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

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.

© 2008 Elsevier Ltd. All rights reserved.

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).

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 natural-ness 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 psy-chophysical 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, color-fulness and overall accuracy. The results showed that bilateral filtering [9] generated more accurate results than other algo-rithms. 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 statisti-cally 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 experi-ments, 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 informa-tion, 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 recon-struction 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. Encour-aged 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 appear-ance. 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 con-venient 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.
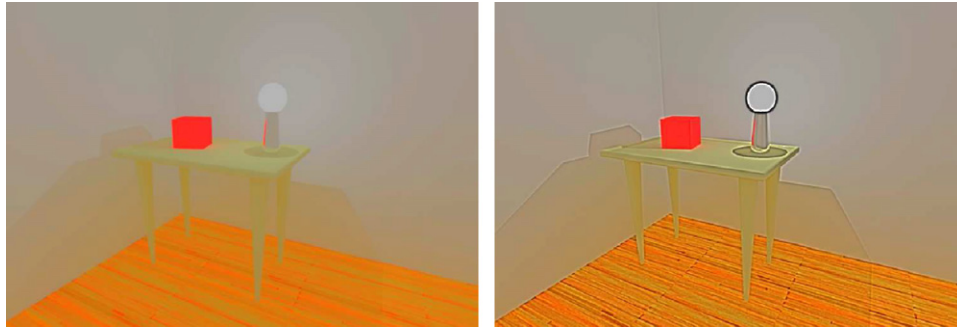
Fig. 4. Halo artifacts and oversaturation. Left: HDR image after successful tone mapping without halo artifacts. Right: the same image after tone mapping using the local method exhibiting a massive amount of halo artifacts. Both images exhibit oversaturation.

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
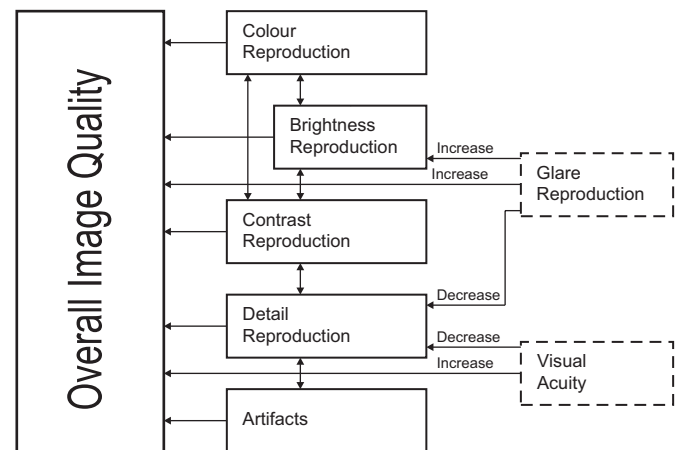

Fig. 6. The relationships between image attributes. The attributes we did not evaluate in subjective perceptual experiments are in dashed boxes.

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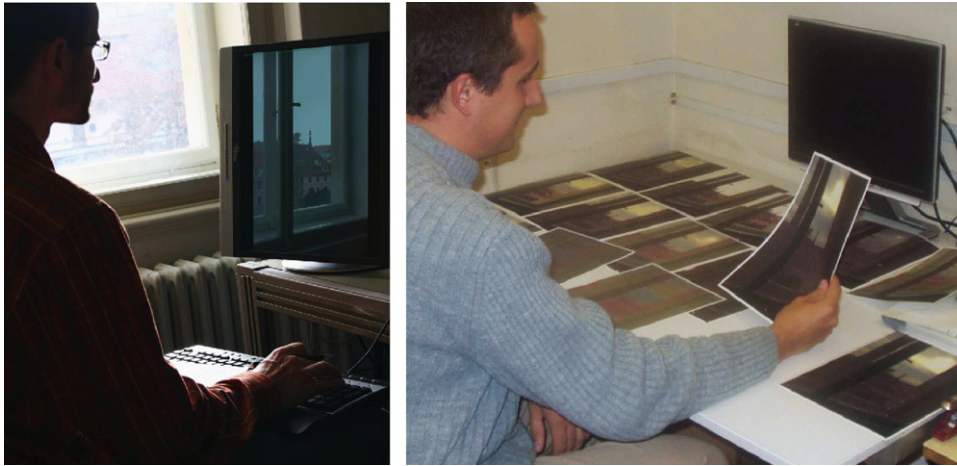
Fig. 7. Example of subjective perceptual experiments setups. Left: rating experiment with real references, Right: ranking experiment without references.

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

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

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).

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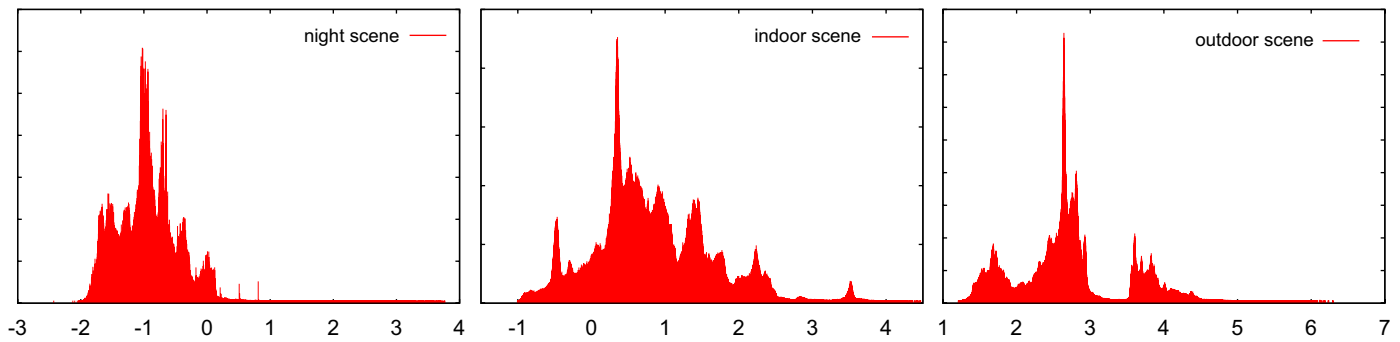
Fig. 8. Luminance histograms ($\log_{10}$) of the experimental HDR images, from left: night scene, indoor scene, outdoor scene.

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 $3_{(scenes)} \cdot (10 + 10)_{(subjects)} \cdot 6_{(attributes)} \cdot 14_{(methods)} = 5040$ values of observation scores.
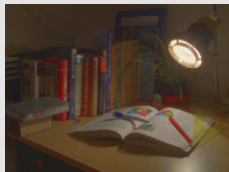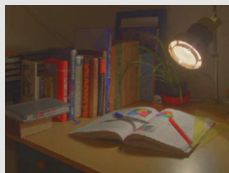
In the first experiment, based on *rating* (see Fig. 7—left), we simultaneously presented an original (*real-world*) HDR scene and the appropriate TM images of this scene to human observers. In order to keep the illumination moderately constant, we performed all the testing procedures at the same time 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
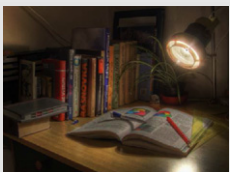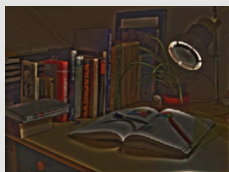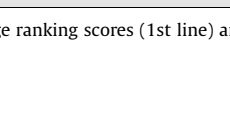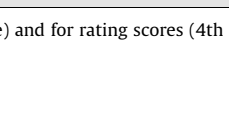
experiment and they were instructed to "Rate the images on how close the particular image attribute matches in appearance to the real-world scene" (attribute reproduction results) and to "Rate the images on how close the overall match in appearance is to the real-world scene" (overall image quality results). 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
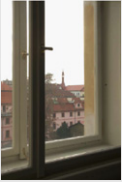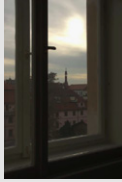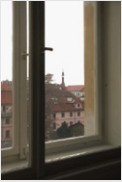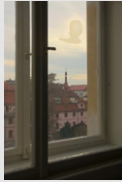
---
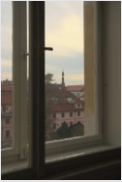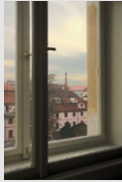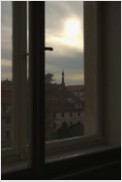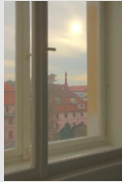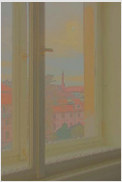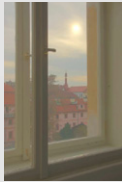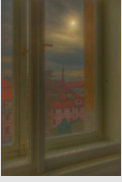
[3] FSC P19-2, 19″ LCD display, with maximum luminance of 280 cd/m². We used manufacturer's ICC profiles (D65) for both the monitor and the camera to perform the colorimetric characterization of the devices.

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

Method	Image	Brightness	Contrast	Details	Colors	Overall quality
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

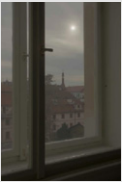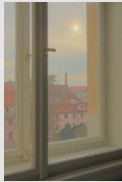| Method	Image	Brightness	Contrast	Colors	Details	Overall quality
LCIS99		**4.1** / 1.5	**6.2** / 2.6	**5.4** / 3.9	**3.4** / 1.2	**4.6** / 1.3
		11.2 / 2.1	**10.8** / 1.9	**9.8** / 3.3	**9.3** / 3.3	**11.9** / 2.4
Pattanaik02		**11.1** / 3.6	**8.9** / 3.6	**12.4** / 2.5	**8.6** / 3.3	**6.8** / 3.3
		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	Method	Image	Brightness	Contrast	Colors	Details	Overall quality
Linear Clip		**12.3** / *2.2*	**13.2** / *0.9*	**12.5** / *1.9*	**13.4** / *0.5*	**13.2** / *0.9*	LCIS99		**9.4** / *2.2*	**7.9** / *1.6*	**8.8** / *1.5*	**7.9** / *2.0*	**7.8** / *1.7*
		10.3 / *3.1*	**10.7** / *2.9*	**9.4** / *3.6*	**10.1** / *3.6*	**11.8** / *2.5*			**7.8** / *3.3*	**7.8** / *3.2*	**8.7** / *2.6*	**8.5** / *2.6*	**7.7** / *3.0*
Ward94		**4.3** / *2.0*	**6.2** / *2.6*	**4.1** / *3.1*	**5.7** / *1.4*	**8.2** / *1.8*	Pattanaik02		**6.1** / *4.1*	**4.1** / *1.8*	**3.4** / *2.2*	**2.1** / *0.3*	**2.1** / *0.3*
		7.4 / *3.1*	**10.2** / *2.3*	**7.2** / *2.3*	**9.1** / *2.3*	**9.7** / *2.2*			**2.3** / *0.8*	**2.1** / *0.8*	**4.8** / *3.4*	**3.1** / *2.1*	**2.3** / *0.7*
Tumblin99		**12.4** / *2.3*	**12.7** / *1.2*	**12.5** / *1.9*	**13.6** / *0.5*	**12.9** / *1.0*	Choudhury03		**8.2** / *1.7*	**5.8** / *1.3*	**7.9** / *1.4*	**7.3** / *1.7*	**6.0** / *1.5*
		10.3 / *2.5*	**8.4** / *2.1*	**7.3** / *4.1*	**9.8** / *3.4*	**11.2** / *1.5*			**6.0** / *2.4*	**6.1** / *3.2*	**7.8** / *3.1*	**7.1** / *2.4*	**7.7** / *3.0*
Reinhard02		**9.2** / *2.0*	**10.6** / *0.9*	**9.0** / *1.6*	**10.1** / *1.1*	**7.9** / *3.2*	Drago03		**3.6** / *2.1*	**4.1** / *2.4*	**5.1** / *2.2*	**5.4** / *2.5*	**6.9** / *2.7*
		13.1 / *1.1*	**12.0** / *1.7*	**9.7** / *4.6*	**11.7** / *1.9*	**9.7** / *2.9*			**6.2** / *2.3*	**6.8** / *2.3*	**6.1** / *1.7*	**5.0** / *3.6*	**5.7** / *1.5*
Schlick94		**9.4** / *2.7*	**10.6** / *1.6*	**10.7** / *2.6*	**10.9** / *0.5*	**11.5** / *1.4*	Ashikhmin02		**8.7** / *2.5*	**6.9** / *1.8*	**7.7** / *2.2*	**6.8** / *1.5*	**4.9** / *1.3*
		10.6 / *2.0*	**10.3** / *2.6*	**9.7** / *2.7*	**9.8** / *2.2*	**11.1** / *2.3*			**6.6** / *2.4*	**6.0** / *2.1*	**7.9** / *4.1*	**7.1** / *2.3*	**5.2** / *2.5*
Ward97		**10.6** / *2.1*	**11.6** / *2.3*	**12.3** / *1.1*	**11.4** / *0.8*	**11.3** / *1.8*	Fattal02		**2.8** / *1.3*	**1.8** / *1.1*	**3.1** / *1.4*	**3.7** / *1.9*	**3.5** / *1.2*
		9.7 / *2.9*	**9.0** / *3.7*	**8.9** / *3.4*	**8.7** / *2.5*	**8.9** / *1.5*			**4.4** / *4.1*	**2.9** / *0.9*	**5.5** / *4.7*	**2.0** / *0.7*	**2.3** / *0.9*
Durand02		**3.6** / *2.0*	**6.0** / *2.9*	**4.8** / *2.6*	**5.6** / *1.3*	**7.7** / *1.6*	Chiu93		**4.4** / *3.8*	**3.5** / *3.0*	**3.1** / *2.3*	**1.1** / *0.3*	**0.3** / *1.1*
		7.8 / *3.4*	**10.4** / *2.5*	**7.0** / *2.0*	**9.0** / *2.4*	**10.1** / *2.7*			**2.9** / *2.2*	**2.6** / *2.6*	**5.3** / *5.1*	**4.4** / *3.8*	**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*

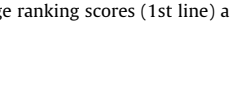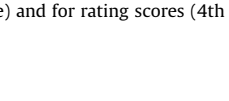| Method	Image	Brightness	Contrast	Details	Colors	Overall quality	Method	Image	Brightness	Contrast	Colors	Details	Overall quality
Linear Clip		**11.3**	**12.8**	**13.2**	**12.2**	**12.9**	LCIS99		**6.5**	**6.2**	**5.8**	**6.1**	**5.7**
		3.6	*1.2*	*1.3*	*2.7*	*1.0*			*2.3*	*1.5*	*0.9*	*2.7*	*0.5*
		10.0	**9.8**	**8.8**	**10.4**	**10.2**			**7.6**	**6.9**	**7.5**	**6.5**	**7.7**
		3.0	*2.1*	*3.3*	*2.9*	*3.1*			*2.2*	*2.9*	*2.1*	*2.0*	*2.6*
Ward94		**10.6**	**12.1**	**12.1**	**11.9**	**12.5**	Pattanaik02		**9.1**	**10.0**	**10.5**	**9.6**	**9.2**
		3.0	*1.8*	*1.5*	*2.2*	*1.4*			*2.0*	*1.0*	*1.8*	*2.5*	*1.6*
		10.2	**11.1**	**9.6**	**11.5**	**11.9**			**10.4**	**12.2**	**12.7**	**11.7**	**12.5**
		3.2	*3.2*	*3.9*	*3.0*	*1.7*			*3.9*	*1.9*	*1.3*	*2.0*	*1.8*
Tumblin99		**7.4**	**7.6**	**8.1**	**8.4**	**8.8**	Choudhury03		**7.1**	**6.6**	**5.3**	**5.9**	**5.1**
		2.3	*1.7*	*1.0*	*1.5*	*1.3*			*2.6*	*2.3*	*0.9*	*2.6*	*1.0*
		8.8	**8.6**	**6.8**	**8.4**	**9.4**			**6.7**	**5.6**	**7.3**	**6.5**	**6.9**
		2.7	*3.1*	*2.9*	*3.4*	*2.7*			*1.9*	*3.4*	*2.9*	*2.8*	*2.5*
Reinhard02		**9.1**	**9.0**	**9.4**	**9.7**	**9.3**	Drago03		**4.9**	**4.9**	**3.3**	**3.8**	**3.4**
		3.6	*2.5*	*1.3*	*1.4*	*1.3*			*4.2*	*3.9*	*0.6*	*2.6*	*1.2*
		6.8	**8.5**	**9.8**	**10.0**	**8.7**			**6.9**	**4.9**	**5.2**	**3.9**	**4.4**
		4.2	*2.4*	*3.7*	*2.3*	*2.6*			*3.9*	*2.5*	*3.0*	*3.2*	*1.9*
Schlick94		**8.8**	**9.3**	**9.7**	**9.5**	**10.4**	Ashikhmin02		**5.1**	**3.6**	**3.6**	**4.5**	**3.3**
		2.6	*2.1*	*1.3*	*1.8*	*1.7*			*3.4*	*1.1*	*1.0*	*2.7*	*1.0*
		7.5	**10.9**	**9.5**	**9.3**	**8.8**			**8.2**	**4.8**	**8.5**	**5.2**	**4.8**
		3.1	*1.8*	*2.5*	*1.9*	*1.9*			*3.2*	*2.2*	*4.4*	*2.6*	*2.3*
Ward97		**8.7**	**7.0**	**7.8**	**8.1**	**7.7**	Fattal02		**4.0**	**2.4**	**2.4**	**2.6**	**2.7**
		2.8	*1.8*	*2.0*	*1.4*	*1.3*			*2.5*	*0.5*	*0.7*	*0.5*	*0.6*
		3.7	**6.3**	**6.5**	**6.9**	**5.9**			**5.7**	**3.7**	**3.0**	**2.6**	**2.1**
		2.8	*2.0*	*2.0*	*2.0*	*1.9*			*4.8*	*3.2*	*3.0*	*1.1*	*0.4*
Durand02		**11.4**	**12.5**	**12.8**	**11.7**	**13.0**	Chiu93		**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
		2.5	*1.4*	*0.4*	*2.2*	*0.6*			*0.0*	*0.0*	*0.0*	*0.0*	*0.0*
		8.9	**10.9**	**8.9**	**10.9**	**11.0**			**3.9**	**1.1**	**1.2**	**1.3**	**1.1**
		3.7	*2.5*	*2.9*	*2.2*	*2.4*			*5.0*	*0.2*	*0.3*	*0.6*	*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.
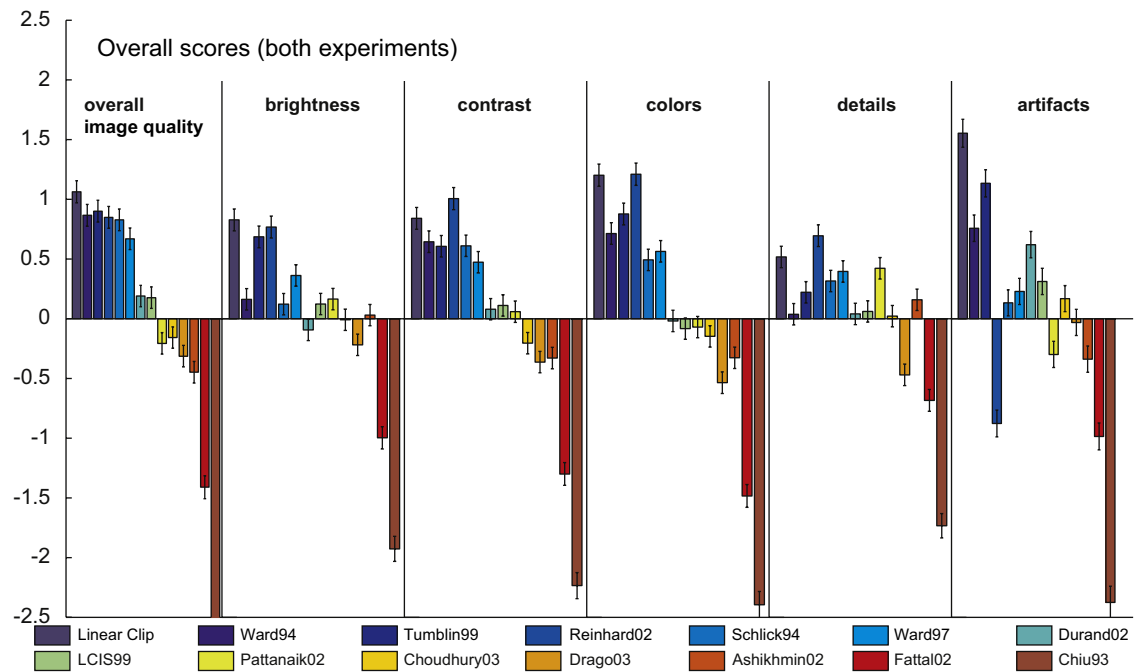
Fig. 9. Overall accuracy scores for all examined TM methods. Left to right: overall perceptual quality, reproduction of brightness, reproduction of contrast, reproduction of details, reproduction of colors, lack of disturbing artifacts. In each chart the higher value represents the higher reproduction 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 $Q_{\text{crit}} = 5.99$

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 $Q_{\text{crit}} = 19.16$

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 ($v = 14 \times 10 \times 3 - 1 = 419$ degrees of freedom) is $\chi^2_{\text{crit}} = 467.73$. 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

---

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

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 $\mathbb{X}_{\text{rat}_i}$ and $\mathbb{X}_{\text{ran}_i}$, where $\mathbb{X}_{\text{rat}_i}$ and $\mathbb{X}_{\text{ran}_i}$ 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: $\mathbb{A}_{\text{rat}_i} = (A_{\text{INDOORrat}_i}, A_{\text{OUTDOORrat}_i}, A_{\text{NIGHTrat}_i})$ and $\mathbb{A}_{\text{ran}_i} = (A_{\text{INDOORran}_i}, A_{\text{OUTDOORran}_i}, A_{\text{NIGHTran}_i})$ where $\mathbb{A}$ denotes a particular image attribute, and $A_{\text{DESK}*_i}, A_{\text{WINDOW}*_i}$, and $A_{\text{NIGHT}*_i}$ 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: $\mathbb{X}_{\text{rat}_i} = (OIQ_{\text{rat}_i}, Bri_{\text{rat}_i}, Con_{\text{rat}_i}, Det_{\text{rat}_i}, Col_{\text{rat}_i}, Art_{\text{rat}_i})$ and $\mathbb{X}_{\text{ran}_i} = (OIQ_{\text{ran}_i}, Bri_{\text{ran}_i}, Con_{\text{ran}_i}, Det_{\text{ran}_i}, Col_{\text{ran}_i}, Art_{\text{ran}_i})$, 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 $V_{\text{crit}}^* = 2.6383$ and the resulting values are $V_{\text{OIQ}}^* = -0.3489$, $V_{\text{Bri}}^* = -0.4791$, $V_{\text{Con}}^* = 0.0565$, $V_{\text{Det}}^* = -0.1409$, $V_{\text{Col}}^* = -0.0404$, $V_{\text{Art}}^* = 0.1727$. 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 $V_{\text{crit}}^* = 2.394$

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.

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 multi-variate 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
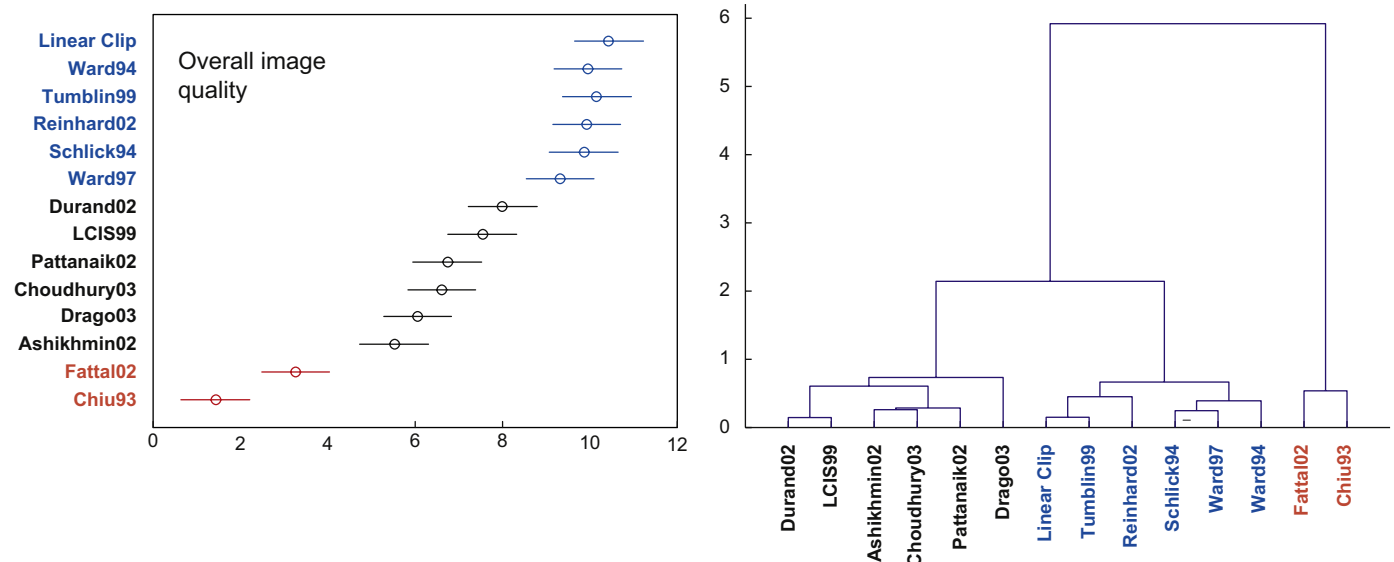

Fig. 10. Left: average overall image quality with confidence intervals. Circles show OIQ means with 95% confidence intervals (horizontal axis)—the higher value the better quality. Right: average Mahalanobis distances of overall image quality for all methods.
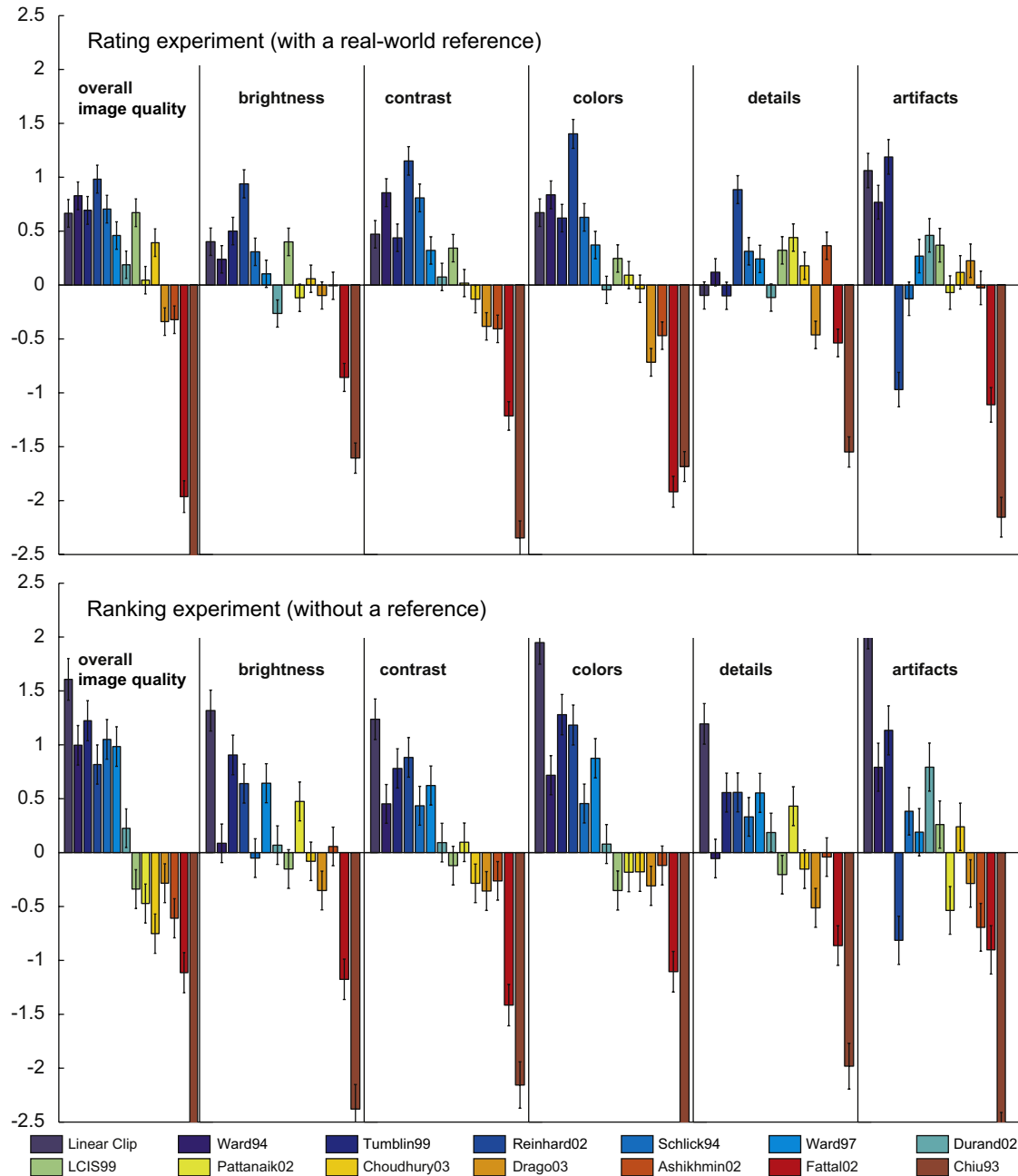
Fig. 11. Accuracy scores for rating (with a reference) experiment (top) and for ranking (without a reference) experiment (bottom) for all examined TM methods. Left to right: overall perceptual quality, reproduction of brightness, reproduction of contrast, reproduction of details, reproduction of colors, 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:

$$OIQ = 0.07 \cdot Bri + 0.37 \cdot Con + 0.06 \cdot Det \\ + 0.36 \cdot Col + 0.21 \cdot Art, \tag{1}$$

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:

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

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-
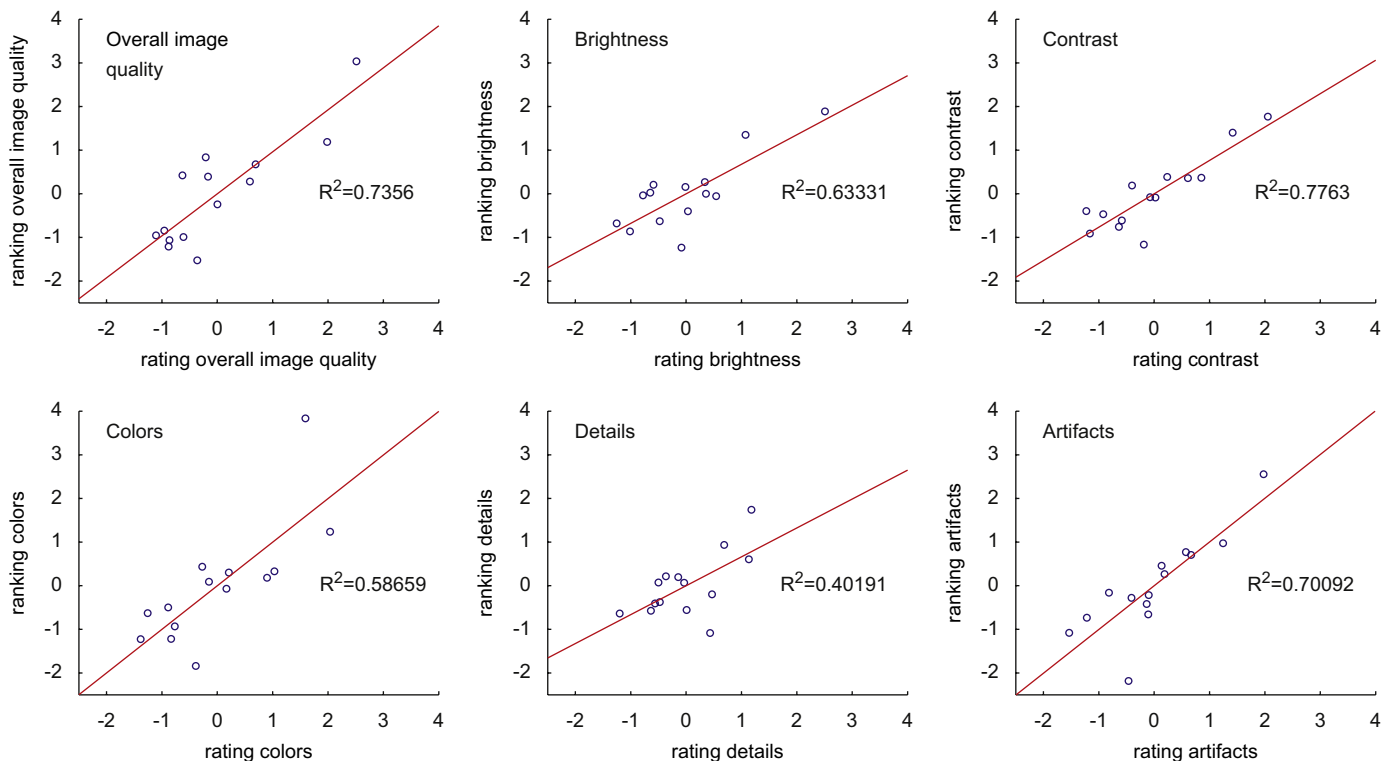

Fig. 12. Relations of the ranking experiment (vertical axes) and rating experiment (horizontal axes) interval scale results for all image attributes.

Overall image quality - rating experiment

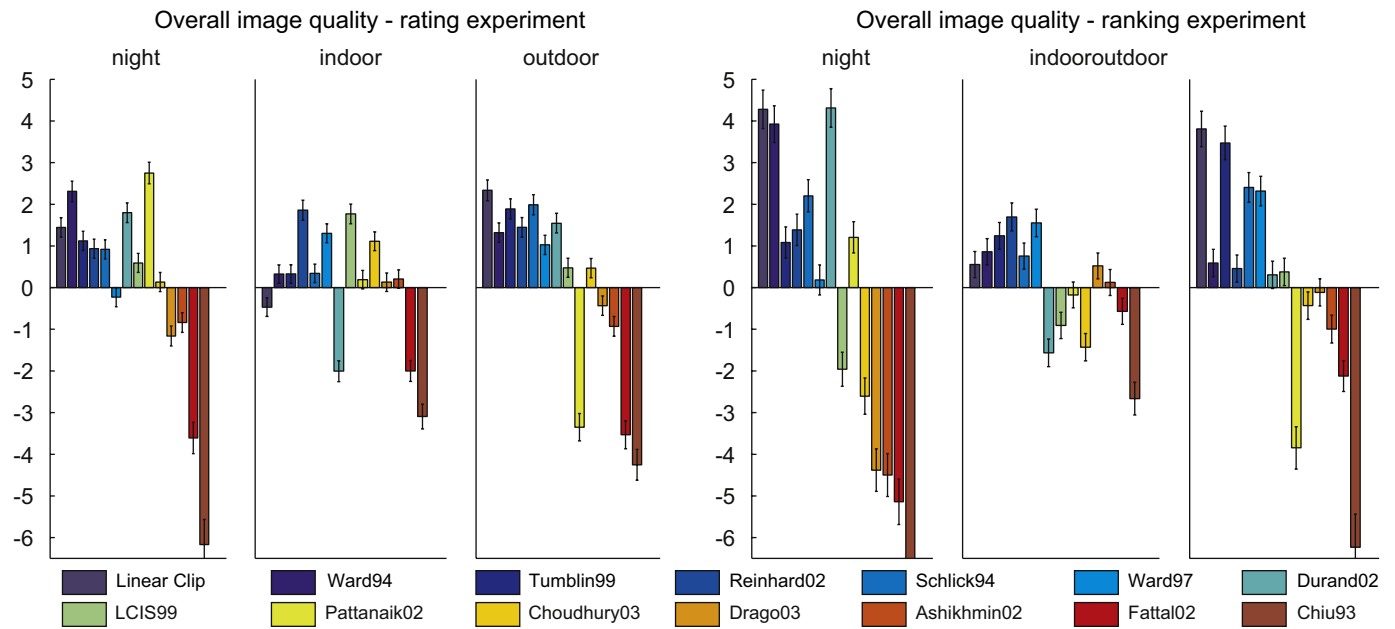## Overall image quality - ranking experiment

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:

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

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

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.

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

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

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

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

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

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

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

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

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

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.