is computationally intractable, we added a few features that mimic the global impact of local distortions on the perception of artifacts that is beyond local pixel statistics.

In the LOCCG dataset we observed that some artifacts are subjectively less severe than others depending on the likelihood that such an artifact pattern could also occur in reference images (e.g., darkening in corners). We denote such a phenomenon as *artifact plausibility*. In order to approximately model artifact plausibility we make use of a larger independent dataset of reference photos (the LIVE and Labelme datasets [SWCB06,RTMF07]) from which we sample random sub-images referred to as *patches* of size 16×16 pixels in a pre-process. Since we are mainly interested in the structural similarity of patches, we make patches contrast and brightness invariant by subtracting the mean luminance and dividing by the standard deviation. Moreover,

to make later searching efficient, we map these contrast-normalized patches to a truncated DCT basis (12 out of 255 AC-coefficients). For this pool of random patches we build an index data structure for efficiently searching the nearest neighbors. Then, for each sample point in a distorted image we extract a patch following the same steps as in the pre-process and query the k-nearest neighbor patches ($k = 16$) in this database using the L1-distance. Given the distance to the k-th nearest neighbor, we compute an estimate of the probability density for the query patch in the world of all images, which becomes a new feature denoted as *plausibility*.

Inspired by non-local means filtering, we additionally estimate the occurrence frequency of a local image patch by searching for the most similar patches contained in the same image rather than in an independent database as for the *plausibility* feature. This way, patches with common structure (e.g., edges, repeating texture) receive higher values than patches with rare patterns in the same image, see Fig. 1. This feature is denoted as *patch frequency*.

We are also interested in analyzing whether the distribution of the locations of artifacts within an image has an effect on the visibility of local artifacts. Therefore, we compute the first central moments of the artifact distribution in the image, i.e., we compute the mean, variance, kurtosis, and skewness of the artifact distance to the center of the image, which is summarized as *location prior* in Table 1.

3.5. Visual Saliency (Eye Tracking)

Another potentially important cue for the perception of local artifacts may be saliency. To estimate its importance for image quality assessment, we explicitly modeled saliency by employing an eye tracker in a user experiment. Low-resolution saliency maps were generated from the recorded gaze points per image that represent the mean visual exploration, which is stored as a feature denoted by *vis attention*. In the experiment, we were showing images from the

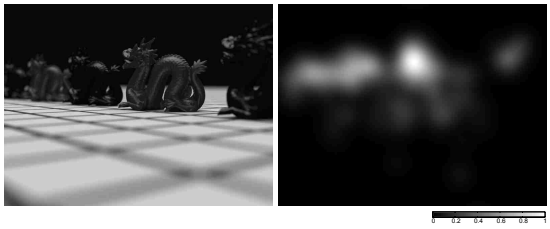


Figure 2: *The new visual attention dataset (examples for scene* dragons*). For each image from the original LOCCG dataset (left), we measure the average saliency map (right).*

LOCCG dataset to observers. The observers were asked to remember the details of the image without any top-level task. The eye tracker collected the gaze data for each image presented for 12 seconds. The answers to these questions were not analyzed and did not affect the results. We calibrated the eye tracker before each set of 5 images to increase the accuracy of the gaze estimation. The observers were asked to use the chin rest to stabilize the head position relative to the display. The experiment was conducted for 13 observers of age 20 to 43 years (12 males and 1 female).

The gaze data represents the positions of the gaze points in screen coordinates. For an individual observer we computed the fixation points based on the I-DT technique [Wid84] (with dispersion and duration equal to 100 pixels and 250 ms respectively). The fixation maps were blurred using a low-pass Gaussian filter (σ=20 pixels) to create the saliency maps called heat maps. These maps were averaged and normalized for all observers to prepare one heat map per stimulus image, see Fig. 2.

Our experimental setup consisted of a P-CR RED250 eye tracker controlled by the proprietary SMI iViewX software (version 2.5) running on a dedicated PC. The RED250 eye tracker was mounted under a 22" Dell E2210 LCD display with screen dimensions 47.5×30 cm and a native resolution of 1680×1050 pixels (60Hz). The results shown in Table 1 indicate that the measured visual saliency maps do not improve the prediction results for the LOCCG dataset. The dataset of the visual attention maps for computer graphics images, however, is interesting for future research and we make it publicly available at the project webpage.

4. Feature Selection

As the number of features we implemented is high (see Table 1), the natural questions we should answer are: first, how significant are particular features to the task of visual distortions prediction, and second, what features should be combined in a joint feature descriptor to give best generalization performance of the new IQM. Optimal feature-subset selection by exhaustive searching is computationally intractable and we experimented with different methods for feature selection where each method provides new information about the strength of individual features.

ROC Analysis One of the easiest ways to rank features is according to area-under-the-curve (AUC) values of their ROC curves [ČHM*12]. Such AUC values are shown in the last column of Table 1. The values show that the dense-sift, masking entropy, and the structural component of SSIM (*ssim struct*) provide the largest predictive power when used alone, though the differences between the best features are moderate. Although ROC analysis identifies strong features, it neither accounts for the correlation of features nor can it detect complementary features that when combined yield the best performance. For that purpose, we attempt three different feature selection strategies.

Greedy Feature Selection This procedure follows in principle the approach proposed in [LSAR10]: among the set of all possible features, we iteratively select the one that gives the smallest cross-validation error when adding it to

the pool of selected features and training a classifier on it. The process is continued until adding new features to the pool does not improve the cross-validation error. Here, for classification we use a non-linear support vector machine [CL11] with radial basis function (RBF) kernel with hyper-parameters optimized by a grid-search.

Decision Forests Another common approach for feature selection is to analyze decision trees [Bre01], which we also use for our metric described in Section 5. Ensembles of decision trees are natural candidates for feature selection [Bre01, TBRT09] since they intrinsically perform feature selection at each node of the tree. The expected frequency that a single feature is chosen for a split in a random tree and the trees impurity reduction due to the node split indicates the relative importance of that feature to the tree model [TBRT09]. This type of feature selection differs from the others in the sense that it only provides an importance weight of the scalar components of individual features.

Stacked Classifiers To this end, we also analyzed the importance of individual features by an embedded SVM classifier with L1-regularization [BM98]. To analyze the non-linear discriminative power of individual features, we build a 2-level stack of classifiers [Bre96b] where the first level consists of k non-linear classifiers (SVM) [CL11], one for each feature, that compute the artifact probability based on a single feature. These probability values are fed forward as k independent input features to the second level, which is a single linear classifier $\mathbf{w}_2 \in \mathbb{R}^k$. The classifier $\mathbf{w}_2$ is then trained on a disjoint training set using a SVM with L1-regularization, which results in a sparse vector $\mathbf{w}_2$ that can be interpreted as a joint feature importance – the higher the absolute weight $w_i = |\mathbf{w}_2(i)|,\ i \in \{1,..,k\}$ the more discriminative the ith feature. Using this procedure the average weights computed on LOCCG dataset with leave-one-out cross-validation are shown in Table 1.

4.1. Feature Selection Results

All feature selection strategies produce a reasonable feature sub-set that generalizes well when tested with leave-one-out cross-validation on trained decision tree ensembles as shown in the last row of Table 1. Although, the ROC analysis does not exploit correlation of features and selects only the best 1-dimensional features the resulting combined feature subset is still performing well. However, when comparing the feature scores (last 4 columns in Table 1) one can observe some discrepancies in the selected feature sets, which result from slightly different objectives of the methods and correlation among the features. For example decision trees can be considered as ensembles of many weak classifiers based on scalar features, whereas the greedy and the stacking approach operate on multi-dimensional features, and the ROC analysis ignores feature combinations altogether. Further, correlation among individual features can produce different sets that, when carefully observed, may actually be similar. An example are the features *hog9 diff* and *dense-sift diff*, which are highly correlated and chosen mutually exclusively by either method. Also, the signed difference (*diff*) is highly correlated with *ad* and is also a linear combination of *lum test* and *lum ref* and therefore not selected in the greedy approach but for the stacking and decision forest. Nevertheless, in agreement with the majority of the methods, the SSIM structure component (*ssim struc*), the bag-of-words (*bow*), the masking entropy (*mask. entropy I/II*), and the signed difference at multiple image scales (*diff*) can be considered as important features for our task of classifying distortions. Further, we can also rule out certain features that either do not improve performance or are simply redundant. These include absolute difference (*ad*), global image statistics (*global stats*), location of artifacts (*location prior*), and visual attention (*vis attention*). In particular, all high-level and global visual features (Section 3.4) perform rather weak in our analysis. However, this does not necessarily conclude their ineffectiveness but rather our too simplistic modeling of the complex high-level human vision.

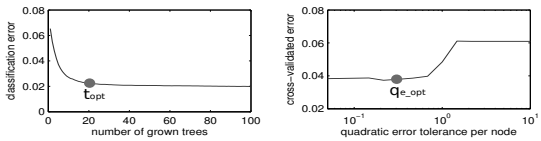


Figure 3: *Optimal parameters of the decision forest. Left: classification error versus number of decision trees t. Right: optimal splitting threshold q_e based on cross-validation.*

5. Data-Driven Metric

We experimented with different classification methods including Naive Bayes classifiers, linear and non-linear support vector machines [CL11], and decision trees [Bre01]. For our data-driven metric we obtained the best results (in terms of ROC area-under-curve) with ensembles of bagged decision trees [Bre01], which we refer to as *decision forest*. Decision forest is a powerful classification and regression tool that is scalable and known for its robustness to noise. Having constructed several random trees by bootstrapping [Bre96a], an observation is classified by traversing each tree from root node to a leaf, which contains the predicted label (artifact/no-artifact) that is averaged across all trees. The path through the tree is determined by comparing single sub-features against learned thresholds in each node. The pruned tree depth and the number of trees controls the accuracy of the classification. Using a cross-validation protocol we empirically set the number of trees to $t = 20$ and the average tree depth to 10 (implicitly controlled by a quadratic error tolerance threshold $q_e = 0.25$ for the node-splitting), which yields good generalization performance (see Fig. 3). We train our metric using the 10 best features as derived in Section 4 (shown in the last but one column of Table 1).

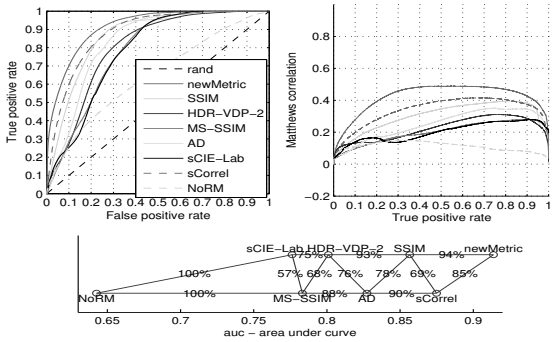


Figure 4: *Quantitative results for quality metrics on LOCCG dataset shown as ROC (top left) and Matthews correlation (top right). The bigger the area under the curve (AUC), the better.* $AUC_{newMetric}$*=0.916,* AUC_{SSIM}*=0.858,* $AUC_{HDRVDP2}$*=0.802,* AUC_{MSSSIM}*=0.786,* AUC_{AD}*=0.832,* $AUC_{sCIELab}$*= 0.783,* $AUC_{sCorrel}$*=0.880,* AUC_{NoRM}*=0.644. Bottom: ranking according to AUC (the percentages indicate how often the metric on the right results in higher AUC when the image set is randomized using a bootstrapping procedure similar to [ČHM*12]).*

5.1. Results

We train our new data-driven metric described above on the LOCCG dataset, which consists of 35 annotated image-pairs that exhibit a variety of computer graphics distortions that are difficult to predict by existing FR IQMs [ČHM*12]. Since the size of the LOCCG dataset is rather small and the images are very diverse showing (combination of) different artifacts and scenes, we do not split it into a train and test set. We instead evaluate our method in a leave-one-out cross validation fashion; i.e., we train it on $n-1$ images and test on the n-th image repeating this process n times. In addition, we validate our metric on a new uncorrelated dataset that is described in Section 5.1.1. We compare the trained metric to 7 state-of-the-art and baseline methods as shown in the quantitative analysis in Fig. 4. Our new metric outperforms all existing FR IQMs on the LOCCG dataset in terms of AUC in Fig. 4 (the higher the AUC the better). Also, the visual results agree with the ground-truth annotation as shown in the color-coded distortion maps for three images of the LOCCG dataset in Fig. 5. Please refer to the supplementary material for all results and a more detailed analysis.

For completeness, we include results of the NR metric NoRM. However, this method was not originally intended to be used for detecting general, mixed image distortions and is tuned for only specific artifacts assuming the depth maps and other cues of the scenes to be available for feature computation. Unfortunately, depth and texture maps are not available in many cases in the LOCCG dataset, and we run NoRM only with color features rendering its performance poor.

We implemented our new metric and feature computation in MATLAB for which the code is available at the project webpage. Reporting the overall computation time of the unoptimized MATLAB code, the data preprocessing and feature computation time per image (800×600) is in the order of a few minutes, the time for training the decision forest on our selected feature set based on 100.000 samples takes less than 1 minute, whereas the distortion prediction using our trained decision forest requires only ≈ 0.5 sec.

5.1.1. Results for New Synthetic Dataset

Even though we report the result for cross validation to avoid over-training, we may expect that some distortions appearing in different images are correlated and the metric just learns the distortions that are specific for that data set. To test against this possibility, we measured another dataset.

The new Contrast-Luminance-Frequency-Masking (CLFM) dataset was measured using a similar procedure as in [ČHM*12]. 13 observers provided localized markings for the visible differences in 14 image pairs. The dataset was designed to cover a wide range of problematic cases for image quality assessment in possibly few images. Such problematic cases included increments of different size and contrast, edges shown at different luminance levels, random noise patterns of different frequency and contrast, several cases of contrast masking, image pairs with pixel misalignment and noise patterns generated with a different seed for the test and reference image (see example stimuli in Fig. 6). The CLFM dataset is available at the project webpage.

Fig. 7 shows the result of the tested metrics for the new dataset. Note that our new metric was trained on the LOCCG dataset and none of the new dataset images was used for training. From the shape of the ROC curves, it is clear that the CLFM dataset is extremely challenging and the metrics mispredict in many cases. But it is interesting to notice that, on average, the proposed metric has the highest AUC value.

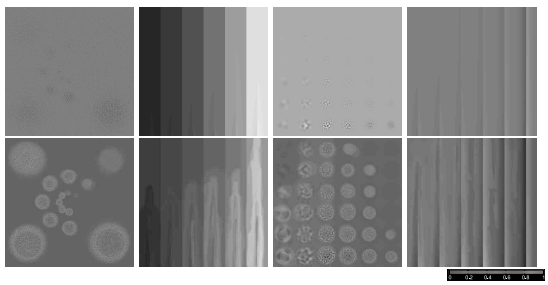


Figure 6: *CLFM: our synthetic validation dataset for testing of IQMs perceptual-masking prediction. Top row: test images containing (from left to right) increments of different size, edges at different luminance levels, and band-limited noise patterns organized in a CSF-like chart. Bottom: subjective data for the corresponding images. (Best viewed in electronic version.)*

Figure 5: *Comparison of distortion maps predicted by the proposed method with the state-of-the-art metrics for the* red kitchen, *and* sponza tree shadows *scenes. From left: subjective ground-truth, prediction of the new metric, SSIM, HDR-VDP-2, and sCorrel. Please see the complete set of results in the supplementary material.*

The performance expressed as Matthew's correlation coefficient is very steady throughout the range of true positive rates, while many other metrics exhibit significant "dips". This means that the new metric is less prone to loss of performance in the worst-case scenario.

It is encouraging to observe that learning the "real-world" distortions (e.g. based on the LOCCG dataset) may enable decent prediction performance even for the synthetic dataset like CLFM. This is different from the "traditional" approach to modeling quality metrics, where the synthetic cases are used to train the metric and the assumption is made that these will generalize for complex "real-world" cases. Interestingly, this correlates with our experience – when we used synthethic CLFM data for training, it did not lead to better predictions of LOCCG than traditional metrics.

6. Optimizing Existing Metrics

The stack of classifiers described in the last paragraph of Section 4 can be used to optimize the parameters of traditional metrics for the testing datasets. We show the results for two metrics (SSIM and HDR-VDP-2) on the LOCCG dataset as an illustration of this approach.

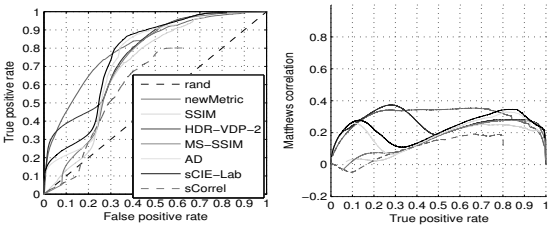


Figure 7: *Quantitative results for the new synthetic dataset (CLFM) for our metric trained on the LOCCG dataset.* $AUC_{newMetric}$*=0.805,* AUC_{SSIM}*=0.695,* $AUC_{HDRVDP2}$*=0.772,* AUC_{MSSSIM}*=0.714,* AUC_{AD}*=0.733,* $AUC_{sCIELab}$*=0.763,* $AUC_{sCorrel}$*=0.624.*

6.1. Training SSIM

The stucture similarity metric (SSIM) consists of 3 terms that were introduced in Eq. (1). The sensitivity or importance of the individual terms is controlled by the parameters α, β, and γ, which are set to 1 by default.

We optimize those 3 parameters on the LOCCG dataset with cross-validation to give the best possible prediction by employing a linear support vector machine [CL11] that computes the optimal 3D weight vector $\mathbf{w} = [\alpha^*, \beta^*, \gamma^*]^T$ for the 3 SSIM terms in the log domain $log(SSIM^*) = \alpha^* \cdot \log(l) + \beta^* \cdot \log(c) + \gamma^* \cdot \log(s) = \mathbf{w}^T \cdot \mathbf{d}^{lcs}$ by minimizing the convex objective function:

$$\arg\min_{\mathbf{w}} \sum_i \max(0, 1 - y_i \cdot \mathbf{w}^T \cdot \mathbf{d}_i^{lcs})^2 + \lambda \|\mathbf{w}\|_2^2, \qquad (3)$$

where y_i are the ground-truth labels in the dataset that are set to -1 or 1 if the distortion for the i-th training sample is visible or not, respectively, and $\mathbf{d}_i^{lcs} \in \mathbb{R}^3$ is the corresponding precomputed vector of the SSIM terms. The regularization is controlled with $\lambda = 1$.

Figure 8: *An illustration of the features of SSIM for the* sala *scene where darker pixels represent more visible distortions. From left: luminance, contrast, and structure term.*

We run this optimization 35 times with randomized set of input images to assess the stability and quality of the coefficients obtained. Interestingly, the results (Fig. 9, left) show a clear tendency towards higher weighting of the structure and contrast components than the luminance component ($\alpha = 0.2$, $\beta = 2.8$, $\gamma = 3.5$). This implies that the structural and contrast components are more important than the luminance for computer graphics artifacts, which agrees

with the results presented in Section 4. Please notice that the performance improvement of the new weighted metric ($SSIM_{learned}$) compared to the original SSIM in Fig. 10. An illustration of improvement of the distortion maps is shown in Fig. 11.

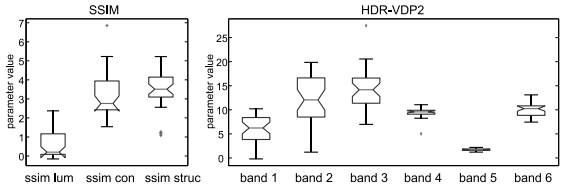


Figure 9: *The results of the optimization of SSIM (left) and HDR-VDP-2 (right) metric parameters. The red mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to extreme data points not considered outliers, and outliers are plotted individually. The notches show 5% level intervals of the median significance.*

6.2. Training HDR-VDP-2

Visible differences predictor for high-dynamic-range images (HDR-VDP-2) [MKRH11] is a perceptual metric that models low-level human vision mechanisms, such as light adaptation, spatial contrast sensitivity and contrast masking. The predicted probability of detecting differences between test and reference images is modeled as psychophysical detection task separately for each spatial frequency band. The cumulative probability is computed as probability summation, which corresponds to summing logarithms of probability values from all bands. To introduce learning component to the HDR-VDP-2, we weighted the logarithmic probabilities before summation. After learning, which used the identical method as for the SSIM (Section 6.1), we found the optimum band weights to be (in decreasing frequency): w_1=6.2, w_2=12.1, w_3=14.2, w_4=9.6, w_5=1.7, w_6=10.2 (Fig. 9, right). Please notice the significant performance gain of the new weighted metric (HDR-VDP-$2_{learned}$) compared to the original HDR-VDP-2 in Fig. 10. The improved distortion maps can be found in Fig. 11.

7. Conclusions and Future Work

In this work we proposed a novel data-driven full-reference image quality metric, which outperforms existing IQMs in detecting perceivable rendering artifacts and reporting their location in a distortion map. The key element of our metric is a carefully designed set of features, which generalize over distortion types, image content, and superposition of multiple distortions in a single image. We also propose easy to use customizations of existing metrics SSIM and HDR-VDP-2 that improve their performance in predicting rendering artifacts. Finally, as the outcome of this work two new datasets have been created, which are potentially

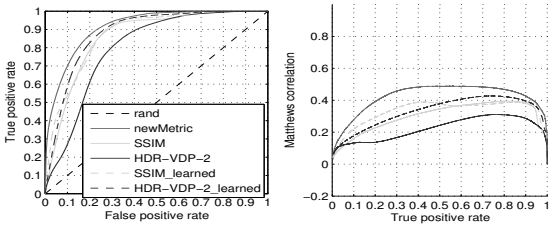


Figure 10: *Comparison of the overall results of optimized and original SSIM and HDR-VDP-2 metrics. Left: ROC, right: Matthews correlations. The bigger the area under the ROC curve (AUC), the better.* AUC_{SSIM}*=0.858,* $AUC_{SSIM_{learned}}$*=0.872,* $AUC_{HDRVDP2}$*=0.802,* $AUC_{HDRVDP2_{learned}}$*=0.883. The result of newMetric (red) is shown here for comparison.*

useful for the imaging and computer graphics communities. The Contrast-Luminance-Frequency-Masking (CLFM) dataset contains a continuous range of basic distortions encapsulated in a few images, with the distortion visibility annotated in a perceptual experiment. The distortion saliency maps captured in the eye tracking experiment could be used for further studies on visual attention, for example as a function of rendering distortion type and its magnitude.

The main limitation of our work is the size of the training dataset, and we expect that the performance of our metric can be still improved when a larger dataset is available. Furthermore, it would be interesting to explore other supervised learning techniques, e.g. [GRHS04], both for feature selection and for FR metric prediction. The eye-tracking features deserve further exploration too: for example the combination of eye-tracking data with other features like absolute difference could indicate where people gaze due to severe artifact.

Acknowledgements

This work was partially supported by the Polish Ministry of Science and Higher Education through the grant no. N N516 508539, and by the COST Action IC1005 on "HDRi: The digital capture, storage, transmission and display of real-world lighting".

References

[BM98] BRADLEY P. S., MANGASARIAN O. L.: Feature selection via concave minimization and support vector machines. In *Proc. of 13th International Conference on Machine Learning* (1998), pp. 82–90. 6

[Bre96a] BREIMAN L.: Bagging Predictors. *Machine Learning 24*, 2 (Aug. 1996), 123–140. 6

[Bre96b] BREIMAN L.: Stacked regressions. *Machine Learning 24* (1996), 49–64. 6

[Bre01] BREIMAN L.: Random forests. *Machine Learning 45*, 1 (2001), 5–32. 6

Figure 11: *An example of the improved predictions of SSIM and HDR-VDP-2 for the* sponza above tree *scene after the parameter optimization. From left: subjective ground-truth, prediction of SSIM, SSIM*$_{\text{learned}}$*, HDR-VDP-2, HDR-VDP-2*$_{\text{learned}}$.

[BZM07] BOSCH A., ZISSERMAN A., MUNOZ X.: Image classifcation using random forests and ferns. *In Proc. of ICCV* (2007), 1–8. 3, 4

[ČHM*12] ČADÍK M., HERZOG R., MANTIUK R., MYSZKOWSKI K., SEIDEL H.-P.: New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts. *ACM TOG (Proc. of SIGGRAPH 2012)* (2012). Article 147. 1, 2, 3, 4, 5, 7

[CL11] CHANG C.-C., LIN C.-J.: LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2* (2011), 27:1–27:27. 6, 8

[DT05] DALAL N., TRIGGS B.: Histograms of oriented gradients for human detection. *Proc. of IEEE Computer Vision and Pattern Recognition* (2005), 886–893. 3, 4

[FFP05] FEI-FEI L., PERONA P.: A bayesian hierarchical model for learning natural scene categories. *Proc. of IEEE Computer Vision and Pattern Recognition* (2005), 524–531. 4

[GRHS04] GOLDBERGER J., ROWEIS S., HINTON G., SALAKHUTDINOV R.: Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17* (2004), MIT Press, pp. 513–520. 9

[HČA*12] HERZOG R., ČADÍK M., AYDIN T. O., KIM K. I., MYSZKOWSKI K., SEIDEL H.-P.: NoRM: no-reference image quality metric for realistic image synthesis. *Computer Graphics Forum 31*, 2 (2012), 545–554. 2

[HS88] HARRIS C., STEPHENS M.: A combined corner and edge detector. *Proc. of the 4th Alvey Vision Conference* (1988), 147–151. 3, 4

[Kov99] KOVESI P.: Image features from phase congruency. *Videre: A Journal of Computer Vision Research 1*, 3 (1999). 3, 4

[LBW11] LI C., BOVIK A. C., WU X.: Blind image quality assessment using a general regression neural network. *IEEE Transactions on Neural Networks 22*, 5 (2011), 793–9. 2

[LSAR10] LIU C., SHARAN L., ADELSON E., ROSENHOLTZ R.: Exploring features in a bayesian framework for material recognition. *Proc. of IEEE Computer Vision and Pattern Recognition* (2010), 239–246. 5

[Lub95] LUBIN J.: *Vision Models for Target Detection and Recognition. ed. E. Peli*. World Scientific, 1995, ch. A Visual Discrimination Model for Imaging System Design and Evaluation, pp. 245–283. 1

[MB10] MOORTHY A., BOVIK A.: A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters 17*, 5 (2010), 513 –516. 2

[MKRH11] MANTIUK R., KIM K. J., REMPEL A. G., HEIDRICH W.: HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM TOG (Proc. of SIGGRAPH 2011)* (2011). Article 40. 1, 2, 3, 9

[NL10] NARWARIA M., LIN W.: Objective image quality assessment based on support vector regression. *IEEE Transactions on Neural Networks 21*, 3 (2010), 515–9. 2

[PH11] PEDERSEN M., HARDEBERG J.: Full-reference image quality metrics: Classification and evaluation. *Found. Trends. Comput. Graph. Vis. 7*, 1 (2011), 1–80. 1, 2

[RFWB07] RAMANARAYANAN G., FERWERDA J., WALTER B., BALA K.: Visual equivalence: towards a new standard for image fidelity. *ACM TOG (Proc. of SIGGRAPH 2007)* (2007). 2

[RGHZ10] REDI J. A., GASTALDO P., HEYNDERICKX I., ZUNINO R.: Color distribution information for the reduced-reference assessment of perceived image quality. *IEEE Trans. on Circuits and Systems for Video Techn. 20*, 12 (2010), 1757–69. 2

[RTMF07] RUSSELL B., TORRALBA A., MURPHY K., FREEMAN W. T.: Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision* (2007). 4

[SBC05] SHEIKH H., BOVIK A., CORMACK L.: No-reference quality assessment using natural scene statistics: JPEG2000. *IEEE Trans. on Image Processing 14*, 11 (2005), 1918–1927. 2

[SBC10] SAAD M., BOVIK A., CHARRIER C.: A DCT statistics-based blind image quality index. *IEEE Signal Processing Letters 17*, 6 (2010), 583 –586. 2

[SSB06] SHEIKH H., SABIR M., BOVIK A.: A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. on Image Processing 15*, 11 (2006), 3440–3451. 1

[SWCB06] SHEIKH H. R., WANG Z., CORMACK L., BOVIK A. C.: LIVE image quality assessment database RLS 2, 2006. 2, 4

[TBRT09] TUV E., BORISOV A., RUNGER G., TORKKOLA K.: Feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of Machine Learning Research 10* (2009), 1341–1366. 6

[TJK11] TANG H., JOSHI N., KAPOOR A.: Learning a blind measure of perceptual image quality. *Proc. of IEEE Computer Vision and Pattern Recognition* (2011), 305–312. 2

[TM08] TUYTELAARS T., MIKOLAJCZYK K.: Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis. 3*, 3 (2008), 177–280. 2, 3

[WB06] WANG Z., BOVIK A. C.: *Modern Image Quality Assessment*. Morgan & Claypool Publishers, 2006. 1, 2, 3

[WBSS04] WANG Z., BOVIK A. C., SHEIKH H. R., SIMONCELLI E. P.: Image quality assessment: From error visibility to structural similarity. *IEEE Trans. on Image Processing 13*, 4 (2004), 600–612. 3, 4

[Wid84] WIDDEL H.: Operational problems in analysing eye movements. *In A.G. Gale & F. Johnson (Eds.), Theoretical and Applied Aspects of Eye Movement Research. 1* (1984), 21–29. 5

[ZW97] ZHANG X., WANDELL B. A.: A spatial extension of CIELAB for digital color-image reproduction. *Journal of the Society for Information Display 5*, 1 (1997), 61. 3

Appendix K

Visually Significant Edges

T. O. Aydın, M. Čadík, K. Myszkowski, and H.-P. Seidel. Visually significant edges. *ACM Transactions on Applied Perception*, Vol. 7, No. 4, pp. 27:1–27:15, 2010.
IF=1.447

Visually Significant Edges

Tunç Ozan Aydın *
MPI Informatik

Martin Čadík †
MPI Informatik

Karol Myszkowski ‡
MPI Informatik

Hans-Peter Seidel §
MPI Informatik

Abstract

Numerous computer graphics methods make use of either explicitly computed strength of image edges, or an implicit edge strength definition that is integrated into their algorithms. In both cases, the end result is highly affected by the computation of edge strength. We address several shortcomings of the widely used gradient magnitude based edge strength model through the computation of a hypothetical human visual system (HVS) response at edge locations. Contrary to gradient magnitude, the resulting "visual significance" values account for various HVS mechanisms such as luminance adaptation and visual masking, and are scaled in perceptually linear units that are uniform across images. The visual significance computation is implemented in a fast multi–scale second generation wavelet framework, which we use to demonstrate the differences in image retargeting, HDR image stitching and tone mapping applications with respect to gradient magnitude model. Our results suggest that simple perceptual models provide qualitative improvements on applications utilizing edge strength at the cost of a modest computational burden.

CR Categories: I.3.3 [Computer Graphics]: Picture/Image Generation—Line and curve generation; I.4.6 [Segmentation]: Edge and Feature Detection; I.4.2 [Enhancement]: Filtering

Keywords: edge strength, visual perception, HDR

1 Introduction

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 (HVS) 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. Except perhaps simple stimuli designed for experimental purposes, visual masking is expected to occur in virtually any complex image and often to have a strong influence on perception. Disregarding the non–linear perception of luminance, especially in HDR images, often leads to overestimations in 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.

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. In this paper, we first summarize related work (Section 2), then discuss the edge awoiding decomposition framework (Sec. 3) and the HVS model (Section 4), than we validate the model (Sec. 5) 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 (Sec. 6).

2 Background

In this section we discuss related work on edge detection, computer graphics applications that utilize edge models, and HVS models for contrast perception. Due to the purely 2D nature of our technique, we do not discuss any line drawing techniques that are capable of localizing edges in a semantically meaningful way, but require 3D information about depicted objects.

Edge Detection
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

*e-mail: tunc@mpi-inf.mpg.de
†mcadik@mpi-inf.mpg.de
‡karol@mpi-inf.mpg.de
§hpseidel@mpi-inf.mpg.de

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].

Applications
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. [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 and Turk 1999]. Gradient domain operators such as [Fattal et al. 2002; Mantiuk et al. 2006], while not explicitly stated, also utilize edges since gradient magnitude operator is essentially an edge detector. Mantiuk et al.'s [2006] method has additionally a perceptual component in the form of a simple contrast transducer.

Contrast Perception
The HVS characteristics involved in contrast perception are quite complex and have been investigated in numerous psychophysical studies. Even in the simple case of *detection experiments*, where the task is to distinguish a sine wave grating from the uniform background, the resulting detection threshold 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. These characteristics are modeled by contrast sensitivity functions (CSF) [Daly 1993; Barten 1999]. Other characteristics of contrast perception are observed in the *discrimination experiments*, whose goal is to determine how the presence of one masking sine [Legge and Foley 1980] or square [Whittle 1986] grating affects the discriminability of another test grating. In some experiments, it turned out that the maskers of weak contrast actually facilitate the discriminability of test grating, and the corresponding discrimination thresholds are even smaller than the detection threshold as measured by the CSF. For high contrast (suprathreshold) maskers an elevation of discrimination thresholds can be observed. This behavior is modeled by *transducer* functions [Legge and Foley 1980; Wilson 1980; Mantiuk et al. 2006], which convert physical contrast of an image to a hypothetical HVS response. Various transducers have been successfully incorporated into the HVS models used in many computer graphics applications including texture masking simulation [Ferwerda et al. 1997], image appearance modeling [Pattanaik et al. 1998], perception-based rendering [Bolin and Meyer 1998], and tone mapping and contrast enhancement [Mantiuk et al. 2006; Mantiuk et al. 2008]. Often, transducer functions limit their modeling to intra–channel masking assuming a certain contrast patch is solely masked by other contrast patches at the same spatial frequency and orientations. A more comprehensive model by Watson and Solomon [Watson and Solomon 1997] also comprises masking from adjacent frequencies (inter–channel masking), in effect contrast patches are subject to masking from other contrast patches within a certain neighborhood. The neighborhood masking model in JPEG2000 is a simpler implementation of the same principle [Zeng et al. 2000].

3 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.

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 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. Fig. 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 1, where m and n are intensities at the current location and some neighboring pixel, respectively:

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

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).

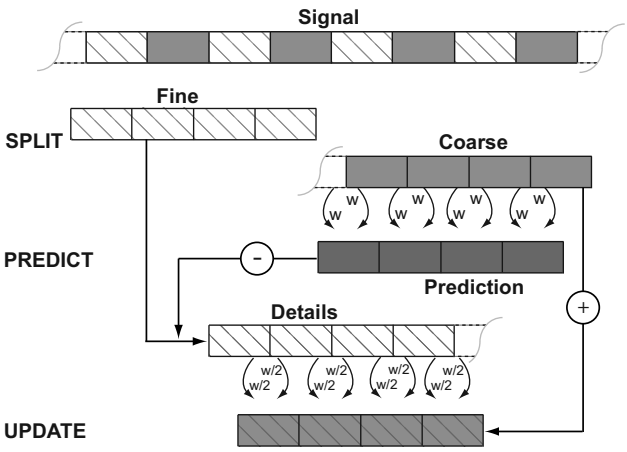


Figure 1: *Illustration of the lifting scheme on a 1D signal. The signal is decomposed into fine and coarse parts by designating odd pixels as fine, and even pixels as coarse components. The fine component is predicted from the coarse component using weights computed by the edge aware function ω, or 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 straightforward extension to the second dimension is to repeat the 1D computation at both dimensions (Fig. 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 (Fig. 2b).

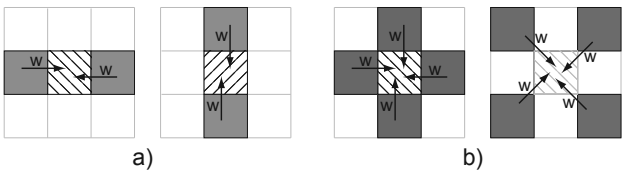


Figure 2: *The lifting scheme can be extended by repeating the 1D computation in X and Y directions (a), or using a lower anisotropy red-black quincunx lattice (b). Only the prediction step is illustrated for brevity.*

4 Human Visual System Model

We extend the EAW framework (Section 3) with an HVS model, where we modify the weighting function (Equation 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 is approximated in the EAW framework by dividing the fine samples by the local mean of the *predictions* of immediate neighbors K (2 and 4 for X-Y splitting and red-black splitting, respectively):

$$C = \frac{Fine}{(\frac{1}{K})\sum_K Prediction_k} - 1. \quad (2)$$

Repeated at each scale, this formulation is similar to the low–pass contrast in [Mantiuk et al. 2006]. The advantage of a contrast based edge strength measure over a gradient based measure is illustrated in Fig.3

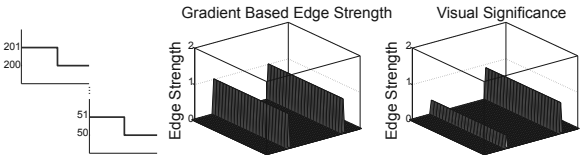


Figure 3: *Edge strength predictions utilizing physical contrast account for the effect of background luminance level. The perceived strength of step edges 200-201 cd/m^2 and 50-51 cd/m^2 (left) are predicted to be the same by the gradient based method, whereas a contrast based method correctly predicts the weaker perceived strenght 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** are its spatial frequency (ρ), and the adaptation luminance (L_a). These effects can easily be observed in the Campbell-Robson chart. We use the CSF from the Visible Differences Predictor [Daly 1993] with corrections as indicated in [Aydın et al. 2008, Equations (10, 11)] to obtain the perceptually linearized contrast $C' = C \cdot CSF(\rho, L_a)$. Fig. 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 (Fig. 4 center). The CSF's scaling results in a more uniform smoothing over edges with similar magnitude of visibility (Fig. 4 right).

Visual masking 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

Figure 4: *The effect of luminance adaptation. 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 overestimed by EAW method in the absence of a model of luminance adaptation. All images are tone mapped [Reinhard et al. 2002] for display purposes.*

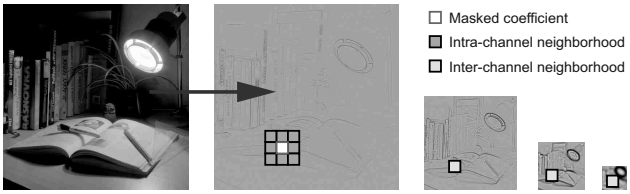


Figure 5: *An illustration of neigborhood masking on detail layers of a multi–scale decomposed image.*

channels contain information from the spatial neighborhood, a contrast patch at a certain location is effectively masked by neighboring contrast patches (See Fig. 5 for an illustration of neighborhood masking.)

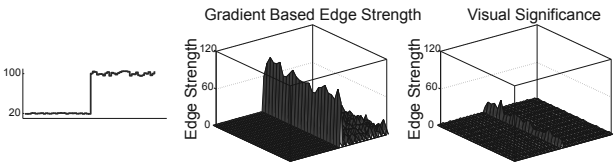


Figure 6: *The visual masking due to the random noise modulated by image luminance in the test stimulus (left), results in lower perceived edge strength then the gradient magnitude (center), as predicted by our method (right).*

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 where the visibility of edges are notably lesser than non–textured regions. To account for that, our ν function (Equation 1) comprises the point–wise extended masking model [Zeng et al. 2000] which, in addition to a compressive non–linearity, also accounts for visual masking from the local neighborhood K:

$$R = \frac{sign(C')|C'|^{0.5}}{(1 + \sum_K |C'_k|^{0.2})} \ . \tag{3}$$

The effect of visual masking on a simple stimulus is illustrated in Fig. 6. Figure 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.

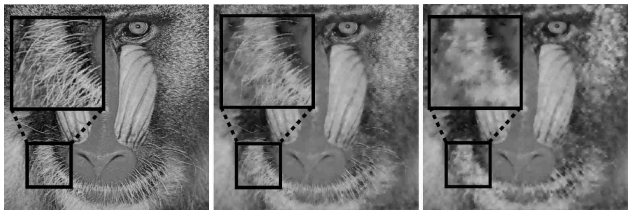


Figure 7: *The effect of contrast masking in a complex image. The original image (left), smoothing with EAW method (center), and smoothing with EAW method using visually significant edges (right). The masking model reduces the strength of the facial hair edges due to the presence of hair in the local neighborhood.*

5 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) 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 used CSF curves [Daly 1993] reflect measurements using the Michelson's definition of contrast, which is slightly different from the implemented definition contrast (Equation 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 after-images, 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 $1.5 - 400cd/m^2$.

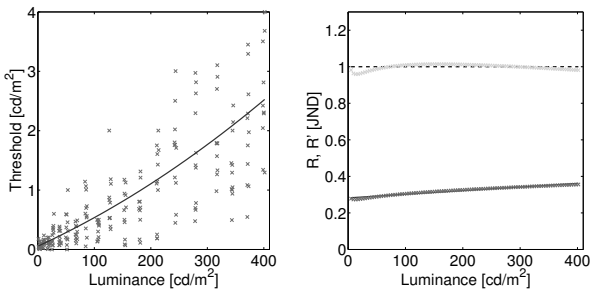


Figure 8: *Perceptual experiment. Left: measured edge detection luminance thresholds as a function of adaptation luminance L_a, right: model predictions before (red crosses) and after the calibration (green crosses). An ideal model response is constatntly 1 JND for the threshold data (dashed line).*

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 $440\mathrm{cd/m^2}$. 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. The observers were both male and female 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 Fig. 8 (left), were approximated by the second order polynomial function (blue curve). Using the polynomial function, we generated 100 input 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 Fig. 8 (right). The threshold prediction of the uncalibrated model (red crosses) 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 fuction (blue curve in Fig. 8 (right)):

$$R' = \frac{R}{0.0002\ L_a + 0.2822}, \tag{4}$$

where L_a is the adaptation luminance in $\mathrm{cd/m^2}$.

As the masking model (Equation 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.

6 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 1 as discussed in Section 3.

6.1 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 9 and 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 (Fig. 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. In images consisting of mostly bright regions ($> 100\ cd/m^2$) a simple logarithmic non-linearity may be sufficient to approximate the perception of luminance. However, this method is less precise in darker regions where Weber's law doesn't hold (compare Fig. 10 (f) and (g)). Moreover, visual masking may have a significant effect in images contating many details (Fig. 9).

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.

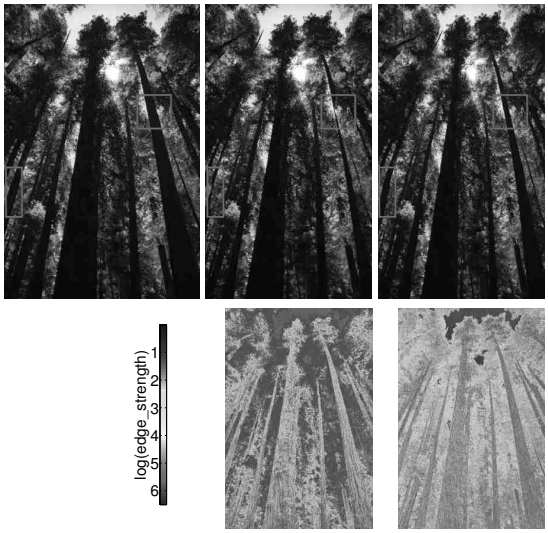


Figure 9: *HDR image shrinking by seam carving (150 pixels horizontally). First row left: original HDR image. Middle: result when the Sobel operator is used for importance map construction. Right: result using the proposed visually significant edges. Images are tone mapped [Drago et al. 2003] for the display purposes. Second row: 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 strenghts (due to the JND scaling), and the structural differences in the edge map (due to the masking).*

Figure 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. 2003] for the display purposes. Second row: (e,f,g) corresponding edge strength maps.*

6.2 HDR Tone Mapping

As mentioned in experimental evaluations [Kuang et al. 2007; 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 appearence. 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 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 and Turk 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 Fig. 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 paper. We thus obtain image decomposition coefficients that are closer to the human visual system response (accounting for phenomena described in Section 4) 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 appearence, see Fig. 11 (right).

6.3 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:

$$E = \max(E_{left} \cdot E_{right}, 0). \tag{5}$$

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 Fig. 12. The source images were inverse tone mapped prior to processing by simple contrast stretching.

7 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 computation time by $30 - 50\%$.

The HVS model is integrated into the edge avoiding wavelet frame-

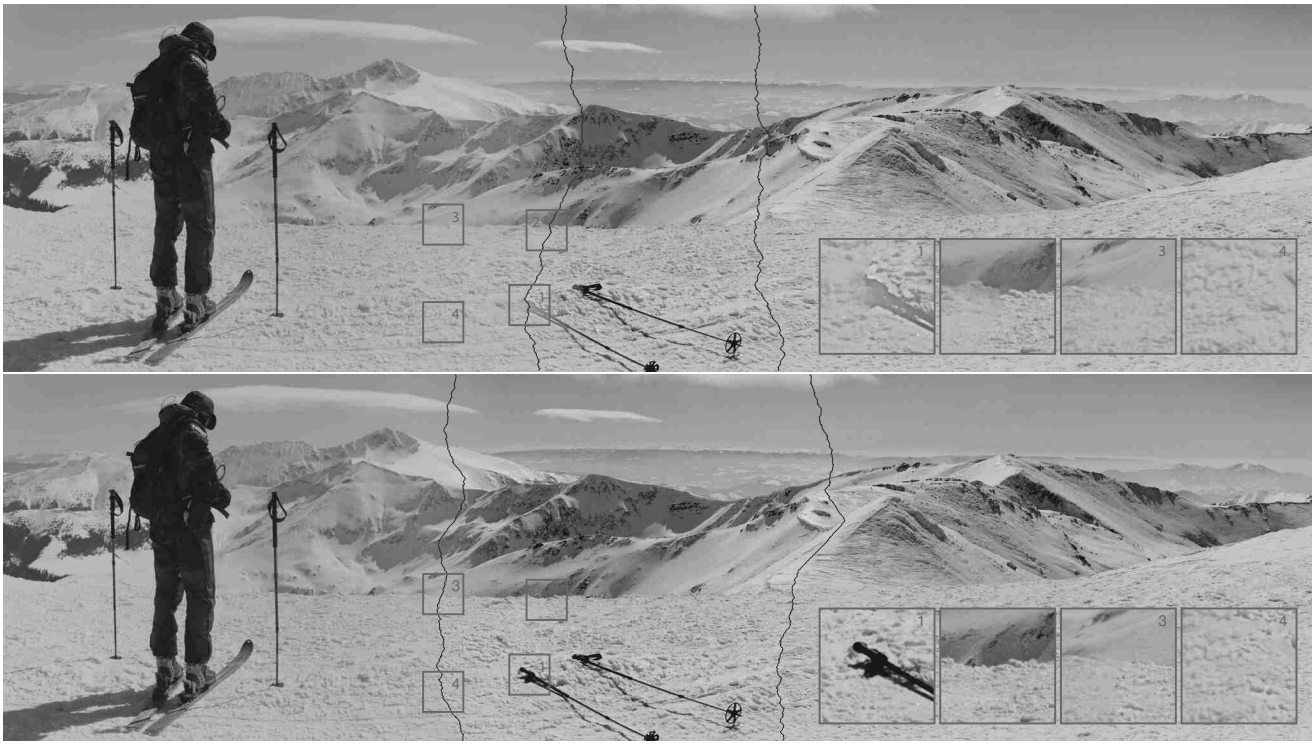


Figure 12: *An HDR panorama stitched from three different, not precisely aligned pictures using Ward's technique [Ward 2006]. Top: the result obtained using Sobel operator, Bottom: the result using the proposed visually significant edges. The images are tone mapped [Reinhard et al. 2002] for display purposes.*

work which provides a convenient basis for edge preserving image decomposition, and also extraction of edges by inverting the edge–stopping citerion. 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 work 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.

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.

References

AVIDAN, S., AND SHAMIR, A. 2007. Seam carving for content-aware image resizing. In *ACM SIGGRAPH*, ACM, New York, NY, USA.

AYDIN, T. O., MANTIUK, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2008. Dynamic range independent image quality assessment. In *ACM SIGGRAPH*, ACM, 1–10.

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

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

ČADÍK, M., WIMMER, M., NEUMANN, L., AND ARTUSI, A. 2008. Evaluation of HDR tone mapping methods using essential perceptual attributes. *Computers & Graphics 32*, 330–349.

CANNY, J. 1986. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell. 8*, 6, 679–698.

COLE, F., GOLOVINSKIY, A., LIMPAECHER, A., BARROS, H. S., FINKELSTEIN, A., FUNKHOUSER, T., AND RUSINKIEWICZ, S. 2008. Where do people draw lines? *ACM Transactions on Graphics (Proc. SIGGRAPH) 27*, 3 (Aug.).

DALY, S. 1993. The visible differences predictor: An algorithm for the assessment of image fidelity. In *Digital Images and Human Vision*, A. B. Watson, Ed. MIT Press, 179–206.

DECARLO, D., AND SANTELLA, A. 2002. Stylization and abstraction of photographs. *ACM Transactions on Graphics 21*, 3 (July), 769–776.

DRAGO, F., MYSZKOWSKI, K., ANNEN, T., AND N.CHIBA. 2003. Adaptive logarithmic mapping for displaying high contrast scenes. *Computer Graphics Forum 22*, 3.

DURAND, F., AND DORSEY, J. 2002. Fast bilateral filtering for the display of high-dynamic-range images. *ACM Transactions on Graphics (Proc. SIGGRAPH) 21*, 3, 257–266.

ELDER, J. H., AND GOLDBERG, R. M. 2001. Image editing in the contour domain. *IEEE Transactions on Pattern Analysis and*

Figure 11: *HDR image tone mapping without (left column) and with our HVS model (right column). The original method [Fattal 2009] preserves as many image details as possible at the cost of overall scene appearence. Our method is more balanced in terms of reproduction of scene appearence and detail preservation.*

Machine Intelligence 23, 3, 291–296.

FARBMAN, Z., FATTAL, R., LISCHINSKI, D., AND SZELISKI, R. 2008. Edge-preserving decompositions for multi-scale tone and detail manipulation. In *SIGGRAPH '08: ACM SIGGRAPH 2008 papers*, ACM, New York, NY, USA, 1–10.

FATTAL, R., LISCHINSKI, D., AND WERMAN, M. 2002. Gradient domain high dynamic range compression. *ACM Transactions on Graphics (Proc. SIGGRAPH) 21*, 3, 249–256.

FATTAL, R. 2009. Edge-avoiding wavelets and their applications. *ACM Trans. Graph. 28*, 3, 1–10.

FERWERDA, J. A., PATTANAIK, S. N., SHIRLEY, P. S., AND GREENBERG, D. P. 1997. A model of visual masking for computer graphics. In *Proceedings of SIGGRAPH 97*, Computer Graphics Proceedings, Annual Conference Series, 143–152.

FREEMAN, W. T., AND ADELSON, E. H. 1991. The design and use of steerable filters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 13*, 9, 891–906.

GEORGESON, M. A., MAY, K. A., FREEMAN, T. C., AND HESSE, G. S. 2007. From filters to features: scale-space analysis of edge and blur coding in human vision. *Journal of Vision 7*, 13.

JANSEN, M. H., AND OONINCX, P. J. 2005. *Second Generation Wavelets and Applications*. Springer.

KUANG, J., YAMAGUCHI, H., LIU, C., JOHNSON, G. M., AND FAIRCHILD, M. D. 2007. Evaluating HDR rendering algorithms. *ACM Transactions on Applied Perception 4*, 2, 9.

LEGGE, G., AND FOLEY, J. 1980. Contrast masking in human vision. *Journal of the Optical Society of America 70*, 12 (Dec.), 1458–1471.

LINDEBERG, T. 1996. Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision 30*, 2, 77–116.

MANTIUK, R., MYSZKOWSKI, K., AND SEIDEL, H. P. 2006. A perceptual framework for contrast processing of high dynamic range images. *ACM Trans. Appl. Percept. 3*, 3, 286–308.

MANTIUK, R., DALY, S., AND KEROFSKY, L. 2008. Display adaptive tone mapping. *ACM Transactions on Graphics (Proc. SIGGRAPH) 27*, 3. Article 68.

MARR, D., AND HILDRETH, E. 1980. Theory of edge detection. *Proc. Royal Soc. Lond. B*, 207, 187–217.

ORZAN, A., BOUSSEAU, A., BARLA, P., AND THOLLOT, J. 2007. Structure-preserving manipulation of photographs. In *Int. Symposium on Non-Photorealistic Animation and Rendering*.

PATTANAIK, S. N., FERWERDA, J. A., FAIRCHILD, M., AND GREENBERG, D. P. 1998. A multiscale model of adaptation and spatial vision for realistic image display. In *SIGGRAPH '98 Proceedings*, 287–298.

PELLEGRINO, F. A., VANZELLA, W., AND TORRE, V. 2004. Edge detection revisited. *Systems, Man, and Cybernetics, Part B, IEEE Transactions on 34*, 3, 1500–1518.

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

PERONA, P., AND MALIK, J. 1990. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence 12*, 629–639.

REINHARD, E., STARK, M., SHIRLEY, P., AND FERWERDA, J. 2002. Photographic tone reproduction for digital images. In *SIGGRAPH '02*, ACM Press, 267–276.

RUBINSTEIN, M., SHAMIR, A., AND AVIDAN, S. 2009. Multi-operator media retargeting. In *ACM SIGGRAPH*, ACM, New York, NY, USA, 1–11.

RUZON, M. A., AND TOMASI, C. 1999. Color edge detection with the compass operator. In *Computer Vision and Pattern Recognition*, vol. 2, 166 Vol. 2.

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

TAYLOR, M. M., AND CREELMAN, C. D. 1967. PEST: Efficient estimates on probability functions. *The Journal of the Acoustical Society of America 41*, 4A, 782–787.

TUMBLIN, J., AND TURK, G. 1999. LCIS: A boundary hierarchy for detail-preserving contrast reduction. *Proceedings of SIGGRAPH*, 83–90.

UYTTERHOEVEN, G., ROOSE, D., AND BULTHEEL, A., 1997. Wavelet transforms using the lifting scheme.

WANDELL, B. 1995. *Foundations of Vision*. Sinauer Associates.

WANG, Y.-S., TAI, C.-L., SORKINE, O., AND LEE, T.-Y. 2008. Optimized scale-and-stretch for image resizing. In *ACM SIGGRAPH Asia*, ACM, New York, NY, USA, 1–8.

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

WATSON, A. B., AND SOLOMON, J. A. 1997. A model of visual contrast gain control and pattern masking. *J. Opt. Soc. Am. A*, 14, 2379–2391.

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

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

ZENG, W., DALY, S., AND LEI, S. 2000. Visual optimization tools in JPEG 2000. In *Proc. of Inter. Conf. on Image Processing*, vol. 2, 37 –40 vol.2.

ZIOU, D., AND TABBONE, S. 1997. Edge detection techniques - an overview. Tech. rep., International Journal of Pattern Recognition and Image Analysis.

Appendix L

Contrast Prescription for Multiscale Image Editing

D. Pająk, M. Čadík, T. O. Aydın, M. Okabe, K. Myszkowski, and H.-P. Seidel. Contrast Prescription for Multiscale Image Editing. *The Visual Computer Journal*, Vol. 26, No. 6-8, pp. 739–748, June 2010.
IF=1.073

Vis Comput (2010) 26: 739–748
DOI 10.1007/s00371-010-0485-3

ORIGINAL ARTICLE

Contrast prescription for multiscale image editing

Dawid Pająk · Martin Čadík · Tunç Ozan Aydın · Makoto Okabe · Karol Myszkowski · Hans-Peter Seidel

Published online: 14 April 2010

Abstract Recently proposed edge-preserving multi-scale image decompositions enable artifact-free and visually appealing image editing. As the human eye is sensitive to contrast, per-band contrast manipulation is a natural way of image editing. However, contrast modification in one band usually affects contrasts in other bands, which is not intuitive for the user. In practice, the desired image appearance is achieved through an iterative editing process, which often requires fine tuning of contrast in one band several times. In this article we show an analysis of properties of multiscale contrast editing frameworks and we introduce the concept of contrast prescription, which enables the user to lock the contrast in selected areas and bands and make it immune to contrast manipulations in other bands.

Keywords Multiscale image editing · Contrast enhancement · Interactive image processing · HDR · Computational photography · Image decomposition

D. Pająk (✉)
Computer Science Department, West Pomeranian University of Technology, Żołnierska 49, 71-210 Szczecin, Poland
e-mail: dpajak@mpi-inf.mpg.de

M. Čadík · T.O. Aydın · M. Okabe · K. Myszkowski · H.-P. Seidel
Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany

M. Čadík
e-mail: mcadik@mpi-inf.mpg.de

T.O. Aydın
e-mail: tunc@mpi-inf.mpg.de

M. Okabe
e-mail: mokabe@mpi-inf.mpg.de

K. Myszkowski
e-mail: karol@mpi-inf.mpg.de

H.-P. Seidel
e-mail: hpseidel@mpi-inf.mpg.de

Fig. 1 An application of prescription idea in multiscale contrast manipulation. Enhancement of medium/high frequency contrast bands produces saturation of some image features existing in lower bands. Prescribing the contrast of unmodified bands prevents the saturation and at the same time allows to effectively increase the contrast according to user's request

1 Introduction

Contrast editing is a common post-processing step in digital photography, usually aimed towards improving the photograph's aesthetical appeal. Research on contrast editing has been focused on two main issues: developing versatile, yet computationally efficient frameworks that produce artifact-free results, and developing user interfaces that provide intuitive interaction with these underlying frameworks. Common to all state-of-the-art contrast editing methods is the involvement of a multiscale image decomposition, through which the image contrast can be edited in an arbitrary number of scales. This tendency is not surprising; since the human visual system (HVS) comprises mechanisms to perceive contrast in multiple spatial frequencies, editing fine and coarse image details separately feels only natural for the end user.

In a multiscale framework, a perfect separation between individual scales such that no two scales have any overlap, can theoretically be achieved by using frequency domain filters with sharp cutoffs. However this simple approach is never applied in practice since it results in heavy ringing artifacts, which can only be avoided by a smooth transition between filters of different scales. As a direct result, an important, but often ignored property of multiscale frameworks is the interaction between contrast at individual scales. Thus, enhancing the contrast of, for example, medium frequency details, indirectly affects the appearance of fine and coarse details due to the overlap between the filters of the neighboring scales in frequency domain. The central idea of this work is "contrast prescription", where the user selects a certain image region for which the contrast at each unmodified scale is locked ("prescribed"), and thus the image details with the desired spatial frequency can be edited independently.

The practical implication of contrast prescription is a more intuitive contrast editing experience. While a set of user controls (often sliders) that control the contrast amplitude at different scales gives the impression of being orthogonal to each other, in reality the changes at one scale propagates to others; an effect which we call "leaking" of contrast. These leaks can effectively be prevented by prescribing the contrast in the image region being edited, which frees the user from iteratively adjusting interface controls to confine the contrast change to the desired scale.

We show that contrast prescription can be implemented in multiple state-of-the-art contrast editing frameworks. Our GPU implementation combined with an intuitive user interface comprising brush and slider controls provides real-time feedback. In this work we focus on editing both low-dynamic range (LDR) and high-dynamic range (HDR) images using ordinary (LDR) display devices. In the rest of the paper we discuss related work on contrast editing and multiscale image decompositions (Sect. 2), give details on the related consequences of multiscale editing (Sect. 3) and how we address them (Sect. 4). Next, we discuss details on the extension of multiple frameworks to handle contrast prescription (Sect. 5), and finally present our results (Sect. 6).

2 Related work

Multiscale image decompositions such as the Laplacian pyramid [3, 11] have been successfully applied to many image processing tasks, including image editing. The practical advantage of considering multiple scales for image editing is the ability to modify the appearance of coarse and fine details separately [17]. On the downside, enhancing fine details disproportionally to coarser details leads in the extreme case to the well-known halo artifacts, resulting in unnatural images often considered not to be aesthetically pleasing.

Edge-preserving image decompositions, on the other hand, minimize halo artifacts by avoiding smoothing across strong edges. The edge-preserving behavior is accomplished through non-linear filters such as weighted least squares [14], anisotropic diffusion [2, 24], or the bilateral filter [27]. Motivated by the anisotropic diffusion, [28] proposed a hierarchical approach called LCIS for HDR tone mapping purposes. The bilateral filter has been widely used in HDR tone mapping as well [5], but also in image fusion [6, 9], example-based transfer of photographic look [1], among others. However, the extra complexity of the filters confine applications of Bilateral filtering to work offline. The performance issue has been addressed by introducing the "Bilateral grid" as a data structure built on Bilateral filter, which enables real-time, multiscale edge-aware image manipulations [4]. The edge-preserving behavior has been further improved by a weighted least squares (WLS) based framework [7], and later another framework based on edge avoiding wavelets [8] has been shown to achieve similar quality results much faster, due to the involvement of the linear time lifting scheme. The principle idea of preserving edges during decomposition has also been used in contrast processing of HDR images [21]. This method relies on performing image editing by first scaling perceptually linearized image gradients, and then reconstructing the image from the new gradient values. Recently, Subr et al. [25] proposed another edge-preserving image decomposition based on local extrema.

One consequence of multiscale image editing is the increased complexity of the editing process from the user's perspective. In fact, interfaces for intuitive image manipulation have been an active topic of research [15, 16, 18]. In our work we used an intuitive brush based interface, using which we were able to generate results in the paper in sessions lasting a few minutes. This was achieved also thanks to the real-time feedback of the underlying contrast processing framework.

3 Consequences of band modification in multiscale image decomposition

Contrast can vaguely be described as the difference between the intensity of an image location with the intensity of some neighborhood around that location, normalized again by the intensity of the same neighborhood. A mathematical formulation of this quantity is possible for simple luminance patterns (such as a foreground–background stimulus with luminance profile defined by a step function) where contrast can be defined as Weber's ratio:

$$W = (L_{\max} - L_{\min})/L_{\min}, \tag{1}$$

or the logarithmic ratio

$$G = \log(L_{\max}/L_{\min}), \tag{2}$$

where $L_{\max}$ and $L_{\min}$ are the intensities of the foreground and the background, respectively. For a sinusoidal luminance pattern, contrast can be computed using Michelson's formula $M = (L_{\max} - L_{\min})/(L_{\max} + L_{\min})$ with $L_{\max}$ and $L_{\min}$ being the sinusoid's peak points. Note that choosing among these "simple" contrast definitions is contextual, since they can trivially be converted to one another if required.

Natural images, however, are much more complex than mere step or sinusoidal intensity patterns, in that they contain various details at multiple scales. Consequently, the computation of the aforementioned "simple" contrast measures is not clear since $L_{\max}$ and $L_{\min}$ are not well-defined. Peli [23] defines contrast in complex images as the ratio of the bandpass image to the low-pass image at multiple scales

$$C_i = \frac{K_{\sigma(i)} * I - K_{\sigma(i+1)} * I}{K_{\sigma(i+1)} * I}, \tag{3}$$

where $*$ denotes the convolution operation between linear luminance and a low-pass Gaussian kernel $K_{\sigma(i)}$ at scale i, and $\sigma(i) = 2^i/\sqrt{2}$ denotes standard deviation, which accounts for frequency band cutoff.

In this paper, we use a multiscale contrast representation, where each contrast sub-band is calculated as a ratio between successive (i and $i+1$) Gaussian-like[1] smoothings of the image I:

$$G_i = \frac{\text{smoothingOperator}(I, i)}{\text{smoothingOperator}(I, i+1)}. \tag{4}$$

To simplify the computations, the decomposition is performed on the logarithm of luminance (roughly approximating the non-linear perception of luminance), for which the ratio can be replaced with a simple subtraction. Such a difference then gives the logarithmic ratio between an image location at some scale and the mean of its neighborhood, as shown in (2). This representation has also the advantage of being computationally more efficient then Peli's contrast, therefore most of the multiscale image editing frameworks [8, 19] follow this simple band decomposition scheme.

The *selection of smoothing operator* is usually the key factor in the overall performance and quality of a decomposition framework. Marr and Hildreth [22] define two main requirements that need to be met for the good smoothing filter. The first is that every multiscale decomposition requires the filter spectrum to be smooth and band-limited in the frequency domain. This allows to reduce the range of scales over which intensity changes take place. We can design a band-pass filter that would be perfectly localized in the frequency domain (sharp cutoff or brick-wall type of filter). However, processing an image with such a filter will induce well known ringing artifacts. Non-oscillating low-frequency parts of the image will yield in global oscillations in the output band representation. To prevent such artifacts, one needs to put the second requirement, a spatial localization constraint on the filter characteristics. This requirement, much more important from image editing point-of-view, can be interpreted in a way that every pixel in the filtered image should be computed from a weighted average of nearby pixels. The constraints above are contradicting in a sense that we are able to increase the spatial locality (reduce ringing) at the cost of frequency domain performance (reduced band-pass behavior).

A good example of such a trade-off in filter design is the Gaussian low-pass filter. It is non-negative and non-oscillatory, hence causes no ringing. The response in the frequency domain is a Gaussian function itself with the mean focused around the middle frequency of the band. This feature has an important consequence when applied in multiscale image processing. Such filter is inevitably causing an *energy leakage* between bands: the energy (modification) in one band leaks out to neighboring bands in the successive multiscale image manipulations. We show an illustrative band manipulation example exploiting Gaussian filter in Fig. 2 (left column), where single sub-band modification generates undesired energy leaking in neighboring sub-bands (Fig. 2, middle column). The energy leakage is prevented when ideal band-pass filter is used (identified by box function in frequency domain), however it results in ringing artifacts as described above. In Fig. 3 we compare the energy leakage of Gaussian-based image decomposition with this "ideal" band-pass decomposition.

The uniform smoothing behavior of the Gaussian filter, which leads to halo artifacts if sub-bands are independently modified, is addressed by so-called edge-preserving smoothing operators. They preserve sharp edges by excluding pixels across image discontinuities from consideration, which

[1] In practice, any type of low-pass (and also edge-preserving) filter can be used as a replacement of Gaussian filter.

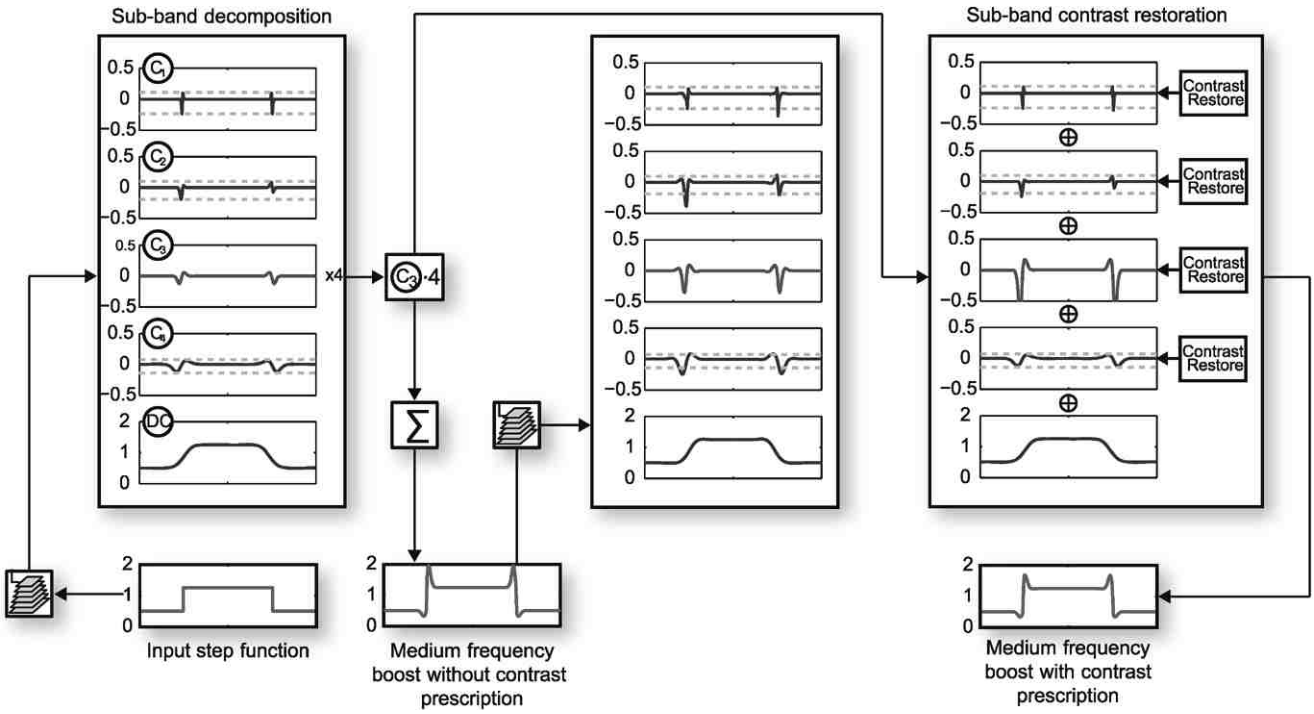


Fig. 2 Step function contrast enhancement. Input function is decomposed into 4 frequency bands and a DC component. Each band is visualized using Peli's definition for the physical contrast. We simulate detail enhancement by multiplying the C_3 band by 4. Decomposing the modified signal again shows that the contrast change of modified band is not proportional to the applied multiplier (*middle column*). Furthermore, due to Gaussian filter *energy leaking* property, all neighboring bands have been modified. Contrast prescription during the computation of modified signal (*right column*), *locally* limits the energy leaking and allows for better preservation of the contrast values in unaffected bands

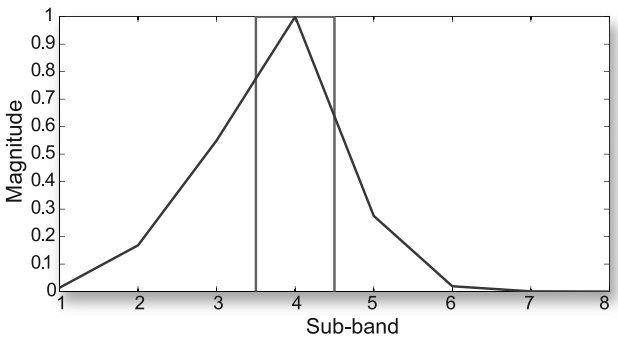


Fig. 3 Comparison of *energy leakage* of a Gaussian based decomposition with an ideal band-pass filter based decomposition. Input impulse signal is decomposed into 8 DoG (Difference-of-Gaussians) bands. While modifying the band we measure the signal magnitude change at each scale, which leads to piecewise linear approximation (*blue plot*). Due to non-ideal band-pass characteristics of the Gaussian filter, boosting the *middle band* results in uncontrolled amplification of features in neighboring scales (*blue plot*). The energy distribution of DoG decomposition resembles Gaussian function itself, but is not symmetric in logarithmic frequency scale. On the contrary, ideal band-pass filter satisfies the localization requirement in frequency domain (*red*). However, when applied to discrete image edges, creates undesired ringing effect in the spatial domain

avoids dividing the energy of the same edge across multiple sub-bands. However, due to imperfect localization in the spatial and frequency domain of the low-pass filter, these halo-free decompositions are still affected by the inter-band energy leakage during band manipulation.

We are not aware of any practical image decomposition scheme which is free of inefficiencies of Gaussian-like low-pass filtering mentioned here. To address this problem we introduce *contrast prescription* which restores unmodified bands physical contrast (i.e. Peli's contrast) during image reconstruction process (see Fig. 2, right column).

4 Contrast prescription

To overcome the inconvenient effect of energy leakage in multiscale based contrast editing applications, we introduce a simple and efficient contrast prescription idea. Our proof-of-concept implementation allows to directly manipulate spatial selection of entire range of image scales. We define prescription-enabled editing as a manipulation in which all unmodified sub-bands are prescribed to keep their contrast values constant. As there are no assumptions made on the type of chosen multiscale decomposition method, the contrasts in prescribed sub-bands are restored while the pyramid-like decomposition is integrated back to the output

image. Adopting contrast representation from (4), the integration is performed by simple addition of all sub-bands:

$$\log(I_{\text{out}}) = \sum_{i=1}^{N} \text{mult}_i \cdot G_i, \tag{5}$$

where mult_i denotes per pixel contrast multiplier for band i and $\cdot$ is an element-wise multiplication operator. Note that multipliers are applied to the logarithmic contrast representation, which is equal to computing a power function on the linear contrast values. In practice, this prevents too strong darkening of the image which otherwise would happen in linear space. However such an operation, in case of detail enhancement, tends to oversaturate the already well-visible details. Our prescription algorithm can counteract this scenario by selectively modulating the contrast values of oversaturated pixels. The overall contrast restoration algorithm is shown as pseudo-code in Fig. 4.

Although we store sub-band contrast in a logarithmic ratio representation, in order to compute contrast changes during successive band manipulations, we utilize Peli's physical contrast measure, as it is a metric that can be reliably applied to complex images and by definition takes into account inter-band dependencies. We employ local sub-band physical contrast ratio to correct the contrast values affected by the cross-band energy leaking. The contrast correction algorithm starts from the lowest frequency band. According to (3), if the band-pass image is constant, the only way to change contrast is to modify the low-pass image of the same band. As we perform interactive multiscale contrast manipulation, the aforementioned situation takes place quite often. In order to correct for this, we simultaneously compute two low-pass images I_0 and I_1 (approximated background luminance), which let us estimate for each sub-band how the contrast has changed. Initially both images are the same, so no correction is applied. However, as we move further, adding up a new sub-band components (Line 9–10 of Fig. 4), the difference between background luminance values becomes more apparent, and directly affects the Peli contrast for prescribed bands. We modify the *mult* set to reflect the background luminance change and locally correct for the suppression or enhancement of prescribed sub-band contrast.

Given a *linear* low-pass image I_1 and its prescribed counterpart I_0, we perform pixel-wise scaling of mult (line 7) by $(\frac{I_0}{I_1})^\beta$ ratio, which for $\beta = 1$ corresponds to restoring Peli's contrast for a certain pixel. The β scaling parameter provides a non-linear control over the performance of contrast restoration. As the sub-band contrast is stored in a low-pass logarithmic format, before changing the contrast value we need to convert it to an intermediate notation for which the linear band-pass signal is expressed in the nominator. For this purpose, we use the Weber fraction computed as $W = \Delta L/L = 10^{|G|} - 1$. After modifying the contrast we apply simple transformation to get back to low-pass logarithmic contrast: $G = \text{sign}(G) \cdot \log(W + 1)$. Corrected mult multipliers are stored in a separate location, so they can be used in construction of reference contrast pyramid G in user's future edits. After applying the contrast restoration formula for all sub-bands we output the image to the display.

The bottom-up approach is motivated by the fact that the most of the energy leaking is due to the large signal amplitudes of low-frequency bands. This is supported by the findings in natural image statistics [10]: most of the image energy is focused around low-frequency bands, which is known as the power law for the amplitudes of frequencies.

1: **ContrastRestore**(MultiScaleMultipliers mult, MultiScaleDecomposition G)
2: $N \leftarrow \text{height}(G)$ // sub-band count
3: $I_0 \leftarrow G_N$ // log(DC)
4: $I_1 \leftarrow G_N$ // log(DC)
5: **for** $i \leftarrow N - 1$ **downto** 1 **do**
6: **for all** pixels **do**
7: $R \leftarrow \text{mult}_i \cdot 10^{(I_0 - I_1)\cdot\beta}$ // corrected multiplier
8: $W \leftarrow 10^{|G_i|} - 1$
9: $I_0 \leftarrow I_0 + G_i$
10: $I_1 \leftarrow I_1 + \text{sign}(G_i) \cdot \log(W \cdot R + 1)$
11: **end for**
12: **end for**
13: $I_{\text{out}} \leftarrow I_1$

Fig. 4 Contrast restoration algorithm. We simultaneously integrate two pyramids using previous and current band multipliers. The integration incorporates a correction fraction for I_1 (output) that locally restores contrast in prescribed bands. The I_0 image serves as a prescribed sub-band adaptation luminance reference which is required for computing the correction factor

5 Extension of edge-preserving decompositions

In this section we discuss extending recent edge-stopping multiscale decomposition frameworks with contrast prescription. Furthermore, we introduce supplementary extension that allows the user for counter-shading (halo editing).

5.1 Weighted least squares decomposition

The contrast prescription algorithm can be implemented in WLS optimization framework in a straightforward manner. The basic idea behind the WLS based decomposition is to keep the complete frequency domain representation of each edge at only one scale. The multiscale image decomposition is achieved by iterative application of an edge-stopping image smoothing operator, which is tuned in subsequent iterations to preserve edges of successively larger contrast. In order to convert such a representation to contrast sub-bands,

we employ (4). The smoothing operator is designed as a Poisson-class linear optimization which minimizes the energy function that penalizes the image gradients (smoothing effect) in the whole image except near strong edges (edge-stopping effect). Due to the existence of high frequency edges in the band-pass image, the WLS filter requires the bands to be stored as full-resolution images.

The weighting function, responsible for edge-stopping behavior is expressed as:

$$w_n(m) = \frac{1}{|L_n - L_m|^\alpha + \epsilon}, \tag{6}$$

where w_n denotes an edge weight between pixel value L_n and its neighbor L_m (α is a model parameter and ϵ prevents division by zero). Contrast restoration mechanism (see Fig. 4) is applied directly on the WLS multiscale image decomposition and the resulting image is obtained by adding up the sub-bands.

5.2 Second generation wavelet decomposition

Recent work [8] showed the application of second generation wavelets to edge-preserving image decomposition. Here, the image is decomposed using edge avoiding wavelets (EAW)—second generation wavelets constructed with a weighting function similar to (6). The computation of second generation wavelet decomposition is performed using the lifting scheme [26].[2]

Our wavelet implementation is based on a Weighted Red-Black decomposition (WRB) [29]. After transforming the image into wavelet representation we cannot directly use the wavelet scaling function coefficients to apply our contrast prescription algorithm. In order to recover DoG-like decomposition of the image, we compute N inverse wavelet transformations. Each inverse transformation sets all scaling coefficients to zero, except the ones which describe features at scale i. Such an algorithm performs an edge-aware interpolation (upsampling) of selected sub-band components. The output image is a full-resolution sub-band which roughly corresponds to the results obtained by the WLS framework mentioned above. As we show in Sect. 6 our GPU implementation of wavelet decomposition is very fast; the entire process takes less than 5 ms on mainstream hardware.

5.3 Perceptual contrast processing framework

Mantiuk et al. [21] presents a framework in which the inter-band dependencies are tightly integrated in the inhomogeneous Laplace equation and the prescription algorithm cannot be applied in the form described earlier. In this framework, especially suitable for processing HDR images, the final image is a result of least square optimization (computed using a Poisson solver) and is not reproduced by simple addition of sub-bands. In order to make our approach applicable we implement contrast restoration as a post-processing step. We use two separate decompositions to track the changes before and after manipulations, constructed by the EAW algorithm described earlier due to its efficiency. Contrast prescription is then realized using sub-band contrasts obtained from these external decompositions.

5.4 Interactive halo editing

The use of counter-shading to enhance the perceived contrast has been known by painters for ages [20]. More recently, [13] proposed an automatic technique for improving contrast perception in digital images by modulating brightness at the edges. In our technique, we allow the user to control the halo effect (counter-shading) *manually*. This is implemented inside the edge-preserving decomposition by means of a minor modification in the weighting function (6) used by both decomposition frameworks. By modulating α coefficient, which is responsible for edge-stopping behavior of decomposition scheme, we can suppress or enhance halo effect near the edges. When α is close to 0, the weights are becoming more spatially uniform, thus the smoothing operator resembles regular Gaussian-like filter. This results in a decomposition that, in case of a local manipulation, is affected by the halo effect. For each sub-band, we define the α pixel-wise. To modify these coefficients we use the same approach as for updating contrast multipliers (see Sect. 6).

6 Results

In this section we present results of a comparison of prescription-enabled editing with regular one on a number of images. Our proof-of-concept software[3] was tested on a mainstream PC equipped with Intel Core2 Duo 3.0 Ghz CPU and NVidia GTX260 GPU. We compared the performance of decompositions presented here. In most cases the decomposition can be done off-line, as a preprocessing step before actual editing session. However, features like halo editing require recreating the sub-bands every time we modify edge weights. Consequently, we chose wavelet decomposition as our benchmark implementation since it is significantly faster than other schemes and the results we obtained are comparable. For a 1 MPixel image, the forward wavelet transform coupled with generation of 8 full-resolution contrast bands takes less than 5 ms. Hence, the framework runs at interactive speeds even

[2] We refer the reader to Jansen and Oonincx [12] for a detailed discussion on second generation wavelets.

[3] The implementation comprises of a platform independent Java UI and native GLSL image processing library.

```
UpdateMultipliers(MultiScaleMultipliers mult, Image mask)
N ← height(mult)   // sub-band count
for i ← 1 to N − 1 do
    B = GetBandMultiplier(i)
    for all pixels do
        mult_i ← max(0, mult_i + mask · B)
    end for
    mask = mask ↓ 2
end for
```

Fig. 5 Contrast multipliers update. The GetBandMultiplier(i) function returns [0, 1] normalized value which indicates mask scaling factor for band i. It can be either user-defined by setting up scale range sliders or computed in fully automatic manner

for large images. In case of Poisson solver, on the other hand, each iteration takes about 3 ms, depending on the amount of modification applied. Note that edits are performed iteratively on a small parts of the image. Therefore, the solver is initialized with a good quality solution, which only needs to be corrected in some selected regions. On average our conjugate gradient based solver requires about 10 iterations to converge. The test software binaries are available for Windows/Linux and can be downloaded from (http://mpi-inf.mpg.de/~dpajak/prescription).

Our implementation allows the user to intuitively manipulate the contrast multipliers. Brushing over selected areas creates a smooth amplitude mask with Gaussian-like fall-off, which is then used to update per band, per pixel multipliers (see Fig. 5). We decided to implement brush based interface as it is still considered to be the most common tool used for manual image retouching. However, this interface can be easily extended by introducing more automated, diffusion-based segmentation as in [19]. For a novice user, manipulating bands with scale range sliders can be challenging. Therefore, we include a brushing mode where band multipliers (see Fig. 5) are computed automatically by measuring image energy[4] for current selection. This approach will always try to selectively boost the bands with smallest energy value.

6.1 Contrast prescription

We demonstrate our approach by performing a set of exemplary, yet typical, contrast editing sessions. First, we show a simple scenario, where only one band range is modified and then we stage more complex manipulation to show how bands interact with each other.

In Figs. 1 and 6, we perform single modification of fine image details using WRB Wavelet decomposition scheme

Fig. 6 Example of contrast enhancement based on WRB wavelet decomposition for $\alpha = 0.8$ (6) and $\beta = 1.0$ (Fig. 4). Boosting fine/medium scales in the ceiling area causes saturation of some image features and results in unnatural appearance. Despite the extreme boost, contrast restoration allows to regain natural look of modified area and get the requested detail enhancement

Fig. 7 Comparison of regular and prescription-enabled image editing in WLS framework ($\alpha = 1.2$). Amplification of coarse features attenuates contrasts in higher bands. Contrast restoration algorithm successfully recovers fine details and boosts coarse image features at the same time

with parameters $\alpha = 0.8$ and $\beta = 1.0$. In both figures, we used the same band range multiplier for prescribed and non-prescribed operation. Due to the energy leakage, features existing across multiple scales are over saturated and by com-

[4]Modulo of a gradient for gradient domain frameworks and absolute amplitude of band-pass contrast for multi-scale decompositions.

paring against the source image we see that their contrast enhancement is spatially inconsistent. Prescription counteracts these situations and allows for more uniform and controllable contrast modification. Note that it is cumbersome to achieve such a result with a series of separate edits since the restoration is spatially local and highly non-linear.

Figure 7 shows an opposite scenario, where user scales low-frequency bands. The modification causes the loss of visibility of trees and plant pot details. Prescription of contrasts in upper bands allows to achieve both goals, increasing the global contrast and maintaining the detail visibility.

In Fig. 8, the image is initially modified by enhancing fine details around the plant area. Next, we boost low-frequency bands, which reduces the visibility of previous edits. Also, high signal amplitudes of low-frequency content exposes the energy leakage issue, resulting in an over-saturated image and decreased perception of unmodified contrasts bands. Contrast prescription visibly restores the details and reduces the saturation while still allowing for large enhancement of low-frequency contrasts.

Finally, we illustrate the complete editing session result, Fig. 10, where we manipulate contrasts in order to transfer the style of a professional photographer to a plain picture of the same location. Despite the obvious difference in source material we managed to properly reflect the style using only local, prescription-enabled contrast modifications.

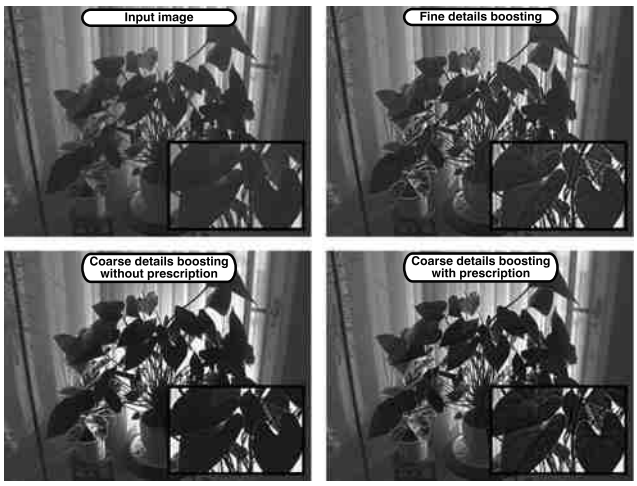


Fig. 8 Contrast editing session in WLS decomposition framework ($\alpha = 1.2$). Contrast prescription prevents the loss of details in unmodified sub-bands, as a result, previously modified fine details are preserved

6.2 Interactive halo editing

As described in Sect. 5.4, we enhanced each of the implemented decomposition frameworks to allow interactive halo manipulation. Figure 9 illustrates simple, low-frequency halo suppression case. Although the local modification of edge-stopping filter behavior usually requires repeating the decomposition, we show that an efficient hardware implementation can deliver an interactive solution even in case of Poisson solver based frameworks.

7 Conclusions and future work

In this work, we analyzed the properties of multiscale image representations used in interactive image editing applications. Direct consequence of Gaussian-like filters commonly used to construct such representations (including edge-aware decompositions), is the effect that we call "energy leaking". When the user modifies one sub-band, part of the "energy" of this modification in effect leaks out to the other sub-bands in successive manipulations. This can affect the perception of already edited parts of the image and leads to non-intuitive, iterative way of image editing. Moreover, change in one band can result in oversaturation of another band (also possibly previously edited), e.g. when user boosts overall contrast, the tiny details might be lost. To overcome those limitations inherently imposed by all existing multiscale image editing frameworks, we propose the concept of contrast prescription, where once edited parts of the edited image retain their prescribed values. The aim is to restore the visibility of each sub-band contrast and limit the effects of cross-band energy leakage. Consequently, reduced number of decomposition related artifacts, allows for more intuitive and controllable multiscale contrast manipulation. To illustrate the concept, we show simple, but efficient and interactive extension of three state-of-the-art multiscale frameworks.

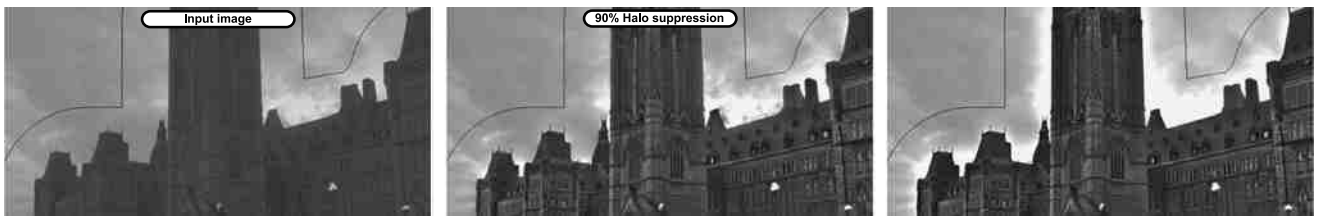


Fig. 9 Halo editing example with a Poisson solver based decomposition framework [21]. The original castle scene is modified by elevating high/medium frequency bands. To artificially modulate contrast along the edges, e.g. for better decomposition of foreground from background, one might locally allow the halo effect to be visible

Fig. 10 Practical demonstration of prescription-enabled contrast editing (*center*). We modified the source image (*left*) to reflect the style and feeling of pictures taken by Ansel Adams (*right*). As the artist style was based mostly on manual dodging and burning we had to perform extreme band manipulations that applied without prescription would result in heavy artifacts. The entire session was about 8 minutes long and included only contrast manipulations on WLS based decomposition

So far, we assumed editing using an ordinary (LDR) display device—in this case the illumination stayed almost constant and the effect of image editing on the user's visual adaptation was very subtle. However, when editing on an *HDR display* device, the change of user's adaptation due to the display luminance is not negligible any more and it can significantly bias user's perception. Modeling of apparent contrast is required to compensate for this effect which suggests a possible extension to this work.

References

1. Bae, S., Paris, S., Durand, F.: Two-scale tone management for photographic look. In: Proc. SIGGRAPH 2006, pp. 637–645 (2006)
2. Black, M., Sapiro, G., Marimont, D., Heeger, D.: Robust anisotropic diffusion. IEEE Trans. Image Process. **7**(3), 421–432 (1998)
3. Burt, P.J., Adelson, E.H.: The Laplacian pyramid as a compact image code. IEEE Trans. Commun. **31**, 532–540 (1983)
4. Chen, J., Paris, S., Durand, F.: Real-time edge-aware image processing with the bilateral grid. In: Proc. SIGGRAPH 2007, p. 103 (2007)
5. Durand, F., Dorsey, J.: Fast bilateral filtering for the display of high-dynamic-range images. In: SIGGRAPH '02: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, pp. 257–266. ACM, New York (2002)
6. Eisemann, E., Durand, F.: Flash photography enhancement via intrinsic relighting. ACM Trans. Graph. **23**(3), 673–678 (2004)
7. Farbman, Z., Fattal, R., Lischinski, D., Szeliski, R.: Edge-preserving decompositions for multi-scale tone and detail manipulation. ACM Trans. Graph. **27**(3) (2008)
8. Fattal, R.: Edge-avoiding wavelets and their applications. ACM Trans. Graph. **28**(3), 1–10 (2009)
9. Fattal, R., Agrawala, M., Rusinkiewicz, S.: Multiscale shape and detail enhancement from multi-light image collections. ACM Trans. Graph. (Proc. SIGGRAPH) **26**(3) (2007)
10. Gibson, J.D., Bovik, A. (eds.): Handbook of Image and Video Processing. Academic Press, Orlando (2000)
11. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Prentice-Hall, New Jersey (2002)
12. Jansen, M.H., Oonincx, P.J.: Second Generation Wavelets and Applications. Springer, Berlin (2005)
13. Krawczyk, G., Myszkowski, K., Seidel, H.P.: Contrast restoration by adaptive countershading. Comput. Graph. Forum **26**(3), 581–590 (2007)
14. Lagendijk, R., Biemond, J., Boekee, D.: Regularized iterative image restoration with ringing reduction. IEEE Trans. Acoust. Speech Signal Process. **36**(12), 1874–1888 (1988)
15. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. SIGGRAPH **23**(3), 689–694 (2004)
16. Li, Y., Sun, J., Tang, C.K., Shum, H.Y.: Lazy snapping. SIGGRAPH **23**(3), 303–308 (2004)
17. Li, Y., Sharan, L., Adelson, E.H.: Compressing and companding high dynamic range images with subband architectures. ACM Trans. Graph. (Proc. SIGGRAPH) **24**(3), 836–844 (2005)
18. Li, Y., Adelson, E., Agarwala, A.: Scribbleboost: adding classification to edge-aware interpolation of local image and video adjustments. Comput. Graph. Forum **27**(4), 1255–1264 (2008)
19. Lischinski, D., Farbman, Z., Uyttendaele, M., Szeliski, R.: Interactive local adjustment of tonal values. ACM Trans. Graph. **25**(3), 646–653 (2006)
20. Livingstone, M.: Vision and Art: The Biology of Seeing. Harry N. Abrams (2002)
21. Mantiuk, R., Myszkowski, K., Seidel, H.P.: A perceptual framework for contrast processing of high dynamic range images. ACM Trans. Appl. Percept. **3**(3), 286–308 (2006)
22. Marr, D., Hildreth, E.: Theory of edge detection. Proc. R. Soc. Lond., Ser. B, Biol. Sci. **207**(1167), 187–217 (1980)
23. Peli, E.: Contrast in complex images. J. Opt. Soc. Am. A **7**(10), 2032–2040 (1990)
24. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell. **12**(7), 629–639 (1990)
25. Subr, K., Soler, C., Durand, F.: Edge-preserving multiscale image decomposition based on local extrema. In: ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2009), Annual Conference Series. ACM, New York (2009)
26. Sweldens, W.: The lifting scheme: a construction of second generation wavelets. SIAM J. Math. Anal. **29**(2), 511–546 (1997)
27. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Proceedings of the Sixth International Conference on Computer Vision (ICCV '98), p. 839. IEEE Computer Society, Los Alamitos (1998)
28. Tumblin, J., Turk, G.: LCIS: A boundary hierarchy for detail-preserving contrast reduction. In: Proc. SIGGRAPH '99, pp. 83–90 (1999)
29. Uytterhoeven, G., Roose, D., Bultheel, A.: Wavelet transforms using the lifting scheme (1997)

Dawid Pająk is a PhD student at West Pomeranian University of Technology. He received MSc Eng degree in computer science from Technical University of Szczecin in 2007. His research interests include image processing, rendering, high performance computing and GPGPU applications. Before moving into research he was a professional game developer.

Martin Čadík is doing research in the fields of image and video quality assessment, high dynamic range imaging and image processing. He received MSc and PhD degrees in computer science from Czech Technical University in Prague in 2002 and 2008, respectively. He is currently a post-doc research fellow at Max-Planck-Institut für Informatik.

Tunç Ozan Aydın is a PhD student at Max-Planck-Institut für Informatik, and holds a BS degree in Civil Engineering from Istanbul Technical University and an MSc degree in Computer Science from Georgia Institute of Technology. His research interests are human visual system modeling, image/video quality assessment and HDR imaging.

Makoto Okabe is an assistant professor at department of computer science, the University of Electro-Communications. He received PhD from department of information science and technology, the University of Tokyo in March 2008. He worked as a postdoc in computer graphics group of Max-Planck-Institut für Informatik from 2008 to 2010. His research interest is user interface for computer graphics and current focus is on image and animation synthesis based on image and video database.

Karol Myszkowski is a senior researcher in the Computer Graphics Group of the Max-Planck-Institut für Informatik. From 1993 to 2000, he served as an Associate Professor at the University of Aizu, Japan. During the years 1985–1992, he worked as a Research Associate and then Assistant Professor at Szczecin University of Technology. He received his MSc degree in electrical engineering from Szczecin University of Technology, and PhD and habilitation degrees in computer science from Warsaw University of Technology in 1983, 1991 and 2001, respectively. His research interests include perception issues in graphics, high-dynamic range imaging, global illumination, rendering, and animation.

Hans-Peter Seidel is the scientific director and chair of the computer graphics group at the Max-Planck-Institut (MPI) Informatik and a professor of computer science at Saarland University. He has published and lectured widely. He has received grants from a wide range of organizations, including the German National Science Foundation (DFG), the German Federal Government (BMBF), the European Community (EU), NATO, and the German-Israel Foundation (GIF). In 2003 Seidel was awarded the 'Leibniz Preis', the most prestigious German research award, from the German Research Foundation (DFG). Seidel is the first computer graphics researcher to receive this award.

Appendix M

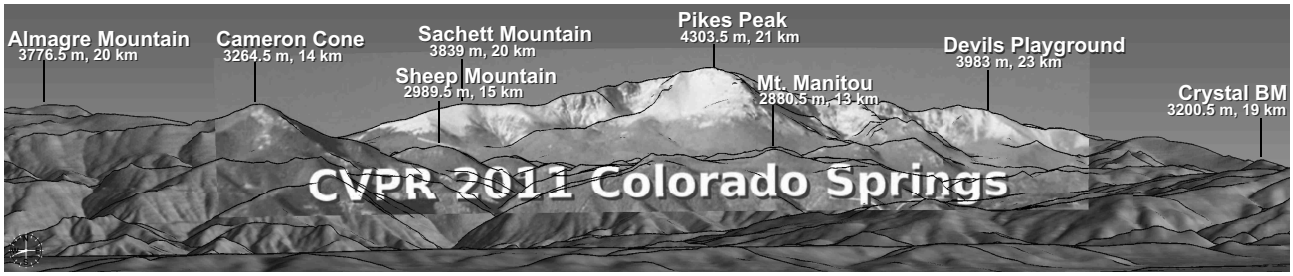
Automatic Photo-to-Terrain Alignment for the Annotation of Mountain Pictures

L. Baboud, M. Čadík, E. Eisemann, and H.-P. Seidel. Automatic Photo-to-terrain Alignment for the Annotation of Mountain Pictures. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR Orals)*, pp. 41–48, IEEE Computer Society, Washington, DC, USA, 2011.
CVPR Oral, acceptance rate 3.5%

Automatic Photo-to-Terrain Alignment for the Annotation of Mountain Pictures

Lionel Baboud* Martin Čadík⋆ Elmar Eisemann† Hans-Peter Seidel⋆

⋆Max-Planck Institute Informatik, †Telecom ParisTech/CNRS-LTCI



Abstract

*We present a system for the annotation and augmentation of mountain photographs. The key issue resides in the registration of a given photograph with a 3D georeferenced terrain model. Typical outdoor images contain little structural information, particularly mountain scenes whose aspect changes drastically across seasons and varying weather conditions. Existing approaches usually fail on such difficult scenarios. To avoid the burden of manual registration, we propose a novel automatic technique. Given only a viewpoint and FOV estimates, the technique is able to automatically derive the pose of the camera relative to the geometric terrain model. We make use of silhouette edges, which are among most reliable features that can be detected in the targeted situations. Using an edge detection algorithm, our technique then searches for the best match with silhouette edges rendered using the synthetic model. We develop a robust matching metric allowing us to cope with the inevitable noise affecting detected edges (*e.g. *due to clouds, snow, rocks, forests, or any phenomenon not encoded in the digital model). Once registered against the model, photographs can easily be augmented with annotations (*e.g. *topographic data, peak names, paths), which would otherwise imply a tedious fusion process. We further illustrate various other applications, such as 3D model-assisted image enhancement, or, inversely, texturing of digital models.*

*{lbaboud, mcadik, hpseidel}@mpi-inf.mpg.de

†elmar.eisemann@telecom-paristech.fr

1. Introduction

The internet offers a wealth of audio-visual content and communities such as Flickr and YouTube make large amounts of photos and videos publicly available. In many cases, an observer might wonder what elements are visible on a certain shot or movie. Especially for natural scenes, the answer to this question can be difficult because only few landmarks might be easily recognizable by non experts. While the information about the camera position is (at least roughly) known in many cases (photographer's knowledge or camera GPS), the camera orientation is usually unknown (digital compasses have poor accuracy).

The principal requirement is then the accurate alignment (registration) of a given photograph or video with a 3D georeferenced terrain model. Interestingly, such a precise localization would be useful in many contexts. Services such as Google StreetView could be extended in an automatic fashion to natural environments by exploiting user-provided shots. Further, the photo can be used to texture virtual terrains such as those in Google Earth. Also, annotations, derived from an annotated 3D terrain model, could be added automatically (highlighting important landmarks) which is of interest when describing or planning a field trip. Because of such applications, cameras start being equipped with GPS in order to automatically track photo locations.

We will focus on a special class of content taken in mountain regions, and provide a solution to automatically derive the orientation that was used for a given shot, assuming that the viewpoint location is known accurately enough, as well as the cameras's intrinsic parameters (*e.g.* field-of-view). It is often complicated or even impossible to access

these regions with cars or robots, making user-provided images an interesting way to collect data. Furthermore, users also benefit from our solution, as it enables them to enhance (and even augment) their photos with supplementary data.

The input of our approach is a single photograph or a video and an indication of where it was taken. Our algorithm then automatically finds the view direction by querying the position against a reference terrain model that we assume to have at disposition. The latter is a smaller constraint because satellites can provide very reliable terrain elevation maps even for less accessible regions. Once the view is matched, we can transfer information from the reference model into the photo.

Our main contribution is the robust matching algorithm to successfully find the view orientation of given photo. This task is far from trivial and many previous approaches attempting to match up an image and 3D content can exhibit high failure rates (Section 2). The reason why our algorithm (Section 3) provides a working solution is that we can exploit the special nature of terrains. Mountain silhouettes are relatively invariant under illumination changes, seasonal influence, and even quality of the camera, therefore we detect these features and make them a major ingredient in our matching metric (Sections 4, 5, 6). Finally, we illustrate the robustness and usefulness of our approach with several of the aforementioned application scenarios (Section 7) before concluding (Section 8).

2. Previous Work

The problem of matching appears in several areas of research, but proves difficult in most cases. Advances in camera engineering (i.e. digital compass and GPS receivers) can facilitate the task in the future, but such data is neither available in most current cameras nor present in video sequences. Furthermore, even when available, such information is not reliable enough for an accurate pose estimation and will not be in a long time because the satellite infrastructure would need to change drastically to allow the precision we seek. Usually, existing GPS and compass-based applications only present distant abstracted depictions (*e.g.* Peakfinder (http://peakfinder.ch), Google Skymap) without considering the actual view content. The same holds for augmented reality applications, such as the Wikitude World Browser (http://www.wikitude.org). In a reasonable time frame only initial estimates of a camera pose, but not the final fine-tune registration will be available. In the context we target, orientation must be known accurately to properly discriminate distant peaks, wereas position accuracy is less crucial (negligible parallax).

Registration comes in many variants, usually, instead of matching an entire image, a first step is to restrict the search to a small set of feature points. Such feature-based (SIFT [13], SURF [1]) techniques work robustly for image to image registration, but are less usable for image-to-model registration [23]. Nonetheless, for applications such as panorama stitching [19], feature-based techniques work well and currently dominate. Unfortunately, our case is more difficult because we have to consider very differing views in a natural scene which exhibits many similar features or features that might depend heavily on the time of the year (*e.g.* snow borders). This constraint also renders statistical methods [24], that are widely used in medical image registration, less successful.

The difficulty of this task is also underlined in the photo-tourism approach [17]. Indoor scenes and landmark shots are handled automatically, while outdoor scenes have to be aligned against a digital elevation map and a user has to manually specify correspondences and similarity transforms to initiate an alignment. Similarly, Deep Photo [9] requires manual registration and the user has to specify four or more corresponding pairs of points.

In our experience, even simpler tasks, such as horizon estimation [6], tend to fail in mountain scenes. Similarly, advanced segmentation techniques [7, 16] proved futile. Maybe for these reasons, existing photogrammetry approaches for mountain imagery, such as GIPFEL (http://flpsed.org/gipfel.html), strongly rely on user intervention.

Robust orientation estimation is a necessary component of localization algorithms for autonomous robots. During missions on moon or mars, it is impossible to rely on standard GPS techniques, but satellite imagery can deliver a terrain model. Many of these algorithms rely on the horizon line contour (HLC) which is the outline of the terrain and specific feature points thereon that are matched with extracted terrain features [2, 21, 8]. Peaks of the HLC are often used as features, but might not correspond to actual peaks in the terrain due to partial occlusion (clouds, fog, or haze), terrain texture (*e.g.* snow), or an incorrect sky detection (see Fig. 3). The latter is very difficult, but particularly crucial for HLC approaches, especially when estimating visibility between peaks in the query image [8]. Learning techniques [8, 14] can often lead to successful segmentations, but they depend on the training set and implicitly assume similar query images (*e.g.* same daytime). Furthermore, even if successful, the localization of peaks in a photo is error prone [2] and can lead to a deviation in the estimate. Hence, sometimes only virtual views are tested [21], or an accurate compass is supposed [20].

Instead of peaks, using all occluding contours leads to more robustness, but previous solutions [18] needed an accurate orientation estimate and assumed that the query image allows us to well-detect all occluding contours. As for the HLC, this property rarely holds because haze, fog or lighting variations often occlude crucial features. Our approach does not penalize missing contours, and the detec-

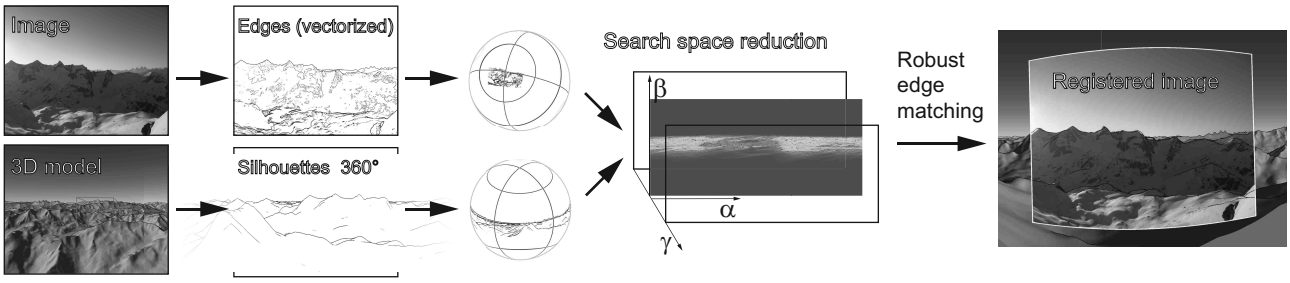


Figure 1. Overview of the proposed technique.

tion robustness does not suffer from false positives.

Interestingly, despite their negative effect on contour detection, haze and fog potentially encode monocular depth information [3] The assumptions on reflectance properties and fog/haze are relatively general and some assumptions such as a ground plane [3] fail in our context. Consequently, the resulting depth estimates are usually coarse and proved insufficient for our purposes.

The area of direct image to model registration is less developed, and most techniques assume some structural elements (*e.g.* straight lines, planes) in the input image [10, 5]. Unfortunately, mountain scenes are highly unstructured making matching very challenging which lead us to develop our approach.

3. Problem setup

Given a photograph, our goal is to estimate its pose relatively to an accurate 3D terrain model based on a digital elevation map (DEM). We assume that the camera's field of view is known, as well as an estimate p_v of the viewpoint position (accuracy is discussed in Section 7). Given these hypotheses, we are looking for the rotation $\tilde{g} \in SO(3)$ that maps the camera frame to the frame of the terrain. The set of images that can be shot from p_v is entirely defined by a spherical image f centered at p_v against which we need to match the query photo.

We target outdoor scenes that do not allow to rely on photogrammetry information, as it can vary drastically. Instead, we rely on silhouette edges that can be obtained easily from the terrain model and can be (partially) detected in the photograph. In general, the detected silhouette map can be error prone, but we enable a robust silhouette matching by introducing a novel metric (Section 4).

Because a direct extensive search on $SO(3)$ using this metric is very costly, we additionally propose a fast preprocess based on spherical cross-correlation (Section 5). It effectively reduces the search space to a very narrow subset, to which the robust matching metric is then applied. The resulting algorithm is outlined in Fig. 1.

3.1. Spherical parameterization

We start by defining some basic notations. The camera frame has its Z axis pointing opposite to the viewing direction, with X (resp. Y) axis parallel to the horizontal (resp. vertical) axis of the image. The terrain frame has its Z axis along the vertical. Rotations of $SO(3)$ are parameterized with the ZYZ Euler angles, *i.e.* an element $g \in SO(3)$ is represented by three angles (α, β, γ) so that $g = R_Z(\alpha)R_Y(\beta)R_Z(\gamma)$, where R_Y and R_Z are rotations around axes Y and Z.

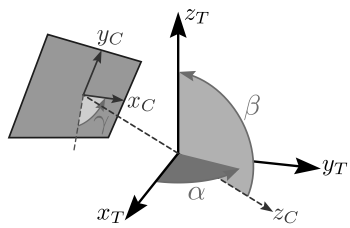


Figure 2. Terrain (x_T, y_T, z_T) and camera (x_C, y_C, z_C) frames.

The synthetic spherical image of the terrain model from p_v will be denoted f, and the spherical representation of the photograph will be denoted p. The corresponding silhouette sets will be denoted $\mathcal{E}_F$ and $\mathcal{E}_P$.

4. Robust silhouette map matching metric

We first address the more costly, but precise fine-matching. In the targeted situation, *i.e.* on photographs of mountainous scenes, results produced by available edge-detection techniques usually contain inaccuracies which can be classified as following (see also Fig. 3):

- some of the silhouette edges are not detected;
- some detected edges are noisy;
- many detected edges are not silhouette edges.

The noisy edges prevent us from using traditional edge matching techniques that often rely on features that are assumed to be present in both images. However the specificity of our problem allows us to derive a robust matching metric.

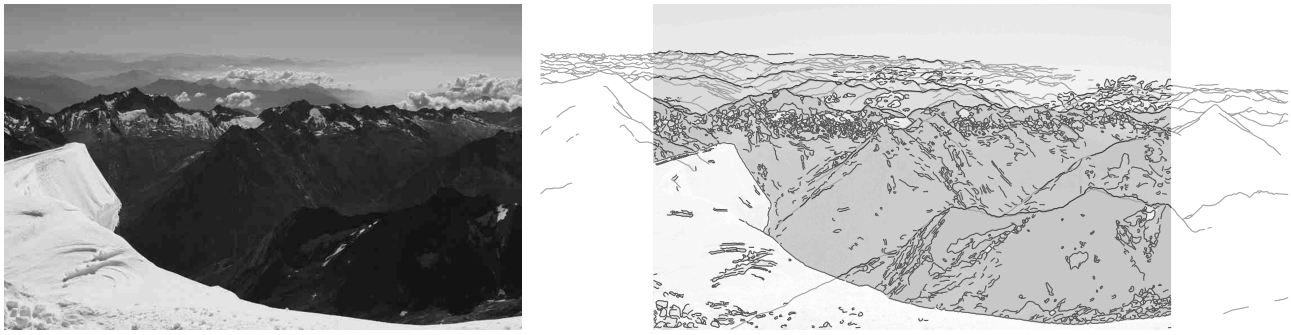


Figure 3. Types of edges detected in mountain scenes: silhouettes encoded (blue) or not encoded in the terrain model (red), noise and non-silhouette edges (green). Reference (*i.e.* synthetic) silhouettes (gray) are not always detected.

Our main observation relates to the topology of silhouette-maps: a feasible silhouette map in general configuration can contain T-junctions, but no crossings. Crossings appear only in singular views, when two distinct silhouette edges align (Fig. 4). Consequently, a curve detected as an edge in the photograph, even if not silhouette, usually follows a feature of some object and thus never crosses a silhouette. This only happens if some object, not encoded in the terrain model, occludes it. The probability for such events remains low, which will render the method more robust despite potentially low-quality edges.

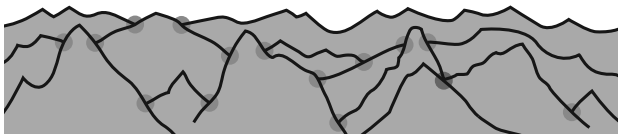


Figure 4. Specific topology of terrain silhouettes: separate edges meet with T-junctions (green), crossings (red) are singular.

To evaluate the likelihood of a given orientation g, the two edge sets (from photo and model) are overlayed according to g. Each edge e_p from $\mathcal{E}_P$ is considered independently and tested against $\mathcal{E}_F$. To account for noise, any potential matching with an edge e_f must be scanned within some tolerance ε_e. When e_p enters the ε_e-neighborhood of an edge $e_f \in \mathcal{E}_F$, four distinct cases can happen, as depicted by Fig 5. A threshold ℓ_{fit} is used to distinguish the case where e_p is following e_f from the case where it crosses it.

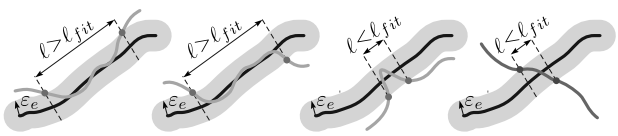


Figure 5. The four possible situations for edge-to-edge matching.

For a single edge e_p, the matching likelihood value is computed as follows:

- parts where e_p stays outside the ε_e-neighborhood of elements of $\mathcal{E}_F$ count as 0;
- if e_p enters the ε_e-neighborhood of an element $e_f \in \mathcal{E}_F$ and exits it after traversing over a length ℓ:
  - if it exits on the same side or if $\ell \geqslant \ell_{fit}$, the fitting energy $\ell^{a_{fit}}$ is added;
  - else, a constant penalty cost c_{cross} is subtracted.

The non-linearity implied by an exponent $a_{fit} > 1$ increases robustness: long matching edges will receive more strength than sets of small disconnected segments of the same total length. Finally, the matching likelihood for $\mathcal{E}_P$ under the candidate rotation g is obtained by summing the values of each of the (accordingly displaced) individual edges of $\mathcal{E}_P$.

In practice the computation is performed as follows: first, $\mathcal{E}_F$ is rasterized with a thickness ε_e into a sufficiently high-resolution spherical image; second, the $\mathcal{E}_P$ edges are warped according to g, traversed and tested against the rasterized $\mathcal{E}_F$ for potential intersections. The cost of this simple approach is $O(mn)$, where m is the resolution of the rasterized $\mathcal{E}_F$ and n the total number of segments of $\mathcal{E}_P$.

Interestingly, the metric relies on all the information available in the detected edges: even non-silhouette edges help to find the correct match by preventing actual silhouette edges from crossing them (Fig. 6). Therefore it would theoretically be possible to find the correct matches even if all silhouette edges were missed.

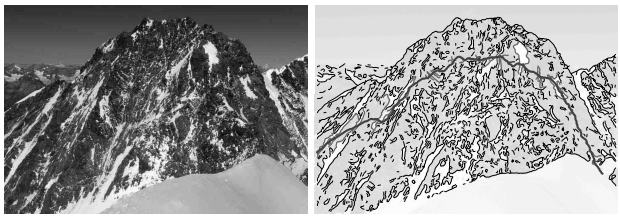


Figure 6. Detected non-silhouette edges also help the matching process: a pose of a reference silhouette (blue) is prevented if it crosses many detected edges (red).

Although this metric allows robust matching (see Section 7), it requires a dense 3D sampling of $SO(3)$, leading to

prohibitive computation times. We avoid this problem with an effective search space reduction preprocess, presented now.

5. Spherical cross-correlation for search space reduction

To address the problem of the high cost implied by a dense sampling of $SO(3)$, we move to the Fourier domain. It is well known that the cross-correlation between two $n \times n$ images can be computed in $O(n^2 \log n)$ using the fast fourier transform (FFT). This has recently been extended to spherical images [11]. The spherical cross-correlation of two complex-valued spherical functions f and p is defined on $SO(3)$ as:

$$\forall g \in SO(3), \quad f \star p\ (g) = \int_{\mathcal{S}^2} f(\omega)\overline{p(g^{-1}\omega)}d\omega,$$

and can be evaluated in $O(n^3 log(n))$ for n^2-sampled spherical functions via FFT algorithms on $SO(3)$ [11].

We could directly apply this to our problem by sampling the two silhouette-maps on the sphere and computing the cross-correlation of these two binary-valued maps (1 on edges, 0 elsewhere). The main problem here is that it completely disregards the relative orientation of edges. With our noise-prone detected edge-maps, the maximum cross-correlation value would be found where most edges overlap, which would only work if the detected edge-map contained all and only the silhouette edges.

5.1. Angular similarity operator

Our goal is to integrate edge orientations in the cross-correlation. The orientation information can be kept by rasterizing $\mathcal{E}_F$ as a 2D real-valued vector field $\mathbf{f}(\omega) = (f_x(\omega), f_y(\omega))$, being the tangent vectors of the edges where they appear, and zero elsewhere (Fig. 8). We define the *angular similarity operator* $\mathcal{M}(\mathbf{f}, \mathbf{p})$ as follows:

$$\mathcal{M}(\mathbf{f}, \mathbf{p}) = \rho_f^2 \rho_p^2 \cos 2(\theta_f - \theta_p),$$

where (ρ_f, θ_f) and (ρ_p, θ_p) are the polar representations of $\mathbf{f}$ and $\mathbf{p}$ (see Fig. 7). The value produced by this operator is:

1. positive for (close to) parallel vectors,
2. negative for (close to) orthogonal vectors,
3. zero if one of the vector is zero.

The matching likelihood between two spherical functions $\mathbf{f}$ and $\mathbf{p}$ can be expressed as:

$$\int_{\mathcal{S}^2} \mathcal{M}(\mathbf{f}(\omega), \mathbf{p}(\omega))d\omega,$$

so that values of ω where edges closely match are counted positively while those where edges cross almost perpendicularly are counted negatively. Furthermore, values of ω where either $\mathbf{f}$ or $\mathbf{p}$ has no edge do not affect the integral.

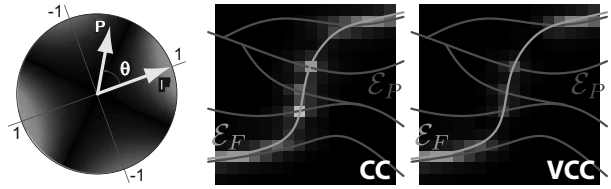


Figure 7. Left: $\mathcal{M}(\mathbf{f}, \mathbf{p})$ as a function of $\mathbf{p}$ (for a fixed $\mathbf{f}$). Classical cross-correlation (CC) disregards orientations, wereas our vector-field cross-correlation (VCC) properly penalizes crossings.

5.2. Spherical 2D-vector fields cross correlation

In order to be used as a matching likelihood estimation, this integral would need to be evaluated for any candidate rotation g, by rotating $\mathbf{p}$ accordingly. However, now that $\mathbf{p}$ values are vectors, we need to take the effect of the rotation into account. Because we defined the transformation of the camera relative to the world frame, we can show that the expression of $\mathbf{p}$ under a rotation $g = (\alpha, \beta, \gamma)$ is:

$$R_{\gamma+\frac{\pi}{2}}.\mathbf{p}(g^{-1}w) \quad \text{with} \quad R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

The formula stems from the fact that in the ZYZ euler angles parametrization we are using, the γ angle corresponds to the rotation of the camera around its viewing direction (the $\frac{\pi}{2}$ offset reflects that a horizontally-looking camera with a zero γ value is tilted by $\frac{\pi}{2}$).

Our operator needs to be modified as follows to take the rotation of $\mathbf{p}$ into account:

$$\mathcal{M}_g(\mathbf{f}, \mathbf{p}) = \rho_f^2 \rho_p^2 \cos 2(\theta_f - (\theta_p + \gamma + \frac{\pi}{2})),$$

and for a candidate rotation g we then define the matching likelihood between $\mathbf{f}$ and $\mathbf{p}$ as follows:

$$\text{VCC}(\mathbf{f}, \mathbf{p})(g) = \int_{\mathcal{S}^2} \mathcal{M}_g(\mathbf{f}(\omega), \mathbf{p}(g^{-1}\omega))d\omega.$$

5.3. Efficient computation

Using the representation of 2D vectors as complex numbers, VCC can be expressed as one spherical cross-correlation operation. Indeed, $\mathcal{M}(\mathbf{f}, \mathbf{p})$ can be rewritten as follows,

$$\mathcal{M}(\mathbf{f}, \mathbf{p}) = Re\left\{\hat{f}^2\overline{\hat{p}^2}\right\},$$

where

$$\hat{f} = \rho_f e^{i\theta_f} \quad \text{and} \quad \hat{p} = \rho_p e^{i\theta_p}.$$

This leads to the following VCC formulation:

$$\begin{aligned}\text{VCC}(f,p)(g) &= Re\left\{\int_{\mathcal{S}^2} \hat{f}^2(\omega)\overline{\left(e^{i(\gamma+\frac{\pi}{2})}\hat{p}(g^{-1}\omega)\right)^2}d\omega\right\} \\ &= -Re\left\{e^{-i2\gamma}\ \hat{f}^2 \star \hat{p}^2\ (g)\right\}.\end{aligned}$$

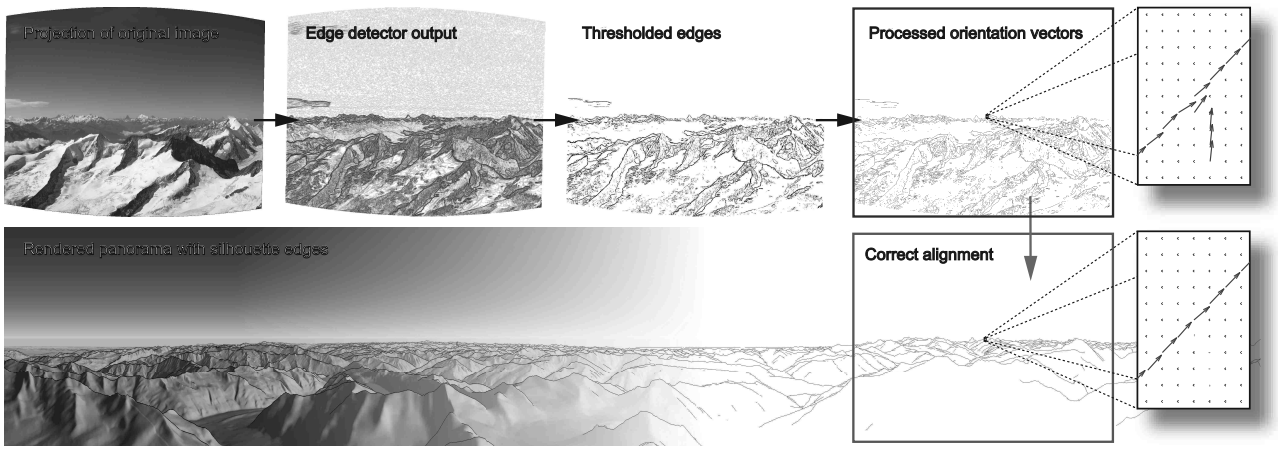


Figure 8. Detection and processing of edges into orientation vectors (blue frame), used to find the optimal registration (red frame).

In other words, we expressed the computation as a cross-correlation between $\hat{f}^2$ and $\hat{p}^2$, that is weighted by $-e^{-i2\gamma}$ and reduced to its real part. The dominant cost of the matching space reduction is therefore the cross-correlation computation, *i.e.* $O(n^3 \log n)$.

6. Implementation details

Terrain model We experimented with two terrain datasets: 1) coverage of the Alps with 24 meters spaced samples (http://www.viewfinderpanoramas.org); 2) National Elevation Dataset (USGS, http://ned.usgs.gov), covering the United States at thrice bigger resolution. Experiments showed the importance of considering the Earth's curvature when rendering the synthetic panoramas.

Image processing The input photograph is first remapped to a rectilinearly projected RGB image with known FOV, using the camera's intrinsic parameters (read from the attached EXIF data, assisted by a camera database if necessary). We then apply the *compass* edge detector [15], parameterized by a radius σ, producing separate maps for edge strengths (Fig 3) and orientations, that are easily combined into a vector field of tangent vectors (Fig. 8). This edge detector has the particularity of fully exploiting the color information, unlike classical ones that handle only grayscale images. The result is then thresholded (parameter τ) to keep only significant edge. The edge map $\mathcal{E}_P$ (a set of vectorized lines) is finally extracted by thinning [12] and vectorization [4]. The following parameters were used without further need of dynamic adaptation: $\sigma = 1, \tau = 0.7$.

Panorama processing Generating silhouettes from the 3D terrain data is a classical computer graphics problem for which several options exist. Exploiting the GPU, we apply raycasting to render the silhouettes into a 2D cylindrical image, which is then vectorized into an edge map $\mathcal{E}_F$.

Efficient matching Because $SO(3)$ has three dimensions, the robust matching metric still needs to be evaluated on many sampled rotation candidates, even after the search space reduction process. Nonetheless, each evaluation being independent, the overall process is highly parallelizable making a GPU mapping possible that cuts down the computation time from several hours to a few seconds.

7. Results

Our approach was implemented on a Dell T7500 workstation equipped with two six-core Intel Xeon processors, one GeForce GTX 480 GPU, and 23GB RAM. With our simple implementation, the overall process takes around 2 minutes, critical parts being compass edge detection (around 1 min.), spherical cross-correlation (less than one minute, with sampling bandwidths of 1024 for $\mathcal{S}^2$ and 256 for $SO(3)$) and final matching metric evaluation (around 20 s. with the GPU implementation). Of a collection containing 28 photographs randomly chosen from Flickr, 86% were correctly aligned by our technique (interestingly, VCC was already maximized at the correct orientation for 25% of the tested examples). We examined two different mountainous regions (Alps in Europe and Rocky Mountains in USA) and found that our approach performs similarly. The matching is generally very accurate, *i.e.* below $0.2°$ (Fig. 1, 9 and 10). Small deviations mostly correspond to imperfections of the 3D model. Experimentally, an accuracy below a few hundred meters for the viewpoint is sufficient.

7.1. Applications

Annotations Our solution enables us to mark a certain peak in all given photos if it is visible. This is a difficult and tedious task that often can only be performed by experts.

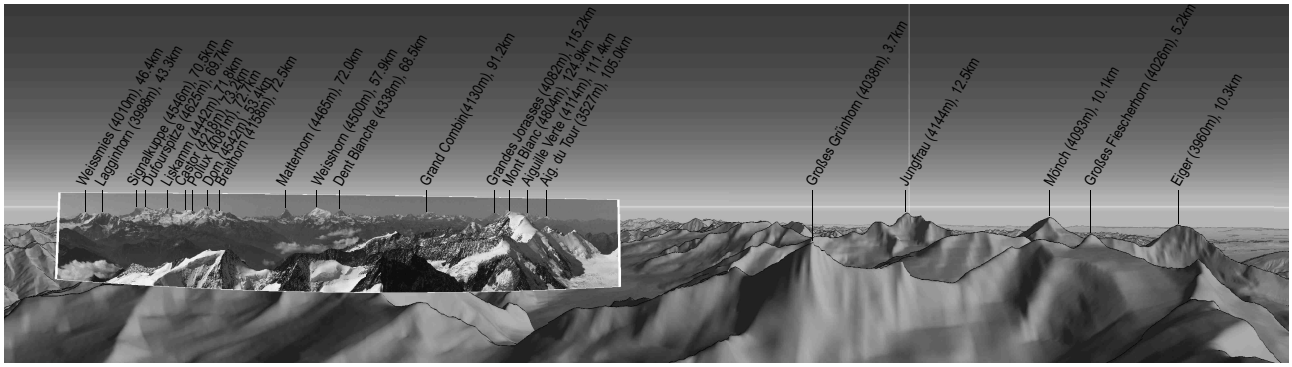


Figure 9. An example of annotated panorama image superposed on synthetic panorama.

By testing the visibility of the corresponding mountain in the 3D terrain model, we can easily decide what part of it shows in the photograph, and how far it is from the camera position. Some results are illustrated in Fig. 9 and 10.

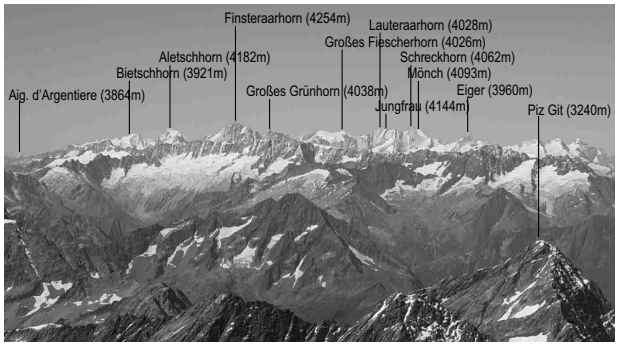


Figure 10. Annotated photo created using the proposed technique.

Augmented reality We can also achieve augmented views of the mountain landscape. Here, we add paths, landmarks and other 3D objects into the 3D terrain model. By transferring only the visible pixels of these models, we can add them into the photograph. Furthermore, we can relight them according to the shot. For this, we can either rely on the time stamp of the photo to deduce the position of the sun and weather conditions from according databases. Alternatively, we can optimize the sun position and illumination by comparing the lit terrain model to the captured photo. We rely on a simple model with a point light (sun) and ambient occlusion (sky). The optimization process is 1D and converges quickly.

Texture transfer Using our approach, photo collections can also easily be used to transfer texture information into a 3D mountain model such as those of Google Earth. Having found the corresponding camera view, it is enough to apply a projective texture mapping (including a shadow map test) to derive which part of the scene was actually visible and could benefit from the image content.

Photo navigation Similarly to photo tourism [17], we can add the photos into the 3D terrain model to enable an intuitive navigation. This allows illustrating or preparing hikes, even when relying on photos of others.

Image Enhancement and Expressive Rendering Using the underlying 3D terrain model, we can enhance an existing image or achieve non-photorealistic effects. E.g., we can perform informed model-based image dehazing (Fig. 11), enhance certain objects, or even mix the view with geological data (*e.g.* using USGS metadata).

Video Matching On a frame-by-frame basis, we can also optimize video sequences. Which is relatively fast because the search space is reasonably reduced by assuming a slow displacement. One could also initialize the search with the frame that gave the highest response in the first search step, but in practice, we found that unnecessary [1].

8. Conclusions and Future Work

We presented a solution to determine the orientation of mountain photographs by exploiting available digital elevation data. Although this is a very challenging task, we showed that our approach delivers a robust and precise result. The accuracy of our solution enabled various interesting applications that we presented in this paper. Our technical contributions, such as the camera pose estimation based on edge-to-silhouette matching could find application in other contexts of more general matching problems.

In the future, we want to explore other cues (*e.g.* the atmospherical scattering, aerial perspective) that might help us in addressing more general environments and improving the edge detection part for these scenarios [22].

[1] Refer to supplemental movie for video matching examples.

Figure 11. Application to image contrast enhancement: the original image (left) is modulated by the diffuse lighting component computed on the synthetic model (particularly profitable for distant mountains, whose contrast is affected by atmospheric effects).

Acknowledgements

Thanks to YouTube user 'towatzek' for his videos (http://www.mounteverest.at), Wikipedia user 'Nholifield' for the Mount Princeton panorama, and ColoradoGuy.com for pictures from Rockies. This work was partially funded by the Intel Visual Computing Institute (Saarland University).

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346 – 359, 2008. Similarity Matching in Computer Vision and Multimedia. 42

[2] F. Cozman and E. Krotkov. Position estimation from outdoor visual landmarks for teleoperation of lunar rovers. In *Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision*, Washington, DC, USA, 1996. IEEE Computer Society. 42

[3] R. Fattal. Single image dehazing. *ACM Trans. Graph.*, 27:72:1–72:9, August 2008. 43

[4] R. C. Gonzalez, R. E. Woods, and S. L. Eddins. *Digital Image Processing Using MATLAB*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2003. 46

[5] R. Grzeszczuk, J. Kosecka, R. Vedantham, and H. Hile. Creating compact architectural models by geo-registering image collections. pages 1718 –1725, sep. 2009. 43

[6] C. Herdtweck and C. Wallraven. Horizon estimation: perceptual and computational experiments. In *APGV '10: Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, pages 49–56, New York, NY, USA, 2010. ACM. 42

[7] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. *ACM Trans. Graph.*, 24(3):577–584, 2005. 42

[8] P. C. N. Jr, M. Mukunoki, M. Minoh, and K. Ikeda. Estimating camera position and orientation from geographical map and mountain image. In *38th Research Meeting of the Pattern Sensing Group, Society of Instrument and Control Engineers*, pages 9–16, 1997. 42

[9] J. Kopf, B. Neubert, B. Chen, M. F. Cohen, D. Cohen-Or, O. Deussen, M. Uyttendaele, and D. Lischinski. Deep photo: Model-based photograph enhancement and viewing. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2008)*, 27(5):116:1–116:10, 2008. 42

[10] J. Kosecka and W. Zhang. Video compass. In *Proceedings of European Conference on Computer Vision*, pages 657 – 673, 2002. 43

[11] P. Kostelec and D. Rockmore. Ffts on the rotation group. *Journal of Fourier Analysis and Applications*, 14:145–179, 2008. 10.1007/s00041-008-9013-5. 45

[12] L. Lam, S.-W. Lee, and C. Y. Suen. Thinning methodologies-a comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14:869–885, September 1992. 46

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 10.1023/B:VISI.0000029664.99615.94. 42

[14] S. Ramalingam, S. Bouaziz, P. Sturm, and M. Brand. Geolocalization using skylines from omni-images. In *Proceedings of the IEEE Workshop on Search in 3D and Video, Kyoto, Japan*, oct 2009. 42

[15] M. A. Ruzon and C. Tomasi. Edge, junction, and corner detection using color distributions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1281–1295, 2001. 46

[16] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. and Mach. Intell.*, 31:824–840, 2009. 42

[17] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3d. In *SIGGRAPH Conference Proceedings*, pages 835–846, New York, NY, USA, 2006. ACM Press. 42, 47

[18] F. Stein and G. Medioni. Map-based localization using the panoramic horizon. In *Robotics and Automation, IEEE Transactions on*, pages 892 – 896, 1995. 42

[19] R. Szeliski. Image alignment and stitching: a tutorial. *Found. Trends. Comput. Graph. Vis.*, 2(1):1–104, 2006. 42

[20] R. Talluri and J. K. Aggarwal. Handbook of pattern recognition & computer vision. chapter Position estimation techniques for an autonomous mobile robot: a review, pages 769–801. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 1993. 42

[21] J. Woo, K. Son, T. Li, G. S. Kim, and I.-S. Kweon. Vision-based uav navigation in mountain area. In *MVA*, pages 236–239, 2007. 42

[22] C. Zhou and B. W. Mel. Cue combination and color edge detection in natural scenes. *Journal of vision*, 8(4), 2008. 47

[23] B. Zitová and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977 – 1000, 2003. 42

[24] L. Zollei, J. Fisher, and W. Wells. An introduction to statistical methods of medical image registration. In *Handbook of Mathematical Models in Computer Vision*. Springer, 2005. 42