COMPUTER GRAPHICS forum

DOI: 10.1111/j.1467-8659.2008.01176.x

COMPUTER GRAPHICS forum
Volume 28 (2009), number 1 pp. 13–25

A Psychophysical Evaluation of Inverse Tone Mapping Techniques

Francesco Banterle[1], Patrick Ledda[1], Kurt Debattista[1], Marina Bloj[2], Alessandro Artusi[1] and Alan Chalmers[1]

[1]Warwick Digital Laboratory, University of Warwick, Coventry, West Midlands, United Kingdom
[2]Department of Optometry, University of Bradford, United Kingdom

Abstract

In recent years inverse tone mapping techniques have been proposed for enhancing low-dynamic range (LDR) content for a high-dynamic range (HDR) experience on HDR displays, and for image based lighting. In this paper, we present a psychophysical study to evaluate the performance of inverse (reverse) tone mapping algorithms. Some of these techniques are computationally expensive because they need to resolve quantization problems that can occur when expanding an LDR image. Even if they can be implemented efficiently on hardware, the computational cost can still be high. An alternative is to utilize less complex operators; although these may suffer in terms of accuracy. Our study investigates, firstly, if a high level of complexity is needed for inverse tone mapping and, secondly, if a correlation exists between image content and quality. Two main applications have been considered: visualization on an HDR monitor and image-based lighting.

Keywords: inverse tone mapping, tone mapping, high dynamic range imaging

ACM CCS: I.3.3 [Computer Graphics]: Picture/Image Generation and Display Algorithms I.4.0 [Image Processing and Computer Vision]: General Image Displays.

1. Introduction

The field of High-dynamic range (HDR) imaging has taken on an important role in computer graphics and is in constant growth. The majority of the work has focused on displaying HDR content and on the creation of such content from a series of low-dynamic range (LDR) images, see for instance [DM97]. Algorithms for the display of HDR images, known as tone mapping operators (TMOs), focus on the compression of the contrast ratio in order to display the resulting images on typical display devices. This, in particular, has been one of the most active areas of research since the widely available viewing displays can only handle LDR imagery. Several papers have been published in the last two decades and an overview of some of these methods is presented in Reinhard *et al.* [RWPD05].

Although HDR imaging is becoming increasingly popular in areas such as games, visual effects and to some extent digital photography, the vast majority of digital content is still limited in luminance range. Digital cameras able to capture multiple exposures, are becoming available to the average consumer, nevertheless, most of current photography is still captured with single exposures. Furthermore, HDR capture might still be problematic in the case of fast moving subjects or where a tripod might be essential. The growing desire to recreate the luminance range of the original content has also led to the development of HDR display devices, see [SHS*04]. Although such displays are still mostly prototypes, there is reason to believe that they will become widely available in the near future.

HDR images have also been used as a form of illumination in order to render computer-generated scenes. Such techniques, known as image-based lighting (IBL), see [Deb98], are important especially in those applications where computer-generated objects are to be integrated in real environments. IBL relies on a reasonably accurate representation

Submitted September 2007
Revised April 2008
Accepted July 2008

of the luminance range, unfortunately, as we mentioned above, it is not always possible to capture such data.

In recent years, algorithms for expanding LDR content to HDR content have been proposed. These algorithms are known as inverse (or reverse) tone mapping operators (iTMOs) and have two main applications: the visualization of images on HDR displays and IBL. Several algorithms have been published, however, it is difficult to establish which methods might be more appropriate for the different applications.

In this paper, we present the results of two psychophysical experiments to evaluate the performance of five iTMOs. In the first experiment, we investigated the similarity of images produced by such models compared to the actual HDR image they were attempting to portray. In the second, the quality of the resultant IBL was assessed. The images generated for both experiments were displayed on an HDR display and presented to a sizeable group of observers.

2. Previous Work

Inverse tone mapping algorithms can broadly be classified into two groups. The first group contains those methods that focus on HDR expansion of single-exposure captured LDR images whereas the other family of operators generate HDR content from the expansion of tone mapped images. In this paper we concentrated on the evaluation of the former as the latter are related to compression of HDR images and videos, for more details see [LSA05; MEMS06, Hat06, OA06].

Only a handful of iTMOs have been proposed to date, most of which we have included in our experiment. One of the first publications on inverse tone mapping was [Lan02], in which an LDRI was expanded using an exponential function for pixel values above a certain threshold. The main problem with this approach is the lack of any methods to avoid discontinuities to attenuate noise during the expansion. These problems can sometimes lead to visible artefacts in the generated images, an aspect which was not taken in account when it was developed for IBL.

A straightforward method to expand single exposure LDRIs for creating specular highlights was presented in [MDS06, MDS07]. Two linear tone scales with thresholding were applied to LDRIs, and discontinuities were solved using Gaussian convolved threshold maps. Psychophysical experiments with the Dolby DR-37P HDR Monitor [Dol05] were carried out in order to validate the images generated by the algorithm.

A more general method for expanding the range of single exposure LDRIs was presented in [BLDC06]. They proposed an iTMO based on the Global Photographic Tone Reproduction Operator (GPTR) [RSSF02]. To avoid quantization problems due to expansion, they introduced an expansion map, which represents the density of light sources derived from sampling the image using the median cut algorithm [Deb05]. Unfortunately, manual selection of some parameters is required in order to generate the map. Furthermore, due to the nature of the median cut algorithm, flickering may occur if applied to a series of frames, limiting this approach to still images only.

A similar, but computationally more efficient, method for performing LDR expansion in real-time for high-definition content was presented in [RTS*07]. Instead of the density estimation of the previous approach, a large Gaussian convolution of pixels over a certain threshold is adopted. An edge-stop function is applied to the expansion map created with Gaussian convolution to avoid expansion discontinuities . Finally, the image is expanded using a linear expansion according to the expansion map.

A series of psychophysical evaluations using the Dolby DR-37P HDR display were recently presented in [AFR*07]. From a perceptual point of view, the conclusion of these experiments was that their proposed linear scaling of LDRIs gives an HDR experience that can equal or even surpass the appearance of a true HDRI. However, a simple linear scale does not take into account quantization problems that can occur when the range of an LDR image is stretched.

Finally, [WWZ*07] proposed a system for adding HDR details to the over-exposed and under-exposed regions of an LDRI. This system transfers texture details from a correct exposed patch to over-exposed and under-exposed patches, followed by an expansion. They termed this method hallucination. The main drawback of this system is the possibility of not finding a patch for transferring detail. Furthermore, due to the manual interaction that this method requires, it only applies to static images.

Other algorithms related to inverse tone mapping such as Farid [Far01] and Lin and Zhang [LZ05], estimate the value of the gamma function applied to an image or more in general the camera response function from a single image. These techniques are usually applied in inverse tone mapping by linearizing the signal at the beginning of the boosting process.

2.1. Perceptual Experiments

In the last few years, the evaluation and validation of images produced by TMOs has become an important issue. Graphics researchers have proposed methodologies and experiments to assess the quality of various algorithms by running either objective or subjective experiments. We discuss some of these methods as they are similar in nature to our work.

Drago *et al.* [DMMS03] published the first specific work on tone mapping evaluation. They asked subjects to compare

tone mapped images and to make preference judgements on how perceptually similar or dissimilar the images were. Based on these preference results, they determined a theoretical preference point which allowed them to ascertain which operators were perceptually perceived as the most similar to this preference point.

A similar experiment, was presented in Kuang *et al.* [KYJF04] where, as above, subjects were asked to rank images produced by several TMOs based on preference. For both experiments the focus was on subjective preference. Participants were not aware of the appearance of the actual real scene the TMOs were attempting to portray.

Yoshida *et al.* [YBMS05] ran an experiment where tone mapped images were compared to two real scenes. This was important work since it concentrated on the evaluation of images with respect to an actual reference scene. Participants were asked to rate how similar the tone mapped images appeared with respect to the real scene.

In the same period, Ledda *et al.* [LCTS05] proposed a psychophysical methodology to evaluate TMOs using a HDR display as a reference instead of real scenes. They ran a large pairwise study where subjects were presented three images: a reference HDR image displayed on an HDR display and two tone mapped images; presented on conventional displays. Subjects were asked to observe the pair of tone mapped images and select which one appeared most similar to the reference.

Yoshida *et al.* [YMMS06] conducted a series of psychophysical experiments aimed at producing a new tone mapping operator based on subjective data. They ran several experiments where participants were asked to manipulate brightness, contrast and saturation to generate an image as close to a real world reference as possible.

Further work based on subjective data was published by Seetzen *et al.* [See06]. They conducted an experiment to study the effects of luminance, contrast and amplitude resolution of HDR displays. The aim of this work was to determine viewer preference for these parameters in order to produce optimal images.

An experiment using IBL was presented in Ramanarayanan *et al.* [RFWB07]. In this work a series of psychophysical experiments was conducted to test materials, geometry and lightprobes under different conditions and distortions which led to the definition of a new image fidelity metric. Recently, further psychophysical experiments on similar topics were presented in [VLD07].

3. Experimental Framework

As we mentioned above, there are several methods that can be used to validate the perceptual quality produced by an algorithm compared to another. We opted to use a similar methodology to the one presented in Ledda *et al.* [LCTS05] since it suited our requirements well.

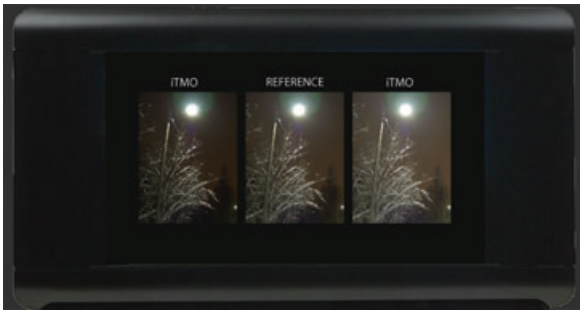

Figure 1: *An example of a paired comparison using the Dolby DR-37P HDR display. Subjects were required to determine which image (left or right) was most similar to the reference (middle).*

A pairwise comparison was conducted where each inverse tone mapped image generated by an iTMO was compared with the inverse tone mapped image generated by the other iTMOs and the reference HDR. At any one time, the viewer was presented with three images displayed side by side on the Dolby DR-37P HDR display. The display was calibrated and had a luminance range of 0.015 cd/m^2 to 3000 cd/m^2. The reference was always displayed in the centre and the images produced by two iTMOs were presented on either side, see Figure 1. The task of the viewer was to observe the two images and indicate which one appeared to be most similar to the reference with respect to a specific criteria which varied across the experiments as explained in Sections 4 and 5.

When data is gathered using a pairwise approach, it is possible to calculate the degree of agreement amongst observers in their preferences. Since each subject was instructed to assess the performance of all possible pairs, the main disadvantage was that a large number of images had to be displayed. On the other hand, this approach is more logical with the added advantage of being straightforward. Although, adopting another approach might have been swifter, due to the limited number of inverse tone mapping algorithms published, such a paired comparison was a suitable choice.

For each reference HDR scene, the total number of possible pairs is $(t(t-1)/2)$ where t is the number of iTMOs being tested. Therefore, in our case, each subject had to evaluate 10 pairs in order to asses all of the combinations for the five iTMOs. For any given pair, the subject had to select which image appeared most similar to the reference. The votes for all participants were then combined into a single preference matrix (reviewers, please refer to additional material).

3.1. Methods for LDR Content Expansion

The five iTMOs selected for our study vary in complexity and computational cost; one of our main goals was to determine whether more complex models are required to solve the LDR to HDR problem. The five operators we compared are Melyan's operator [MDS07], Akyüz's operator [AFR*07], Banterle's operator [BLDC06], Wang's operator [WWZ*07], Rempel's operator [RTS*07].

3.2. Stimuli Generation and Setup

All of the HDR images selected for our experiments were distributed around five orders of magnitude Table 1. This is within the contrast ratio limits of the Dolby DR-37P display. We generated LDR images from the original HDR ones as input for iTMOs using an automatic exposure algorithm [LJKK01]. This algorithm determines the proper exposure factor λ from the HDR scene's luminance values; each HDR image was scaled by λ, then clamped in the range $[0,1]$ and finally discretized in $[0,255]$. This automatic exposure algorithm is similar to the typical one implemented in common digital cameras and video cameras.

In our experiments we did not use tone mapped images as starting LDR images for iTMOs, since all iTMOs have been designed and tested for content generated with traditional capturing device (cameras and video-cameras) and not for tone mapped images. Indeed an iTMO generally works in under-exposed and over-exposed areas of the image which, in tone mapped images, are typically well reproduced.

Each operator was subsequently applied to the generated LDR images. Since each operator required different parameters, what follows are the settings we used for the various algorithms:

- Akyüz *et al.* [AFR*07] (**A**): we used $\gamma = 1$, since it was shown, in the original paper, to be a good parameter. We set k, the maximum input intensity, equal to $L_{\max}$ which is the 95th percentile of the luminance of the original HDR image, thus avoiding outliers.
- Banterle *et al.* [BLDC06] (**B**): for the range expansion, the parameter $L'_{\max}$ (maximum output luminance) was set equal to $L_{\max}$ and L_{white} equal to $2L_{\max}$ as suggested by the authors. For the generation of the expansion map we used values provided in the paper: the radius for the density estimation $r_t = 16$ and the threshold for the number of samples was 6.
- Meylan *et al.* [MDS07] (**M**): we used the same values for thresholds presented in the paper. $Y_{\max}$, the maximum value of the highlight, that was set equal to $L_{\max}$.
- Rempel *et al.* [RTS*07] (**R**): the maximum luminance, was set equal to $L_{\max}$ instead of 1200 cd/m^2 as in the original paper. If the reference has a much lower or higher peak than 1200 cd/m^2 the generated image will appear much darker or brighter than the reference and lead to unfair comparisons.
- Wang *et al.* [WWZ*07] (**W**): we used the same parameters presented in the original paper. In this case manual interaction was needed for generating images. We hallucinated all possible under-exposed and over-exposed regions; however, in some cases it was not possible to find patches for hallucinating from these regions in the same image.

Note that since HDR images are used for generating LDR using a simple linear scale, there is no need to perform linearization of the image's signal for each iTMO.

3.3. Consistency and Agreement

There are several methods to analyse paired data. An approach would be to use Thurstone's Law of Comparative Judgments [Thu27], which is a measuring model for pairwise comparisons. Such a model, however, is more appropriate in the case where one would assume that there are perceptual differences between iTMOs. Our approach, on the other hand, was to analyse the experimental data by primarily testing two 'properties': the individual *consistency* of subjects and the overall *agreement* amongst them.

Consistency or transitivity is an important aspect to consider with paired comparisons. Due to the nature of the experiment a participant can make inconsistent choices when observing paired data. In the simpler case of evaluating three iTMOs for example, if the participant voted $iTMO_i$ to be closer to the reference than $iTMO_j$ and the latter closer than $iTMO_k$, then one would assume, by logic, $iTMO_i$ to be better than $iTMO_k$. In the case of a straightforward ranking, this is what would happen. On the other hand, paired comparisons allow for the case where $iTMO_k$ is voted to be better than $iTMO_i$ thus making an inconsistent choice. To determine such inconsistencies the *Kendall Coefficient of Consistency* can be used [SC88]. This is defined as follows:

$$\zeta = 1 - \frac{24c}{t(t^2-1)}, \tag{1}$$

where t is the number of iTMOs and c is the number of inconsistencies per subject. ζ is calculated on a per-subject and per-image basis and $\zeta \in [0,1]$. If ζ is 1, then there are no inconsistencies and the data could be directly expressed as ranks. ζ will move towards zero as the number of inconsistencies increases. From the above, it may appear that straightforward ranking or rating might be more appropriate as it avoids transitivity issues, nevertheless asking an observer to rank an algorithm from first to last is not always natural behaviour. Measuring inconsistency is useful as it provides a clear indication of the similarity of the algorithms. If there are small differences in the iTMOs, we would expect high inconsistency as the task is hard for the observers.

The overall agreement amongst participants was also tested. This indicates the similarity of votes between observers. A high agreement indicates that subjects classify the iTMOs similarly. Kendall and Babington-Smith have proposed a *Coefficient of Agreement* [SC88] defined as:

$$u = \frac{2\Sigma}{\binom{s}{2}\binom{t}{2}} - 1, \tag{2}$$

where, s is the number of subjects. If we denote a_{ij} the number of times $iTMO_i$ was preferred to $iTMO_j$ then $\Sigma = \sum \binom{a_{ij}}{2}$. Σ is the sum of number of agreements between pairs of observers. Such a coefficient is a suitable measure of association or correlation between a set of ranks. u represents the amount of agreement among the participants and $u \in [-1, 1]$. It has a maximum value of 1 in case of complete agreement (note that u can also assume negative values).

4. Experiment 1: Image Visualization

In this experiment, we wanted to investigate the quality of images generated by the iTMOs compared to a reference HDR image. A selection of eight common HDR images was used for this psychophysical study; these images represent indoor and outdoor scenes in various lighting conditions, from very dark to sunlight, see Figure 2. We conducted three sets of sub-experiments. In the first one, we were concerned with the overall appearance of iTMOs with respect to the reference whereas in the second and third we focused on under and over-exposed areas. iTMOs stretch the luminance channel especially in over-exposed and under-exposed regions of an LDRI, maintaining luminance values in well-exposed regions. We therefore wanted to test such regions separately in order to have a better understanding of individual algorithms. In the first sub-experiment we asked: *Which of the two images is the most similar to the reference?* In the second/third: *Observe the bright/dark areas only; which of the two images is the most similar to the reference?*

4.1. Setup

During the actual experiment a random sequence of all possible pair comparisons for all operators was generated. At any one time three images were presented to the participants: in the centre the HDR image and on either side an inverse tone mapped LDR image, chosen randomly. Each pair combination was only presented once. The Dolby DR-37P display has a resolution of 1920 × 1080 pixels, therefore, we had to resize the three images in order to fit them on the monitor simultaneously. Image width was scaled to 600 pixels whilst maintaining the height-width ratio; in between each image, a 50-pixels black band was inserted.

The HDR images were reproduced with physically correct luminance values following the method suggested in Akyuz *et al*. [AFR*07] using the measured data presented in Ruppertsberg *et al*. [RBBC07].

For each pair the participants had a maximum of 30 seconds to observe the three stimuli before answering. This was shown in a pilot experiment to be sufficient time to successfully complete the task. Since the total number of pairs was 80 (8 images × 10 pairs), we split each participant's experiment in two sessions of 20 minutes reducing the risk of observers getting tired or bored. A group of 24 naive participants (18 men, 6 women) between 21 and 50 years old (mean age 28) with normal or corrected to normal vision took part in this experiment. The display was placed in a dark room minimising the effects of ambient light, and participants were seated 1 metre away from the monitor. Prior to the experiment, each

Figure 2: *The eight images used for Experiment 1, reproduced with the automatic exposure. In the top row from left to right: Scene 1, Scene 2, Scene 3, and Scene 4. In the bottom row from left to right: Scene 5, Scene 6, Scene 7, and Scene 8.*

Table 1: *The dynamic ranges of the HDR images used in the first experiment Figure 2*

Image	Dynamic Range
Scene 1	3.2
Scene 2	3.8
Scene 3	5.5
Scene 4	5.4
Scene 5	3.2
Scene 6	5.3
Scene 7	4.4
Scene 8	3.2

subject was given five minutes to adapt to the environment and was presented with three trials in order to familiarize themselves with the experiment. The same group of participants took part in all three sub-experiments. Presentation order of the three sub-experiments was randomized for each subject.

4.2. Experiment 1: Results

In this section we present the results for the three experiments on image visualisation.

4.2.1. *Overall similarity*.

The results for the first sub-experiment are shown in Table 2. The rows represent data obtained for each of the eight scenes.

In the first column of Table 2, the average consistency is tabulated. As can be seen, the consistency of participants was reasonably high implying that the task was sufficiently simple for the subjects. Such a result also indicates that the algorithms are different enough to be clearly assessed.

The coefficient of agreement u, as discussed above, is an indication of the agreement amongst participants. The value of u for each scene is tabulated. The higher the value the greater the agreement between observers. To determine the statistical significance of u, we can test the null hypothesis H_0 that there is no agreement amongst the raters implying that each iTMO is perceptually equivalent causing difficulties in judging which image is most similar to the reference. In other words, we can test whether subjects allot their preferences at random. To test the significance of u we may use a large-sample approximation to the sampling distribution. For each image, X^2 is compared to a critical value tabulated in statistics tables. Details of the test statistic are included in the Appendix. As shown in Table 2, u is significant at the standard $\alpha = 0.05$ level. Typically, the agreement was much higher ($p < 0.001$) and therefore we can, with confidence, state that observers generally agree with each other when assessing the algorithms. Thus we reject the null hypothesis H_0 that the iTMOs are perceptually the same.

To reinforce our analysis, another test to determine overall significance in the data that is analogue to analysis of variance (ANOVA), was conducted [Dav88]. With ordinal data, this test is more meaningful than a typical ANOVA. The test, instead of considering the agreement amongst subjects u, determines whether there are differences in the scores a_i obtained by the iTMOs. More specifically, it determines whether differences in scores are purely by chance or if indeed the differences are significant. For each image, the standardized sum of squares of the scores S is calculated and it is compared to a critical value S_c for a desired significance level ($\alpha = 0.05$). See the Appendix for more detailed information. As can be seen in Table 2, the results are highly significant ($p < 0.05$) and we can reject H_0.

Typically, following a significant ANOVA, a series of post-hoc tests are conducted to determine *where* the differences lie and if the score between any two iTMOs can be

Table 2: *The results for Experiment 1 Overall Similarity. The first column shows the average coefficient of consistency. The second the coefficient of agreement. In the third column the value of X^2 is presented, which is approximately distributed as χ^2. From this value the significance u is determined, fourth column. The standardized sum of the squares of the scores S and its significance are in column five and six. Finally, the obtained ranks are presented. Any two iTMOs within the same circle cannot be considered perceptually different. The R value is equal to 21 for a single scene and equal to 60 for the Average, see Appendix*

	Ave. Coeff. Cons. ς	Coeff. Agr. u	X^2	Sign. u	S	Sign S	1st	2nd	3rd	4th	5th
Scene 1	0.700	0.210	58.333	$p < .05$	50.133	$p < .05$	B	R	W	A	M
Scene 2	0.508	0.059	23.667	$p < .05$	15.000	$p < .05$	A	B	R	W	M
Scene 3	0.792	0.476	119.500	$p < .05$	109.867	$p < .05$	W	B	R	M	A
Scene 4	0.917	0.650	159.500	$p < .05$	145.933	$p < .05$	B	W	R	M	A
Scene 5	0.858	0.149	44.333	$p < .05$	41.600	$p < .05$	B	A	W	R	M
Scene 6	0.867	0.392	100.167	$p < .05$	94.467	$p < .05$	B	M	W	R	A
Scene 7	0.692	0.263	70.500	$p < .05$	69.533	$p < .05$	M	B	R	W	A
Scene 8	0.500	0.058	23.333	$p < .05$	9.867	$p < .05$	R	B	W	A	M
Average	0.729	0.282	74.917	$p < .05$	67.050	$p < .05$	B	W	R	M	A
							531	434	411	295	249

considered significant. For this purpose, significance tests of the score differences were additionally performed. This procedure is based on the range of the scores obtained by the five iTMOs and it is equivalent to Tukey's method used with ANOVA [Dav88]. We visualize these differences by grouping iTMOs in circles as shown in the tables. If the difference in scores a_i of any two iTMOs is larger than a critical value R (see Appendix), the algorithms are assigned to different circles and may be considered distinguishably different. We can therefore conclude that iTMOs within the same circle are considered perceptually similar. For clarity purposes we only present the average scores across the entire experiment which is tabulated in the last row (reviewers, please refer to additional material).

Although results are scene dependent, a reasonably clear pattern can be seen. iTMO B always ranks in the first group.

It is also fairly clear that iTMOs M and even more so A, did not perform as well. One or the other was always ranked last for most scenes. The poor performance of A can be explained by the fact that the algorithm is a simple linear scale which, especially in dark HDRIs, leads to over bright images. Finally, the performance of iTMOs R and W was more or less the same exchanging position throughout the eight test scenes.

We refer the reader to the Appendix and [Dav88] for a discussion on the specific tests to be performed.

4.2.2. *Bright and dark areas*

The results for the comparison of bright regions are presented in Table 3. As in the previous experiment, iTMO B was always ranked in the first group. The performance of iTMO W was also very positive being ranked in the first group for many of the scenes and in the second for the rest. iTMO M performed, once more, poorly. As in the overall comparison, the agreement u was always statistically significant. In this experiment the significance was even higher at the $\alpha = 0.001$ level. This result is also verified analysing the score differences which are highly significant ($p < 0.05$).

In Table 4, the results of the comparison of dark regions are presented. Interestingly, in this instance we noticed a complete reverse in performance. iTMO M, which performed relatively badly in the two previous sub-experiments, was ranked in the first group for the majority of test scenes and

Table 3: *The results for Experiment 1 Bright Areas. The R value is equal to 21 for a single scene and equal to 60 for the Average, see Appendix*

	Ave. Coeff. Cons. ς	Coeff. Agr. u	X^2	Sign. u	S	Sign S	1st	2nd	3rd	4th	5th
Scene 1	0.808	0.578	143.000	p < . 05	118.733	p < . 05	B	W	A	R	M
Scene 2	0.717	0.328	85.333	p < . 05	75.867	p < . 05	B	A	W	R	M
Scene 3	0.708	0.268	71.667	p < . 05	64.667	p < . 05	B	W	R	M	A
Scene 4	0.775	0.257	69.000	p < . 05	61.867	p < . 05	B	W	M	R	A
Scene 5	0.858	0.374	96.000	p < . 05	88.133	p < . 05	B	W	A	R	M
Scene 6	0.833	0.399	101.833	p < . 05	93.133	p < . 05	M	B	A	R	W
Scene 7	0.625	0.193	54.333	p < . 05	42.133	p < . 05	M	B	A	W	R
Scene 8	0.642	0.228	62.333	p < . 05	59.800	p < . 05	B	R	W	A	M
Average	0.746	0.328	85.437	p < . 05	75.542	p < . 05	B	W	R	A	M
							554	427	350	320	269

Table 4: *The results for Experiment 1 Dark Areas. The R value is equal to 21 for a single scene and equal to 60 for the Average, see Appendix*

	Ave. Coeff. Cons. ς	Coeff. Agr. u	X^2	Sign. u	S	Sign S	1st	2nd	3rd	4th	5th
Scene 1	0.625	0.219	60.333	p < . 05	55.200	p < . 05	M	R	B	A	W
Scene 2	0.783	0.459	115.500	p < . 05	106.867	p < . 05	M	A	R	W	B
Scene 3	0.892	0.612	150.833	p < . 05	137.533	p < . 05	W	R	B	M	A
Scene 4	0.883	0.601	148.167	p < . 05	140.200	p < . 05	B	W	R	M	A
Scene 5	0.658	0.218	60.167	p < . 05	54.200	p < . 05	M	A	W	B	R
Scene 6	0.808	0.273	72.833	p < . 05	64.867	p < . 05	B	W	R	M	A
Scene 7	0.600	0.208	57.833	p < . 05	52.667	p < . 05	M	B	W	R	A
Scene 8	0.475	0.107	34.500	p < . 05	11.867	p < . 05	M	R	W	A	B
Average	0.716	0.337	87.521	p < . 05	77.925	p < . 05	M	B	R	W	A
							477	411	405	382	245

was first on average. A performs poorly which may be related to the fact that linear scaling increases the luminance in the dark regions. The significance level of the data is the same as the Bright Areas experiment.

5. Experiment 2: image-based lighting

In a second experiment, we tested the performance of inverse tone mapping with regards to a common application: image-based lighting. Subjects had to make pairwise assessments of images generated using IBL with the five iTMOs and compare them to images generated using an HDR lightprobe. IBL consists of the evaluation of the following equation at point x with view direction $\vec{\omega}$:

$$L_o(x,\vec{\omega}) = L_e + \int_\Omega f_r(x,\vec{\omega},\vec{\omega}')V(x,\vec{\omega}') \times L_i(x,\vec{\omega}')(n\cdot\vec{\omega}')d\vec{\omega}', \quad (3)$$

where L_e is the emitted radiance, n is the normal at x, f_r the bidirectional reflectance distribution function (BRDF), L_i is the incoming radiance (in the case of our experiment the lightprobe), $V(x,\vec{\omega}')$ a binary function that is true when a ray from position x with direction $\vec{\omega}'$ does not intersect an object. Since the appearance of an object depends strongly on the BRDF, we tested IBL with various materials. For example, a BRDF that represents a perfect specular mirror shows more details of a lightprobe; therefore, differences between iTMOs might be perceptually more noticeable, because the integral degenerates into a single weighted value, $L_i(x,\omega')$. On the other hand, a pure diffuse material could mask these differences because the integral degenerates to an equal distribution on the hemisphere, that can be thought of as an averaging process. We investigated the performance of the five iTMOs for three classes of materials: pure diffuse, glossy and pure specular reflective.

5.1. Stimuli generation and setup

Six HDR lightprobes were selected as lighting data for the IBL experiment with different orders of magnitude, see Table 6. The lightprobes are shown in Figure 3. Each lightprobe was scaled using the same technique presented in Section 3.2. The resulting LDR lightprobes acted as input to the five iTMOs. A pilot study was conducted with three simple objects of different complexity to determine whether they would have an impact on the outcome of the experiment, see Figure 4. None of these objects contained any high-frequency local geometry that could have disturbed the appearance of the illumination, such as the complex geometry which made it hard to distinguish illumination as shown in [RFWB07]. We rendered images with six lightprobes (five iTMO and one reference lightprobe) using IBL with a glossy material and conducted paired comparisons. The outcome of our pilot was that the difference in the geometric complexity of our three scenes did not have a significant impact on the outcome; the performance of the iTMOs was not affected.

For the actual experiment, a simple 3D scene consisting of a teapot (the object of medium complexity in the pilot) was lit with the generated lightprobe images, see Figure 5. Each teapot had a different material assigned to it: a pure diffuse, a Ward glossy BRDF [War92] ($\alpha_v = 0.1, \alpha_u = 0.25$), and perfect specular mirror. A Monte Carlo ray tracer rendered the images for each lightprobe, material and iTMO. As in Experiment 1, the task was to select which image appeared to be closer to a reference image rendered with the original HDR lightprobe. For the IBL experiment, we only studied overall performance as distortions due to reflections would make it hard to distinguish dark and bright areas. With 10 comparison for each lightprobe, the total image pairs to be evaluated were 60. A randomly generated sequence of such pairs was presented using the same approach as in the previous experiment.

For each possible pair, each subject again had a maximum of 30 seconds in which to give an answer. The experiment was split for each material (pure diffuse, glossy, pure specular). A group of 24 naive participants (16 men, 8 women) between 19 and 56 years of age (mean age 22) took part in this experiment. All exhibited normal or corrected to normal vision. Subjects were given 5 minutes to adapt to the environment and were allowed three test trials.

Table 5: *The results for Experiment 2 Diffuse Material. The R value is equal to 21 for a single scene and equal to 52 for the Average, see Appendix*

	Ave. Coeff. Cons. ς	Coeff. Agr. u	X^2	Sign. u	S	Sign S	1st	2nd	3rd	4th	5th
Scene 1	0.658	0.073	26.833	$p < .05$	15.000	$p < .05$	R	B	A	M	W
Scene 2	0.725	0.322	84.000	$p < .05$	78.933	$p < .05$	B	R	W	M	A
Scene 3	0.658	0.100	33.000	$p < .05$	30.600	$p < .05$	W	B	R	M	A
Scene 4	0.600	0.116	36.667	$p < .05$	28.467	$p < .05$	W	B	R	M	A
Scene 5	0.608	0.154	45.500	$p < .05$	38.067	$p < .05$	B	A	M	W	R
Scene 6	0.700	0.069	25.833	$p < .05$	20.067	$p < .05$	B	A	M	W	R
Average	0.658	0.139	41.972	$p < .05$	35.189	$p < .05$	B	W	R	M	A
							355	305	282	280	218

Table 6: *The dynamic ranges of the HDR images used in the second experiment, see Figure 3*

Image	Dynamic Range
Scene 1	5.4
Scene 2	3.7
Scene 3	7.1
Scene 4	4.6
Scene 5	3.7
Scene 6	3.2

5.2. Experiment 2: Results

In this section, we present the results for the three experiments on image-based lighting.

5.2.1. *Diffuse material*

The results for the experiment with the diffuse material are shown in Table 5. The rows represent data obtained for each of the six lightprobes. B performed better overall being placed in the first group on all occasions. W, R and M obtained similar overall results. A performed quite poorly overall and placed last. As seen in Experiment 1, A performs badly because it is a simple linear scale. This linear process is equivalent to evaluating IBL using a normal LDR and subsequently scaling the resulting image. The outcome of the experiment was statistically significant ($p < 0.05$).

5.2.2. *Specular materials*

The results for the experiment with the glossy material are shown in Table 7. In this case, we have similar results to the one obtained for diffuse material shown in Table 5.

Finally the results for the perfect specular mirror material are shown in Table 8. In this experiment B again performed better. W, R and M had very similar performances. It is worth noting that M performed well for both specular materials for the Scene 3, possibly echoing the results from Experiment 1 where it performed best in darker areas.

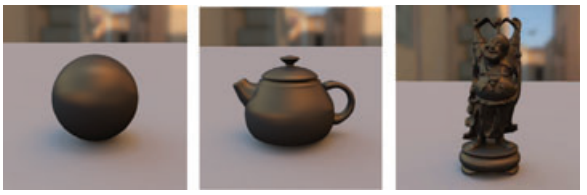

Figure 4: *Models used in the pilot from right to left: a sphere, a teapot model, a more complex model, the Happy Buddha.*

6. Discussion

For the first experiment, the monotonically increasing functions B, W and R that enhance contrast non-linearly perform better overall and in many of the results are grouped together. The linear method A, and to a lesser extent M, perform worse overall, reflecting that for still images complex methods recreate HDR perceptually better.

For Experiment 2, the diffuse results show few differences. This is mostly due to the fact that rendering with IBL consists of the evaluation of an integral and during integration small details may be lost. This is less true for perfectly mirror-like or high glossy materials. However, in these cases details of the lightprobe reflected in the objects may be too small to be seen as is shown by the large groupings in the results. With certain scenes, for example Scene 4 and Scene 5 where the upper hemisphere of the lightprobe is uniformly well lit, the linear operators perform well. For more complex lightprobes, we revert to the previously found ranking. Overall it is still clear that the operators that perform best, as with Experiment 1, are the non-linear operators.

Judging from the average results in our experiments, Operator B performs better overall than the other operators, having been ranked in the top group for all sub-experiments except for Experiment 1, Dark Areas. This result may come at a cost since it is also the most expensive to compute. Nevertheless, when the most accurate results are required, this operator is currently the most likely to recreate most closely the original dynamic range of an image. W and R are commonly paired together in the average results. There is only one instance, Experiment 1, Bright Areas, when W outperforms R in the

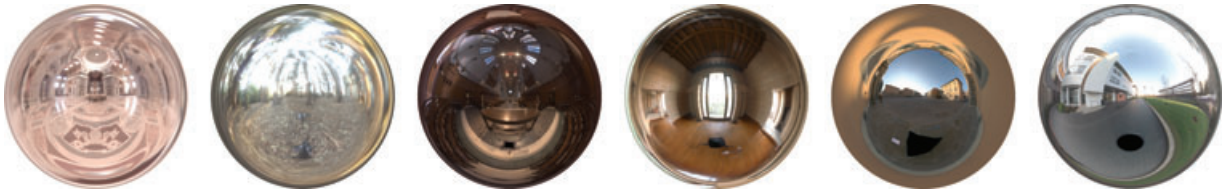

Figure 3: *Lightprobes used for the Experiment 2, reproduced with the automatic exposure. From the left: Saint Peter's Cathedral (Scene 1), Eucalyptus Grove (Scene 2), Grace Cathedral (Scene 3) and Ennis (Scene 4), Pisa's square (Scene 5), and Campus (Scene 6).*

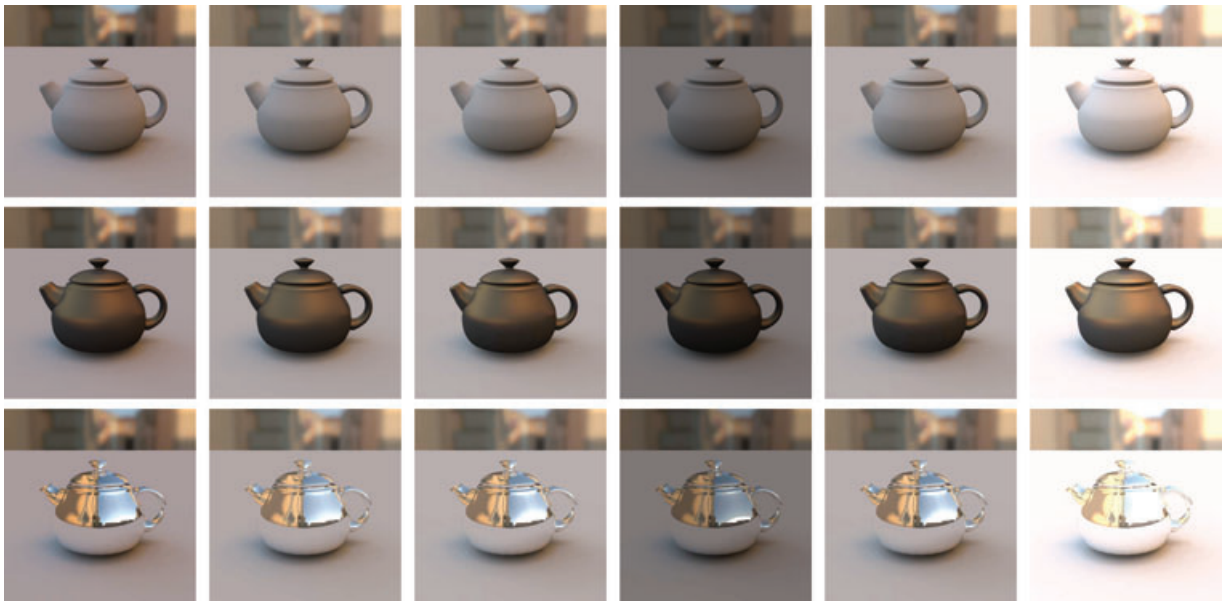

Figure 5: *A sample of the images used in Experiment 2 using the lightprobe Pisa's square at exposure 0. In the first row diffuse material example, from left to right: reference HDR lightprobe, R operator, B operator, M operator, W operator and A operator. In the second row glossy material example, and in the third row pure specular material, the order is kept the same as in the first row.*

Table 7: *The results for Experiment 2 Glossy Material. The R value is equal to 21 for a single scene and equal to 52 for the Average, see Appendix*

	Ave. Coeff. Cons. ς	Coeff. Agr. u	X^2	Sign. u	S	Sign S	1st	2nd	3rd	4th	5th
Scene 1	0.733	0.166	48.167	$p < .05$	41.933	$p < .05$	B	R	M	W	A
Scene 2	0.742	0.225	61.833	$p < .05$	60.267	$p < .05$	R	W	B	M	A
Scene 3	0.542	0.015	13.500	$p < .05$	11.667	$p < .05$	M	B	R	W	A
Scene 4	0.608	0.063	24.500	$p < .05$	12.667	$p < .05$	B	R	W	M	A
Scene 5	0.617	0.073	26.833	$p < .05$	19.533	$p < .05$	B	A	M	W	R
Scene 6	0.633	0.076	27.500	$p < .05$	18.133	$p < .05$	B	W	A	M	R
Average	0.646	0.103	33.722	$p < .05$	27.367	$p < .05$	B	W	R	M	A
							353	304	294	281	208

Table 8: *The results for Experiment 2 Mirror Material. The R value is equal to 21 for a single scene and equal to 52 for the Average, see Appendix*

	Ave. Coeff. Cons. ς	Coeff. Agr. u	X^2	Sign. u	S	Sign S	1st	2nd	3rd	4th	5th
Scene 1	0.675	0.172	49.667	$p < .05$	45.400	$p < .05$	B	R	W	M	A
Scene 2	0.692	0.250	67.500	$p < .05$	59.000	$p < .05$	B	R	M	W	A
Scene 3	0.683	0.070	26.167	$p < .05$	17.667	$p < .05$	M	B	R	W	A
Scene 4	0.533	0.048	21.000	$p < .05$	12.667	$p < .05$	M	B	A	W	R
Scene 5	0.800	0.353	91.167	$p < .05$	84.667	$p < .05$	W	B	A	R	M
Scene 6	0.717	0.220	60.500	$p < .05$	52.200	$p < .05$	W	B	A	M	R
Average	0.683	0.186	52.667	$p < .05$	45.267	$p < .05$	B	W	M	R	A
							381	315	275	274	195

average results. W is not as expensive as B since the expansion is localized to only the required under-exposed and over-exposed regions of the image. However, it requires a substantial amount of manual work. The results for R are very promising since it is fast enough to compute in real time using graphics hardware and while our experiment accounts only for still images, unlike B and W, it also adapts well to HDR video. In the overall, M ranks somewhere between, W and R and A. In particular, for all the average IBL results it is grouped perceptually with W and R. Furthermore, most importantly it performed best in our experiments for dark regions. This conforms to the design of the operator, since it was primarily designed to enhance the specular highlights when expanding LDRs to HDRs. These results are further strengthened by the operator's performance in the pure specular mirror material experiment where it came first in the two darkest scenes. Another advantage of M is that it is faster to compute than the other, non-linear, operators. Finally, A always ranks in the last group, except in the case of Experiment 1, Bright Areas. While A was designed to show that HDR images can be enhanced in a simple way to give a good HDR appearance as shown in [AFR*07], when attempting to recreate the original lost data this approach is insufficient. The advantages of this operator lie with its fast execution times and the possibility of using it with uniformly well-lit lightprobes.

The experiments have shown that more advanced iTMOs, that cater for quantization errors introduced during expansion of an LDR image, such as B, R and W, can perform better than simple techniques that apply single or multiple linear scale expansions, such as A and M. The more computationally expensive methods B, R and W, are better at recreating HDR than simple methods, avoiding most artefacts. Even if a linear scale can elicit an HDR experience in an observer, as shown in [AFR*07], it does not correctly reproduce the perception of the original HDR image.

7. Conclusions

In this paper, we presented a series of psychophysical studies for inverse TMOs. The first study focused on the visualization of LDR content on an HDR display. The focus of the second experiment was the quality of IBL for inverse TMOs using different materials. Our experiments have shown that simple single or multiple linear scales do not perform well in comparison with a reference HDR image. More complicated algorithms present better performance compared to simpler ones. Furthermore, the operators can be distinctly ranked based on their performance. For image-based lighting applications, while differences are not that absolute, a distinction can still be made between non-linear algorithms and purely linear expansions. These results, while not conclusive, demonstrate that a significant improvement in quality can be achieved by investing into more computationally complex inverse tone mapping methods. Future work should look at improving computational performance of these non-linear algorithms. It also indicates that further investigation into novel inverse tone mapping algorithms could prove fruitful.

8. Appendix

The method for determining whether the coefficient of agreement u is significantly different from the value that would be obtained if the comparisons were randomly made is presented below. If the number of observers s is small, probability tables have already been tabulated and can be found in various statistics books, see [SC88] (table U). For a large number of s and t the following equation can be used to compute X^2 which is approximately distributed as chi-square χ^2:

$$X^2 = \frac{t(t-1)(1+u(s-1))}{2} \tag{4}$$

The significance of X^2 for degrees of freedom $\binom{t}{2}$ may be determined from tables of critical values of the chi-square distribution, see [SC88] (table C). If, the probability obtained is greater or equal to the tabulated values at a specific level of significance α we can reject the null hypothesis and confirm that there is strong agreement amongst observers when observing the various algorithms compared to the reference. At $\alpha = .05$ and 10 df the critical value as tabulated is 18.31.

The overall test of equality is a second test which was conducted to determine significance in the data. For each image we compute the standardized sum of squares of the scores S and compare it with S_c, the upper α significance value of the χ^2 distribution with $t-1$ degrees of freedom:

$$S = 4\frac{\sum_{i=1}^{t} a_i^2 - \frac{1}{4}ts^2(t-1)^2}{st} \tag{5}$$

where a_i^2 is the sum of squares of the scores. For an explanation of the derivation of Equation 5, see [Dav88]. If the observed S value is greater than or equal to the corresponding critical value S_c, reject H_0. From χ^2 tables, for $t = 5$ at 0.05 significance level, the critical value $S_c = 9.45$.

Finally, the equation for the multiple comparison test, where the significance of each algorithm is tested against the others follows. If the difference in scores between two iTMO is greater than R, then we can assume the difference to be significant.

$$R = \frac{1}{2}W_{t,\alpha}\sqrt{st} + \frac{1}{4} \tag{6}$$

where $W_{t,\alpha}$ is the upper significance point of the W_t distribution. At a significance level of 0.05 and $t = 5$ iTMOs $W_{t,\alpha} = 3.86$. See Pearson [PH88], table 22.We performed the multiple comparison range test on the results of the preference experiment, for each image individually (s = number subjects) and for the combined preference over all images

(Average) (s = number subjects $\times$ number images). In the first case the value of $R = 21$ for both experiments; in the second case the value of $R = 60$ for Experiment 1 and $R = 52$ for Experiment 2.

Acknowledgements

We thank Elena Sikudova for help with the experimental setup. Furthermore, we thank Ahmet Akyüz, Kimmo Roimela and Paul Debevec for HDR images used in this paper. Also we thank anonymous reviewers for their comments.

The work reported in this paper has formed part of EPSRC grant EP/D032148 whose funding and support is gratefully acknowledged.

References

[AFR*07] AKYÜZ A. O., FLEMING R., RIECKE B. E., REINHARD E., BÜLTHOFF H. H.: Do hdr displays support ldr content?: A psychophysical evaluation. In *SIGGRAPH '07: ACM SIGGRAPH 2007 papers* (New York, NY, USA, 2007), ACM Press.

[BLDC06] BANTERLE F., LEDDA P., DEBATTISTA K., CHALMERS A.: Inverse tone mapping. In *ACM GRAPHITE '06* (New York, NY, USA, 2006), ACM Press, pp. 349–356.

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

[Deb98] DEBEVEC P.: Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *SIGGRAPH '98: ACM SIGGRAPH 1998 papers* (New York, NY, USA, 1998), ACM Press, pp. 189–198.

[Deb05] DEBEVEC P.: A median cut algorithm for light probe sampling. In *ACM Siggraph 2005 Posters* (2005) (New York, NY, USA, 2005), ACM.

[DM97] DEBEVEC P. E., MALIK J.: Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH '97: ACM SIGGRAPH 1997 papers* (New York, NY, USA, 1997), ACM Press/Addison-Wesley Publishing Co., pp. 69–378.

[DMMS03] DRAGO F., MARTENS W. L., MYSZKOWSKI K., SEIDEL H.-P.: Perceptual evaluation of tone mapping operators. In *SIGGRAPH '03: ACM SIGGRAPH 2003 Sketches & Applications* (New York, NY, USA, 2003), ACM, pp. 1–1.

[Dol05] Dolby: http://www.dolby.com/promo/hdr/technology.html, 2005.

[Far01] FARID H.: Blind inverse gamma correction. *IEEE Transactions on Image Processing 10*, 10 (2001), 1428–1433.

[Hat06] HATEREN J. H. V.: Encoding of high dynamic range video with a model of human cones. *ACM Transactions on Graphics 25*, 4 (2006), 1380–1399.

[KYJF04] KUANG J., YAMAGUCHI H., JOHNSON G. M., FAIRCHILD M. D.: Testing hdr image rendering algorithms. In *Color Imaging Conference* (2004), pp. 315–320.

[Lan02] LANDIS H.: Production-ready global illumination. In *Siggraph Course Notes 16*, 2002.

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

[LJKK01] LEE J.-S., JUNG Y.-Y., KIM B.-S., KO S.-J.: An advanced video camera system with robust af, ae, and awb control. *IEEE Transactions on Consumer Electronics* 47, 3 (August 2001), 694–699.

[LSA05] LI Y., SHARAN L., ADELSON E. H.: Compressing and companding high dynamic range images with subband architectures. *ACM Trans. Graph. 24*, 3 (2005), 836–844.

[LZ05] LIN S., ZHANG L.: Determining the radiometric response function from a single grayscale image. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2* (Washington, DC, USA, 2005), IEEE Computer Society, pp. 66–73.

[MDS06] MEYLAN L., DALY S., SÜSSTRUNK S.: The reproduction of specular highlights on high dynamic range displays. In *IS&T/SID 14th Color Imaging Conference*, 2006.

[MDS07] MEYLAN L., DALY S., SÜSSTRUNK S.: Tone mapping for high dynamic range displays. In *Electronic Imaging*, vol. 6492, 2007.

[MEMS06] MANTIUK R., EFREMOV A., MYSZKOWSKI K., SEIDEL H.-P.: Backward compatible high dynamic range mpeg video compression. *ACM Transactions on Graphics 25*, 3 (2006), 713–726.

[OA06] OKUDA M., ADAMI N.: Two-layer coding for high dynamic range images based on inverse tone mapping. In *International Conference on Signal Processing 2006, Japan*, IEEE, 2006.

[PH88] PEARSON E. S., HARTLEY H. O.: *Biometrika tables for statisticians 3rd ed., vol. 1*. Cambridge University Press, 1988.

[RBBC07] Ruppertsberg A. I., Bloj M., Banterle F., Chalmers A.: Displaying colourimetrically calibrated images on a high dynamic range display. *Journal of Visual Communication and Image Representation 18*, 5 (2007), 429–438.

[RFWB07] Ramanarayanan G., Ferwerda J., Walter B., Bala K.: Visual equivalence: Towards a new standard for image fidelity. *ACM Transactions on Graphics 26*, 3 (2007), 76.

[RSSF02] Reinhard E., Stark M., Shirley P., Ferwerda J.: Photographic tone reproduction for digital images. *ACM Transactions on Graphics 21*, 3 (2002), 267–276.

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

[RWPD05] Reinhard E., Ward G., Pattanaik S., Debevec P.: *High Dynamic Range Imaging: Acquisition, Display and Image-Based Lighting*. Morgan Kaufmann Publishers, December 2005.

[SC88] Siegel S., Castellan N. J.: *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill International, 1988.

[SLY*06] Seetzen H., Li H., Ye L., Heidrich W., Whitehead L., Ward G.: Observation of contrast, brightness, and amplitude resolution of displays. In *SID INTERNATIONAL SYMPOSIUM*, June 2006. The society for information display.

[SHS*04] Seetzen H., Heidrich W., Stuerzlinger W., Ward G., Whitehead L., Trentacoste M., Ghosh A., Vorozcovs A.: High dynamic range display systems. *ACM Transactions on Graphics 23*, 3 (2004), 760–768.

[Thu27] Thurstone L. L.: A law of comparative judgment. In *Psychological Review 34* (1927), pp. 273–286.

[VLD07] Vangorp P., Laurijssen J., Dutré P.: The influence of shape on the perception of material reflectance. *ACM Transactions on Graphics 26*, 3 (2007), 77.

[War92] Ward G. J.: Measuring and modeling anisotropic reflection. *SIGGRAPH Computer Graphics 26*, 2 (1992), 265–272.

[WWZ*07] Wang L., Wei L.-Y., Zhou K., Guo B., Shum H.-Y.: High dynamic range image hallucination. In *Proceedings of Eurographics Symposium on Rendering*, June 2007.

[YBMS05] Yoshida A., Blanz V., Myszkowski K., Seidel H.-P.: Perceptual evaluation of tone mapping operators with real-world sceness. In *Human Vision and Electronic Imaging X, IS&T/SPIE's 17th Annual Symposium on Electronic Imaging (2005)* (San Jose, USA, January 2005), vol. 5666 of *SPIE Proceedings Series*, SPIE, pp. 192–203.

[YMMS06] Yoshida A., Mantiuk R., Myszkowski K., Seidel H.-P.: Analysis of reproducing real-world appearance on displays of varying dynamic range. In *EUROGRAPHICS 2006 (EG'06)* (Vienna, Austria, September 2006), vol. 25 of *Computer Graphics Forum*, Eurographics, Blackwell, pp. 415–426.