Evaluation of Tone Mapping Operators using a High Dynamic Range Display

Patrick Ledda*
University of Bristol

Alan Chalmers
University of Bristol

Tom Troscianko †
University of Bristol

Helge Seetzen ‡
Sunnybrook Technologies

Abstract

Tone mapping operators are designed to reproduce visibility and the overall impression of brightness, contrast and color of the real world onto limited dynamic range displays and printers. Although many tone mapping operators have been published in recent years, no thorough psychophysical experiments have yet been undertaken to compare such operators against the real scenes they are purporting to depict. In this paper, we present the results of a series of psychophysical experiments to validate six frequently used tone mapping operators against linearly mapped High Dynamic Range (HDR) scenes displayed on a novel HDR device. Individual operators address the tone mapping issue using a variety of approaches and the goals of these techniques are often quite different from one another. Therefore, the purpose of this investigation was not simply to determine which is the "best" algorithm, but more generally to propose an experimental methodology to validate such operators and to determine the participants' impressions of the images produced compared to what is visible on a high contrast ratio display.

CR Categories: I.3.3 [Computer Graphics]: Picture/Image Generation—Display Algorithms; I.4.0 [Image Processing and Computer Vision]: General—Image Displays

Keywords: Tone Mapping, High Dynamic Range, Psychophysics

1 Introduction

In the real world, our visual system is presented with a wide range of colors and intensities. A surface lit by starlight might have a luminance level of around 10^{-3} cd/m^2, while daylight scenes are close to 10^5 cd/m^2. A well-designed CRT or LCD monitor, even in a darkened room, is only able to achieve a maximum luminance of around 150 cd/m^2 and a contrast ratio of not more than two orders of magnitude. In 1993, Tumblin and Rushmeier proposed a framework to map real world luminances to target display luminances [Tumblin and Rushmeier 1993]. Such tone mapping operators (TMOs) generate images visually similar to a real scene, Figure 1, by carefully mapping to a set of luminances that can be displayed on a low contrast ratio display or printed. These TMOs are capable of enormous reductions in contrast to fit the displayable range and manage to produce satisfying and visually appealing results. However, until now, there has not been a thorough evaluation as to just how accurate the results of the TMOs are, compared with the real scene they are intending to portray.

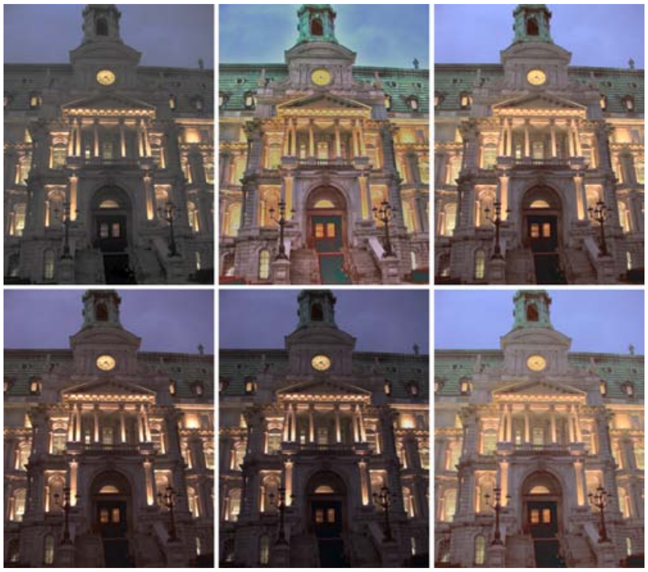


Figure 1: Six tone mapped images which were compared to a reference scene (*Scene 8*) displayed on the HDR device.

2 Previous work

In 2003, Drago et al. [Drago et al. 2003b] asked subjects to make preference judgements on how perceptually similar or dissimilar tone mapped images were. They analyzed this data such that it formed a stimulus space in which the coordinates in each dimension correlated to a stimulus describing differences between images. Based on these preference results, they determined a preference point in the stimulus space which they then used as a reference. This allowed them to determine what operators were perceptually perceived as the most similar to this preference point. In 2004, Kaung et al. [Kuang et al. 2004] ran a similar test where participants were required to judge different images again based on preference. More recently, Yoshida et al. [Yoshida et al. 2005] carried out an experiment where they compared six algorithms to two real scenes. They asked participants to rate images generated by the different operators in terms of how similar they appeared compared to the actual scene. Although direct comparison with real scenes is important, it can introduce uncontrolled variables.

All of these earlier studies were based on rating a series of tone mapped images which, we believe, may not be the most appropriate approach since it is very difficult to quantify such test scenes and thus draw any conclusive results. This is discussed in more detail in section 4.

The work described in this paper differs from the above in several ways. Firstly, we undertake our validation of tone mapping operators by making comparison with a reference scene displayed on an HDR monitor [Seetzen et al. 2003]. This is a novel approach and eliminates many of the uncontrolled variables present in previous studies. The HDR technology, which has previously been validated against reality [Ledda et al. 2004a], allows us to make direct comparisons with many scenes of various stimuli which simplified the

*e-mail: [ledda] [alan]@cs.bris.ac.uk

†e-mail: tom.troscianko@bris.ac.uk

‡e-mail: helge.seetzen@sunnybrooktech.com

validation process because subjects had a specific reference when making image judgements. Adopting the HDR display allows participants to match a 2D reference image to a 2D tone mapped image instead of comparisons with a 3D real scene. It has the advantage of controlling screen resolution, dimensions, colorimetry, viewing distance and ambient lighting.

Our methodology, which could also be valid for real scenes, is based on paired comparisons where each subject is presented three stimuli at any one time, the reference and two tone mapped images. The closest image to the reference, depending on the experimental condition being tested, is then chosen. This is a simpler task than having to rate a series of images presented in one go. Previous work had also the limitation that only a few scenes were considered and a low number of participants actually took part in the experiments. We conducted a study with 109 participants and 23 test scenes.

3 Tone Mapping Operators

Tone mapping operators can be classified into different categories depending on how they attempt to reduce contrast in a HDR image. Those models such as [Ward 1994; Schlick 1994; Tumblin et al. 1999] that apply the same mapping function across the image are known as *global* operators. These algorithms, although not very computationally expensive, do not cope well with huge contrast ratios. Those operators in which the mapping varies spatially depending on a neighborhood of a pixel are known as *local*. Local operators, for example [Pattanaik et al. 1998; Tumblin and Turk 1999; Fattal et al. 2002], are generally capable of a greater contrast reduction allowing significant compression of the dynamic range of a scene. However, a major concern with spatially-varying operators is that contrast reversals (or "halos") artifacts can appear around high contrast edges. In addition, some operators try to mimic the human visual system adopting mapping functions that closely resemble aspects of human vision [Upstill 1985; Tumblin and Rushmeier 1993; Larson et al. 1997; Pattanaik et al. 1998]. A few perceptual operators [Ferwerda et al. 1996; Pattanaik et al. 2000; Durand and Dorsey 2000; Ledda et al. 2004c], also model other effects such as the time course of adaptation, loss of color and visual acuity at different illumination levels. The aim is to produce images which are even closer to what an observer would perceive in reality.

Below, we briefly introduce the six different algorithms that were chosen for this investigation. The italic letters in brackets included with each TMO description is the name we shall use to refer to the operator in our results. Most of the images used for the validation were either generated from code kindly donated by the authors or computed using available source code. We obviously understand that some of the operators' performance could be improved by modifying various parameters. Whenever possible, we attempted to use their default settings as presented in the respective papers. A more detailed discussion of these, and other TMOs, can be found in [Devlin et al. 2002].

Histogram Adjustment (*H*) [Larson et al. 1997]
The operator aims to produce images which preserve visibility in HDR scenes. It also includes models of human contrast sensitivity, color sensitivity, visual acuity and glare, producing images which match the viewer's experience of the real scene.

Bilateral Filter (*B*) [Durand and Dorsey 2002]
This local, detail-preserving operator attempts to display HDR images by decomposition of the image into a base and detail layer. In the base layer the contrast is compressed by an edge-preserving filter known as the bilateral filter.

Photographic Reproduction (*P*) [Reinhard et al. 2002]
This operator simulates the dodging and burning technique used in traditional photography allowing different exposures across the image to be printed. We considered the more complex local model rather than the simpler global version.

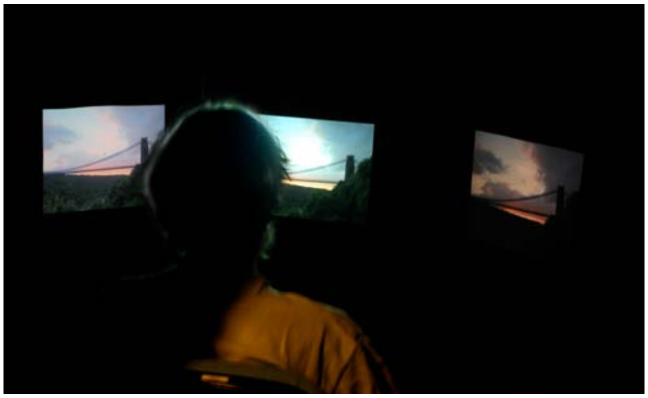


Figure 2: Experimental setup. In the center is the linearly mapped reference image on the HDR display. Left and right are two tone mapped images.

iCAM model (*I*) [Johnson and Fairchild 2003]
iCAM is an image appearance model which has been extended to render HDR images for display. iCAM attempts to determine the perceptual response towards spatially complex stimuli and can predict the appearance of HDR images.

Logarithmic Mapping (*L*) [Drago et al. 2003a]
This method reduces the contrast ratio by a logarithmic compression of luminance values, imitating the human response to light.

Local Eye Adaptation (*A*) [Ledda et al. 2004c]
This recent operator, similarly to [Pattanaik et al. 2000], computes the eye's retinal response to luminance however in this case, the process is entirely localized allowing for a good dynamic range compression.

4 Experimental Framework

Due to the nature of the experiment and task we opted to avoid methods, used in earlier studies, such as rating or ranking. Our approach, on the other hand, was to present each subject with a pair of tone mapped images in addition to the reference image on the HDR display (in the center), as illustrated in Figure 2. Participants were then required to indicate a preference, based on a specific property being tested, for one of the two images compared to the reference. This technique is known as *paired comparisons*. Rating or ranking all of the images would, in this case, be an unnatural task for the observer leading to distorted results [Siegel and Castellan 1988]. The validity and reliability of rating data are problematic to establish without very large numbers of trials and participants and if subjects have not been trained, prior to the trial, on a series of test images [Kendall 1975]. Additionally, rating data are subject to drift effects, leading to likely order effects given relatively small numbers of presentations. Given this, one can argue that ranking may be more appropriate. In ranking a number of test images are arranged in order according to some quality which they all possess to a varying degree. Ranking may be regarded as a less accurate method of expressing ordered relationship of test images (and indirectly tone mapping algorithms) since it does not tell us how close the various images may be to each other. On the other hand, rankings are invariant under stretching: what ranking loses in accuracy

	tmo_1	tmo_2	tmo_3	tmo_4	tmo_5	tmo_6	Score
tmo_1	-	1	0	0	1	1	3
tmo_2	0	-	0	1	1	0	2
tmo_3	1	1	-	1	1	1	5
tmo_4	1	0	0	-	0	0	1
tmo_5	0	0	0	1	-	1	2
tmo_6	0	1	0	1	0	-	2

Table 1: Example preference matrix for one subject when shown six tone mapped images of a given scene. Each tone mapped image in a row, tmo_i is compared with another tmo_j in each column.

it gains in generality for when we stretch the scale of measurement the ranking remains unaltered [Kendall 1975]. Ranking a series of images, however, is still a complicated task for participants, especially when they are simply asked to rank the TMOs in order of overall similarity to the reference, as the participants' judgment will very likely be based on different factors.

The advantage of our approach is not only simplicity, since subjects only have to make straightforward judgments, but it also allows an evaluation of the *transitivity*, that is, the within-subject consistency of the data, as well as the between-subject consistency. This will be discussed in more detail in section 4.2.2. All the experiments described in this paper were carried out using this method and, as will be shown in the following sections, it was a reliable and useful technique.

For clarity, from now on, we will use the term *Scene* to refer to the different images displayed on the HDR display. In our experiment we had 23 scenes and a total of 138 different images (6 tone mapped images per scene).

4.1 Experimental Design

We conducted what is known as a *balanced design paired comparison* test, where each subject was instructed to evaluate all possible comparison pairs taken from the test set. Although this means the trial is large and time consuming, it makes it easier to evaluate and compare the performance of each test subject.

Let us suppose that t is the number of tone mapping operators that we wish to compare against each other and the reference HDR scene (in our case 6). For a given scene, each subject is presented $\binom{t}{2} = 15$ pairs (($t(t-1)/2$), all possible combinations of TMOs. For each pair, the subject's "vote" is recorded. Once all of the pairs have been presented, we may record the results in a $t \times t$ matrix.

As an example, consider the results shown in table 1. The cell in column tmo_2 and row tmo_3 has a value of 1 signifying that the subject considered the image generated with tmo_3 to be more similar to the reference than the image generated with tmo_2. We may also write this as $tmo_3 \rightarrow tmo_2$. From Table 1 we also see that tmo_1 is considered more similar to the reference than tmo_2, tmo_5 and tmo_6 giving an overall score of 3. If we denote p_i as the number of preferences scored by tmo_i ($i = 1, 2..t$), then the overall score per scene per subject is:

$$\sum_{i=1}^{t} p_i = \frac{t(t-1)}{2} = 15 \tag{1}$$

The votes for all s subjects performing the task are then combined into a single *preference matrix* per scene. If all of the subjects completely agreed in their paired comparisons, then $t(t-1)/2$ cells would have a value of s and the remaining cells would have 0. Note that the central diagonal is never considered since we do not compare the same image against itself.

4.2 Statistical Analysis

4.2.1 Kendall Coefficient of Agreement

As just described, there will be a complete agreement if all subjects vote the same way. Suppose that p_{ij} is the number of times that tmo_i is preferred to tmo_j. Now, let

$$\Sigma = \sum_{i \neq j} \binom{p_{ij}}{2} \tag{2}$$

the summation extending over $t(t-1)$ terms (excluding the diagonal), where, as above, t is the numbers of TMOs. Σ is the sum of the number of agreements between pairs. Kendall and Babington-Smith [Kendall and Babington-Smith 1940] have proposed a *coefficient of agreement* among the subjects defined as:

$$u = \frac{2\Sigma}{\binom{s}{2}\binom{t}{2}} - 1 \tag{3}$$

where, s is the number of subjects. As with other correlation methods, u will be equal to 1 if all s subjects made identical choices during the test. The smaller the agreement between subjects, the smaller u will become. The minimum value that u can assume when the scores are evenly distributed across the matrix, is $-1/(s-1)$ when the number of subjects is even or $-1/s$ when s is odd. Therefore, for each scene we can compute the coefficient of agreement which will give us a good indication about the similarity of votes between subjects. Paired comparison data is often analysed using Thurstone's Law of Comaparative Judgments [Thurstone 1927]. Thurstone's method would be appropriate if one would assume that there exist perceptible differences between the TMO's presented for comparison. Kendal coefficient of agreement makes it possible to dispense with such assumptions and precautions.

4.2.2 Coefficient of Consistency

One important aspect to consider when carrying out paired comparisons is *consistency* or *transitivity*. If, when evaluating three objects (TMOs in our case) A, B and C for example, an observer expresses his/her judgment as $A \rightarrow B$ (the arrow means A is closer to the reference than B), $B \rightarrow C$ and $A \rightarrow C$ we define the vote as consistent. On the other hand if the observer makes an inconsistent choice such as $C \rightarrow A$ then we call the triad *circular* and say that the pair comparison is intransitive.

Although inconsistency is not ideal, it can frequently happen, especially in cases, such as ours, where a typical ranking approach is problematic. Inconsistency does not necessarily mean that the data are erroneous. On the contrary, it can provide the tester with very useful information about the experiment. If, on average, most of the participants are inconsistent, we can conclude that the tone mapped images being evaluated are very similar and thus it is difficult to make consistent judgements. (ranking or rating would not consider this). On the other hand, if the inconsistencies are present in only a small proportion of participants, then we can conclude that they are not capable of making a consistent judgment and we have greater justification for not including their scores. In this study, however, no scores had been discarded as the vast majority of participants were consistent in their judgments. We may define a *coefficient of consistency* ζ by the equation [Kendall and Babington-Smith 1940]:

$$\zeta = 1 - \frac{24c}{t^3 - 4t} \tag{4}$$

where c is the number of circular triads observed per participant per test scene. If ζ is 1, then there are no circular triads, the data in this case could be ranked. The number of circular triads is determined as follows [David 1969]:

$$c = \frac{t}{24}(t^2 - 1) - \frac{1}{2}T \tag{5}$$

where $T = \sum(p_i - (t-1)/2)^2$.

For an even t, the maximum number of circular triads is $\frac{1}{24}(t^3 - 4t)$. The coefficient ζ will move to zero as the number of circular triads, and thus the inconsistencies, increases.

Having computed the coefficient of agreement u and consistency ζ for each subject it is important to test the significance values by considering the distribution they would have if all the preferences had been chosen at random. Details of these tests can be found in the Appendix.

5 The Study

We conducted two separate experiments. In the first experiment, the subjects were asked to select which image was overall the most like the reference image, while in the second experiment, participants were asked to make their judgment based on detail reproduction.

The selection of a suitable set of test scenes is crucial to this investigation, because the TMO performance might be scene dependent. One key issue in earlier studies by [Drago et al. 2003b; Ledda et al. 2004b; Kuang et al. 2004; Yoshida et al. 2005] was the limited number of test scenes, between 2 and 8, which might have missed correlation with scene content. We chose to use a larger test set of 23 scenes from a variety of categories, day, night, indoor and rendered. In addition, the scenes contained a variety of different dynamic ranges, all of which were within the limits of the HDR display. The scenes used are all shown in Figure 6. We shall call the scenes *Scene 1, Scene 2, ..., Scene 23*.

5.1 Method

Each scene was tone mapped with the six algorithms and each possible pair combination for a scene shown to the participants. For six TMOs this was a total of 15 images per scene. On every occasion the participant saw three images: in the center the reference scene displayed on the HDR display and to the left and right the tone mapped versions of the reference generated with different operators.

The tone mapped images were displayed on two 15" Viglen LCD monitors which had been calibrated. The resolution of the two LCDs and HDR display was 1024×768 at 60 Hz which is the upper limit of the HDR device used. Both LCDs and HDR display were measured in a dark room with a Minolta CS-100A photometer. The readings were taken by displaying respectively a pure white and black image at full screen. The measurements, made at a 0° angle, were taken in five regions and averaged. These are the readings that we obtained: LCD Max Lum=87.3 cd/m^2, Min Lum=0.65 cd/m^2 (very similar for both displays). Therefore contrast ratio of approx 135:1. For the HDR display, we measured the following: Max=2600 cd/m^2 and Min=0.04, contrast=65,000:1

The viewing angle of the HDR display is around 40° horizontal and 15° vertical, which is very small compared to the 160° horizontal and 120° vertical of the LCD displays. We ensured that each participant's eyes were aligned with the centre of the HDR screen, which was easily achieved by adjusting the height of the chair they were sitting on. The viewing distance was 80 cm. The adjacent displays were positioned at the exact same height and distance but slightly rotated, around the vertical axis. To observe the tone mapped images on these displays, they simply rotated their head enough to form a 90° angle between viewing direction and screen. All of the screens were behind a dark gray mask.

The experiment was conducted in a dark room to avoid any effects of ambient lighting. We allowed each participant to adjust to the environment for 5 minutes before commencing the actual experiment. Presentation order and location were randomized in order to remove any order effects. The maximum time allowed to make a choice between the two tone mapped images was 13 seconds which was decided after a pilot study. It is important to allow the same amount of time for each participant and it should be long enough for the participant to make an informed choice without analyzing every single detail in the image. The total number of participants for the entire study (Experiments 1 and 2) was 109, 44 female and 65 male. All of the participants were between 20 and 34 years old and had normal or corrected to normal vision. The participants had taken at one computer graphics course and, therefore, had a clear understanding of their task. No subject took part in more than one experiment.

	P	*H*	*B*	*L*	*I*	*A*	Total
P	-	24	46	42	10	32	154
H	24	-	44	32	8	12	120
B	2	4	-	8	2	4	20
L	6	16	40	-	4	12	78
I	38	40	46	44	-	38	206
A	16	36	44	36	10	-	142

Table 2: Preference matrix for *Scene 8*.

5.2 Experiment 1: Overall Similarity

In Experiment 1, each participant was presented with all of the scenes, thus looking at a total of $15 \times 23 = 345$ image pairs. Because of the large amount of data, we split each participant's experiment in three sessions of 30 minutes reducing the risk of observers getting tired or bored. Participants were asked to make judgements of the TMOs based on *overall similarity*. For each pair, participants were instructed to observe the two tone mapped images and select the one they believed was the most similar to the reference HDR scene. This was deliberately a vague task. We wanted to investigate whether the outcome would be somewhat random, indicating that various attributes such as color, contrast or detail cancel each other out or if, even in this generic situation, some consistency in the data would be noticed. 48 subjects took part in this 90 minute experiment.

5.3 Experiment 1: Results

The detailed nature of this study has, of course, resulted in a significant amount of data. To illustrate our methodology and statistical analysis, we will discuss in detail the results for one scene, *Scene 8*. The overall comparison results for the 23 scenes are shown in Table 3. We have also used colors for each of the TMOs to make it easier to discern patterns in the data. The complete data set, including all multiple comparison scores, will be made available on a website.

The outcome of the paired comparison data from the 48 subjects was tabulated in a preference matrix. The preference matrix for *Scene 8* is shown in Table 2. The numbers in each cell represent the number of times that a specific tone mapped image was regarded as being closer, in overall similarity, to the reference. For example, the fourth cell in the first row is 42 indicating that algorithm *P* was judged 42 times out of 48 closer to the reference than algorithm *L* (Note that *L*, row four, thus has the value of 48-42=6).

Prior to preparing the preference matrix, each participant's results were analyzed for consistency using Equation 4. For the case presented in Table 2, the *average* coefficient of consistency $\zeta = 0.842$ which, given degrees of freedom and p value, is high and statistically significant. A high value of ζ indicates that the

	Coeff Agr u	*Coeff Cons (ave) ζ*	χ^2	*significance p, 15 df*	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**
Scene 1	0.050	0.533	50.0	< 0.1	*P*	*B*	*A*	*H*	*I*	*L*
Scene 2	0.214	0.692	166.0	< 0.001	*I*	*P*	*H*	*A*	*B*	*L*
Scene 3	0.254	0.817	194.0	< 0.001	*P*	*I*	*A*	*H*	*L*	*B*
Scene 4	0.172	0.767	136.0	< 0.001	*P*	*L*	*I*	*A*	*H*	*B*
Scene 5	0.523	0.933	384.0	< 0.001	*I*	*H*	*A*	*P*	*L*	*B*
Scene 6	0.429	0.892	317.6	< 0.001	*I*	*H*	*A*	*P*	*L*	*B*
Scene 7	0.189	0.692	148.0	< 0.001	*I*	*A*	*P*	*H*	*B*	*L*
Scene 8	0.429	0.842	317.6	< 0.001	*I*	*P*	*A*	*H*	*L*	*B*
Scene 9	0.062	0.650	58.6	< 0.05	*P*	*A*	*L*	*H*	*B*	*I*
Scene 10	0.112	0.725	94.0	< 0.05	*I*	*P*	*H*	*A*	*B*	*L*
Scene 11	0.228	0.775	175.6	< 0.001	*I*	*A*	*H*	*P*	*B*	*L*
Scene 12	0.337	0.858	252.6	< 0.001	*I*	*P*	*H*	*A*	*L*	*B*
Scene 13	0.081	0.642	72.0	< 0.001	*P*	*I*	*H*	*A*	*L*	*B*
Scene 14	0.314	0.979	236.3	< 0.001	*I*	*P*	*A*	*H*	*L*	*B*
Scene 15	0.074	0.617	67.3	< 0.05	*H*	*P*	*I*	*A*	*L*	*B*
Scene 16	0.310	0.850	233.3	< 0.001	*H*	*P*	*I*	*A*	*B*	*L*
Scene 17	0.241	0.850	184.6	< 0.001	*I*	*H*	*P*	*A*	*B*	*L*
Scene 18	0.220	0.875	170.0	< 0.001	*P*	*A*	*H*	*I*	*L*	*B*
Scene 19	0.148	0.725	119.3	< 0.001	*I*	*H*	*A*	*P*	*B*	*L*
Scene 20	0.182	0.658	143.0	< 0.001	*P*	*I*	*A*	*B*	*L*	*H*
Scene 21	0.138	0.717	112.3	< 0.001	*P*	*H*	*A*	*I*	*B*	*L*
Scene 22	0.107	0.575	90.6	< 0.05	*P*	*I*	*A*	*H*	*B*	*L*
Scene 23	0.282	0.808	214.0	< 0.001	*P*	*A*	*H*	*I*	*L*	*B*

Table 3: Overall Similarity results for Color images.

	Coeff Agr u	*Coeff Cons (ave) ζ*	χ^2	*significance p, 15 df*	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**
Scene 3	0.352	0.789	57.2	< 0.001	*P*	*I*	*A*	*B*	*H*	*L*
Scene 4	0.338	0.933	55.5	< 0.001	*P*	*L*	*A*	*H*	*I*	*B*
Scene 7	0.184	0.911	37.1	< 0.001	*I*	*A*	*H*	*P*	*B*	*L*
Scene 8	0.310	0.889	52.2	< 0.001	*I*	*P*	*A*	*H*	*L*	*B*
Scene 11	0.177	0.689	36.3	< 0.001	*I*	*A*	*H*	*P*	*L*	*B*
Scene 13	0.380	0.767	60.5	< 0.001	*I*	*P*	*A*	*L*	*H*	*B*
Scene 14	0.498	0.956	74.8	< 0.001	*I*	*P*	*A*	*L*	*H*	*B*
Scene 16	0.359	0.844	58.0	< 0.001	*H*	*I*	*P*	*A*	*L*	*B*
Scene 18	0.400	0.822	63.1	< 0.001	*P*	*L*	*H*	*A*	*I*	*B*
Scene 21	0.066	0.756	22.9	< 0.001	*P*	*H*	*A*	*B*	*L*	*I*
Scene 23	0.428	0.889	66.4	< 0.001	*P*	*H*	*L*	*A*	*I*	*B*

Table 4: Overall Similarity results for Greyscale images.

TMOs, at least for this scene, could be indirectly ranked. Having established good consistency in the data, we compute, using Equation 3, the coefficient of agreement u amongst subjects. For *Scene 8*, $u = 0.429$. Complete agreement would exist if half of the cells in Table 2 contained 48 and the remaining 0; the data, however, could still be inconsistent. The significance test, as described in the Appendix, shows that we may reject the null hypothesis H_0 at $\alpha = 0.05$ level for $\binom{6}{2} = 15$ degrees of freedom (df) and we can thus conclude that there is some agreement amongst observers when comparing the different tone mapped images to the reference. This implies that there is indeed a perceptual difference between the TMOs, unfortunately it does not tell us where these differences lie. This is the equivalent problem of running an ANOVA test, where usually post-hoc tests need to be conducted to compare each test object against each other. The same concept applies here. Significance test of the *score differences* need to be performed in order to determine whether the perceptual quality of any two algorithms from the test set is perceived as different. Otherwise, we have to conclude that the perceived quality of the two operators is similar. From the *multiple comparison score* test, described in the Appendix, we can test the score of each operator against the others.

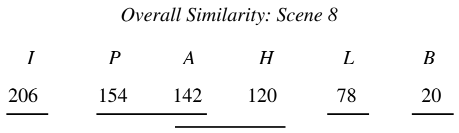


Figure 3: Multiple comparison score for *Scene 8*. Any TMOs whose scores are underlined are considered perceptually similar to the reference. So, for example, P is perceptually similar to A, but not to any of the others.

For $t = 6$ algorithms, 15 comparisons are made. One method to represent such results is as shown in Figure 3 where any TMOs that are underlined by the same line may be considered perceptually indistinguishable. This procedure needs to be repeated for all 23 scenes. There is no succinct way to show the results from the multiple comparisons scores.

Table 3 contains the complete results for the overall similarity experiment. For each scene, the table shows the values of u, average ζ, significance p and the TMO ranking. From these results it is reasonably clear that, on average, TMO *I* followed by *P* were considered the closest to the reference scene in overall similarity. The ranking is very similar across the 23 scenes. We can also see that *B* and *L* did not perform as well. Although the task was intentionally vague, the outcome was consistent and the results are in the vast majority of cases statistically significant.

Figure 4 illustrates the multiple comparison score results when *all* of the scores from the 23 different scenes were added together. Apart from *H* and *A*, which can thus be considered perceptually similar, there is a strong statistically significant difference between the other TMOs and thus these TMOs can clearly be ranked.

Overall Similarity: Color

I	*P*	*H*	*A*	*L*	*B*
3712	3402	2994	2852	1902	1696

Figure 4: Comparison of the scores across all 23 scenes.

5.4 Overall Similarity: Grayscale Images

Following the first experiment, we conducted another experiment with $s = 18$ subjects and a random subset of 11 of the 23 scenes, where the task was identical as previously, however, on this occasion, the images presented were grayscale. These images were generated from the luminance channel of the color images. We wanted to verify if color played a significant role when comparing the various algorithms. The results for this experiment are presented in Table 4. On the whole the outcome had a similar trend although the significance between the individual algorithms was smaller. In particular we noticed that in this experiment *P* performed slightly better (although not significantly) than *I*. This reversal of rank could be explained by the fact that the *I* algorithm is a fairly sophisticated color appearance model. Therefore, when color is present, the tone mapped image generated preserves the color information closer to the reference than other algorithms such as *P*. On average, however, the scores between the color and grayscale experiments are comparable indicating that color has no significant effect on the overall ranking of the TMOs.

5.5 Experiment 2: Detail Reproduction

The results of Experiment 1 showed that there was a strong and consistent agreement between subjects. In a second experiment we wanted to investigate, if when the focus on the image was more specifically on the *visibility and reproduction of detail*, whether the outcome would be analogous. Most tone mapping operators aim to reproduce the maximum amount of detail in the scene sacrificing contrast at times. Humans, on the other hand, may make assessments of scenes mainly based on contrast and color and therefore, we would expect a different outcome when isolating detail as the attribute to judge. Local TMOs, although theoretically capable of better detail reproduction, tend to suffer in terms of contrast when compared to spatially-uniform algorithms. We conducted two sub-experiments with 48 participants. In the first part of Experiment 2, using the same methodology of paired comparisons, participants were required to assess images based on reproduction of detail in the bright regions of the images. In the second, the task was to assess the tone mapped images based on reproduction of detail in the dark regions. The purpose of this experiment was to ensure that TMOs which generate too much detail should be considered as poor as those which do not reproduce enough detail.

5.6 Experiment 2: Bright Regions Results

Table 5 shows the results for the reproduction of detail in bright regions. We can immediately notice that although the ranking for the first two positions are fairly obvious, the remaining positions are not. More specifically we can see that the *P* algorithm, which performed very well in both *overall similarity* experiments, does not perform well when considering the reproduction of detail in brighter regions of the scenes. Not surprisingly *L*, being a spatially-uniform model, performed on average the worse followed by *H*. TMO *B*, which perhaps surprisingly performed poorly in the previous experiment, does much better here. As in Experiment 1, although the overall results for each scene are statistically significant, the score differences between individual algorithms might not be. On average, however, we observed statistically significant differences between operators ranked more than one position apart.

5.7 Experiment 2: Dark Regions Results

Table 6 shows the results of the reproduction of detail in dark regions experiment. Unlike the bright detail experiment, there is an improvement in performance of TMO *P*. As with bright regions, on average *A* obtained high scores and was typically ranked in the first two positions. The *B* algorithm was in most cases ranked the worst.

6 Conclusion

The overall results of our investigation are summarized in Figure 5. The values represent the sum across all subjects and test scenes. Each algorithm was then tested against each other to verify whether they belonged to the same perceptual group or, if indeed one was perceived to be closer to the reference than the other.

In the first experiment we asked subjects to observe pairs of tone mapped images and choose the one which appeared closer to the reference in overall similarity. Given the purposefully inexact definition of the task, leaving participants to set their own judgement criteria, it is surprising how well participants agreed. Furthermore, the value of ζ was high enough to conclude that the task of indirectly ranking the algorithms was possible (average coefficient of consistence $\zeta_{ave} \approx 0.7$). From this first experiment we can conclude that the iCAM (*I*) and Photographic (*P*) operators consistently performed very well. When the same experiment was conducted with grayscale scenes ($\zeta_{ave} \approx 0.7$), the outcome was similar, although we noticed that on average *P* performed slightly better than *I*. We might explain this by the fact that the iCAM is a good color appearance model and when participants observe the images as a whole, color plays a role. When color was absent, participants used other attributes such as contrast to assess the images, in this case *I*'s performance suffered slightly. For the other operators color did not have a major impact in assessing image quality.

In the second experiment, the reproduction of features and detail was tested. Firstly, from Tables 5 and 6, we can see that the average consistency for each scene was higher than when the overall comparisons were made ($\zeta_{ave} \approx 0.9$ both bright and dark experiments); even the overall agreement u was much higher. This is not surprising since the task was more specific, allowing less freedom in judgements. As it can be seen from Table 5 and Figure 5, when the bright details were tested, the iCAM *I* model scored higher than the others followed by the local adaptation model *A*, while the performance of the photographic *P* algorithm was not as good as in previous experiments. In this case the bilateral filter *B* algorithm performs slightly better although surprisingly low once more. In the dark detail experiment *A* again achieved a high score indicating that this algorithm is capable of accurate detail reproduction at the expense of lower contrast, which was confirmed from the overall similarity experiments.

	Coeff Agr u	*Coeff Cons (ave) ζ*	χ^2	*significance p, 15 df*	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**
Scene 4	0.115	0.661	52.8	< 0.02	*H*	*A*	*I*	*P*	*B*	*L*
Scene 7	0.295	0.878	112.3	< 0.001	*I*	*A*	*B*	*H*	*P*	*L*
Scene 11	0.275	0.765	105.7	< 0.001	*I*	*A*	*B*	*H*	*P*	*L*
Scene 12	0.302	0.861	114.7	< 0.001	*I*	*A*	*P*	*B*	*L*	*H*
Scene 13	0.214	0.783	85.5	< 0.001	*L*	*A*	*P*	*I*	*B*	*H*
Scene 14	0.331	0.887	124.1	< 0.001	*I*	*B*	*A*	*L*	*H*	*P*
Scene 16	0.311	0.904	117.5	< 0.001	*H*	*P*	*I*	*B*	*A*	*L*
Scene 18	0.381	0.870	140.8	< 0.001	*I*	*A*	*H*	*P*	*L*	*B*
Scene 21	0.351	0.878	130.7	< 0.001	*I*	*A*	*P*	*B*	*H*	*L*
Scene 22	0.274	0.791	105.3	< 0.05	*I*	*P*	*B*	*A*	*H*	*L*

Table 5: Results of the Reproduction of Detail in Bright regions experiment.

	Coeff Agr u	*Coeff Cons (ave) ζ*	χ^2	*significance p, 15 df*	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**
Scene 4	0.444	0.9	141.4	< 0.001	*P*	*A*	*L*	*B*	*I*	*H*
Scene 7	0.434	0.89	138.8	< 0.001	*A*	*P*	*I*	*H*	*L*	*B*
Scene 11	0.233	0.89	81.4	< 0.001	*A*	*P*	*L*	*H*	*I*	*B*
Scene 12	0.382	0.9	123.8	< 0.001	*I*	*P*	*A*	*L*	*H*	*B*
Scene 13	0.255	0.86	87.8	< 0.001	*H*	*I*	*A*	*P*	*L*	*B*
Scene 14	0.341	0.93	112.2	< 0.001	*I*	*P*	*H*	*A*	*L*	*B*
Scene 16	0.258	0.89	88.4	< 0.001	*A*	*I*	*P*	*H*	*L*	*B*
Scene 18	0.291	0.88	97.8	< 0.001	*P*	*A*	*L*	*H*	*I*	*B*
Scene 21	0.232	0.88	81.0	< 0.001	*A*	*P*	*L*	*B*	*H*	*I*
Scene 22	0.300	0.85	100.4	< 0.001	*P*	*A*	*L*	*I*	*B*	*H*

Table 6: Results of the Reproduction of Detail in Dark regions experiment.

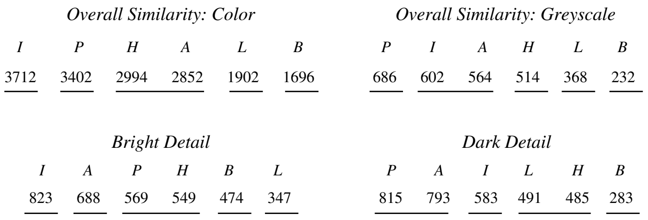


Figure 5: Summary of the entire investigation. Any two TMOs whose scores are underlined are considered perceptually similar.

From our results it appears that the iCAM *I* model generally performs better than other algorithms when participants are asked to make simple comparisons with the ideal reference image. Color is probably a major factor here as when grayscale images were presented to participants, the photographic *P* model performed better. This possibly signifies that *P* has a very good contrast appearance whereas *I*, although still good in contrast reproduction, has the advantage of a better color model. When the more specific attributes of details were examined, we noticed a different trend. The local eye adaptation model *A* generally can be considered the algorithm which best preserves details. Overall the histogram *H* operator had a quite good performance especially considering that it is a global model. Less impressive was the logarithmic mapping operator *L*, although it is capable of good detail reproduction in dark regions. Surprisingly, the bilateral algorithm *B* performed very poorly, the images generated by this TMO tend to have very high contrast and detail, even more than the reference image, which explains the poor score. However, when the images generated by this algorithm are viewed without a reference, subjects tend to prefer these images compared to other algorithms. This was also reported in [Kuang et al. 2004]. Finally, from our results, we did not find any correlation between tone mapping operators and test scenes.

Future work will evaluate further TMOs using not only the methodology we have proposed, but also other approaches including the use of eye-tracking to compare how participants observe the different scenes. This may aid us in comprehending the reasons behind our results. The knowledge gained from these results will be used to develop a new tone mapping operator for traditional displays providing the best scene preservation of all attributes including contrast, color and detail. In addition, tone mapping operators will also be developed for HDR displays.

7 Acknowledgements

We would like to thank the authors of the TMOs for their help with the code and images and the participants for taking part in the study. We would also like to thank the reviewers for their comments and Peter Shirley, Teresa, Mashhuda, Eva and MV for reading early drafts of the paper. The work reported in this paper has formed part of the Rendering on Demand (RoD) project within the 3C Research program whose funding and support is gratefully acknowledged.

References

David, H. A. 1969. *The Method of Paired Comparisons*. Charles Griffin and Company. London.

Devlin, K., Chalmers, A., Wilkie, A., and Purgathofer, W. 2002. Star report on tone reproduction and physically based spectral rendering. In *Eurographics 2002*,

Drago, F., Myszkowski, K., Annen, T., and Chiba, N. 2003. Adaptive logarithmic mapping for displaying high contrast scenes. In *Proc. of EUROGRAPHICS 2003*, P. Brunet and D. W. Fellner, Eds., vol. 22 of *Computer Graphics Forum*, 419–426.

Drago, F., , Martens, W. L., Myszkowski, K., and Seidel, H.-P. 2003. Perceptual evaluation of tone mapping operators. In *ACM SIGGRAPH Conference Abstracts and Applications*.

Durand, F., and Dorsey, J. 2000. Interactive tone mapping. In *Rendering Techniques 2000 (Proceedings of the Eleventh Eurographics Workshop on Rendering)*, B. Peroche and H. Rushmeier, Eds., 219–230.

Durand, F., and Dorsey, J. 2002. Fast bilateral filtering for the display of high-dynamic-range images. In *In Proceedings of ACM SIGGRAPH '02*, ACM Press, 257–266.

Fattal, R., Lischinski, D., and Werman, M. 2002. Gradient domain high dynamic range compression. In *In Proceedings of ACM SIGGRAPH '02*, ACM Press, 249–256.

Ferwerda, J. A., Pattanaik, S. N., Shirley, P., and Greenberg, D. P. 1996. A model of visual adaptation for realistic image synthesis. In *In Proceedings of ACM SIGGRAPH '96*, ACM Press, 249–258.

Johnson, G., and Fairchild, M. 2003. Rendering hdr images. *In Proceedings of IS & T/SID 11th Color Imaging Conference*, 36–41.

Kendall, M. G., and Babington-Smith, B. 1940. On the method of paired comparisons. *Biometrika 31*, 324–345.

Kendall, M. 1975. *Rank Correlation Methods, 4th ed.* Griffin Ltd.

Kuang, J., Yamaguchi, H., Johnson, G., and Fairchild, M. 2004. Testing hdr image rendering algorithms. In *In proceedings of IS and T/SID 12th Color Imaging Conference*.

Larson, G. W., Rushmeier, H., and Piatko, C. 1997. A visibility matching tone reproduction operator for high dynamic range scenes. *IEEE Transactions on Visualization and Computer Graphics 3*, 4, 291–306.

Ledda, P., Chalmers, A., and Seetzen, H. 2004. Hdr displays: a validation against reality. *In IEEE International Conference on Systems, Man and Cybernetics*.

Ledda, P., Chalmers, A., and Seetzen, H. 2004. A psychophysical validation of tone mapping operators using a high dynamic range display (poster). *Symposium on Applied Perception in Graphics and Visualization*.

Ledda, P., Santos, L. P., and Chalmers, A. 2004. A local model of eye adaptation for high dynamic range images. In *In Proceedings of ACM AFRIGRAPH '04*, ACM Press, 151–160.

Pattanaik, S. N., Ferwerda, J. A., Fairchild, M. D., and Greenberg, D. P. 1998. A multiscale model of adaptation and spatial vision for realistic image display. In *In Proceedings of ACM SIGGRAPH '98*, ACM Press, 287–298.

Pattanaik, S. N., Tumblin, J., Yee, H., and Greenberg, D. P. 2000. Time-dependent visual adaptation for fast realistic image display. In *In Proceedings of ACM SIGGRAPH 2000*, ACM Press, 47–54.

Pearson, E. S., and Hartley, H. O. 1988. Biometrika tables for statisticians, 3rd ed. Cambridge University Press, vol. 1.

Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J. 2002. Photographic tone reproduction for digital images. In *In Proceedings of ACM SIGGRAPH '02*, ACM Press, 267–276.

Schlick, C. 1994. Quantization techniques for the visualization of high dynamic range pictures. In *Eurographics Workshop on Rendering*, 7–20.

Seetzen, H., Whitehead, L., and Ward, G. 2003. A high dynamic range display system using low and high resolution modulators. In *Proc. of the 2003 Society for Information Display Symposium*.

Siegel, S., and Castellan, N. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGrall-Hill International.

Thurstone, L. L. 1927. A law of comparative judgment. *Psychological Review 34*, 273–286.

Tumblin, J., and Rushmeier, H. 1993. Tone reproduction for computer generated images. *IEEE Computer Graphics and Applications 13*, 6, 42–48.

Tumblin, J., and Turk, G. 1999. LCIS: A boundary hierarchy for detail-preserving contrast reduction. In *In Proceedings of ACM SIGGRAPH 99*, A. Rockwood, Ed., 83–90.

Tumblin, J., Hodgins, J., and Guenter, B. 1999. Two methods for display of high contrast images. *ACM Transactions on Graphics 18*, 3, 56–94.

Upstill, S. D. 1985. *The realistic presentation of synthetic images: image processing in computer graphics*. PhD thesis.

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

Yoshida, A., Blanz, V., Myszkowski, K., and Seidel, H.-P. 2005. Perceptual evaluation of tone mapping operators with real-world scenes. In *SPIE*.

Appendix

To test the significance of the coefficient of agreement u, we may test the null hypothesis H_0 that there is no agreement amongst the subjects and the alternative hypothesis H_1 that the degree of agreement is greater than if the evaluation of the comparisons had been done randomly. To determine the significance of u we may use the large sample approximation to the sampling distribution. The chi-squared (χ^2) test statistics is [Siegel 1998]:

$$\chi^2 = \frac{t(t-1)(1+u(s-1))}{2} \tag{6}$$

which is asymptotically distributed as χ^2 with $t(t-1)/2$ degrees of freedom. We can determine whether the values we obtained were statistically significant from tables of probability value for χ^2 as given in [Siegel 1998](Table C, page 323). Note that a high agreement u between subjects does not necessarily imply a high consistency ζ. It is possible that all the judges agree and at the same time are completely inconsistent.

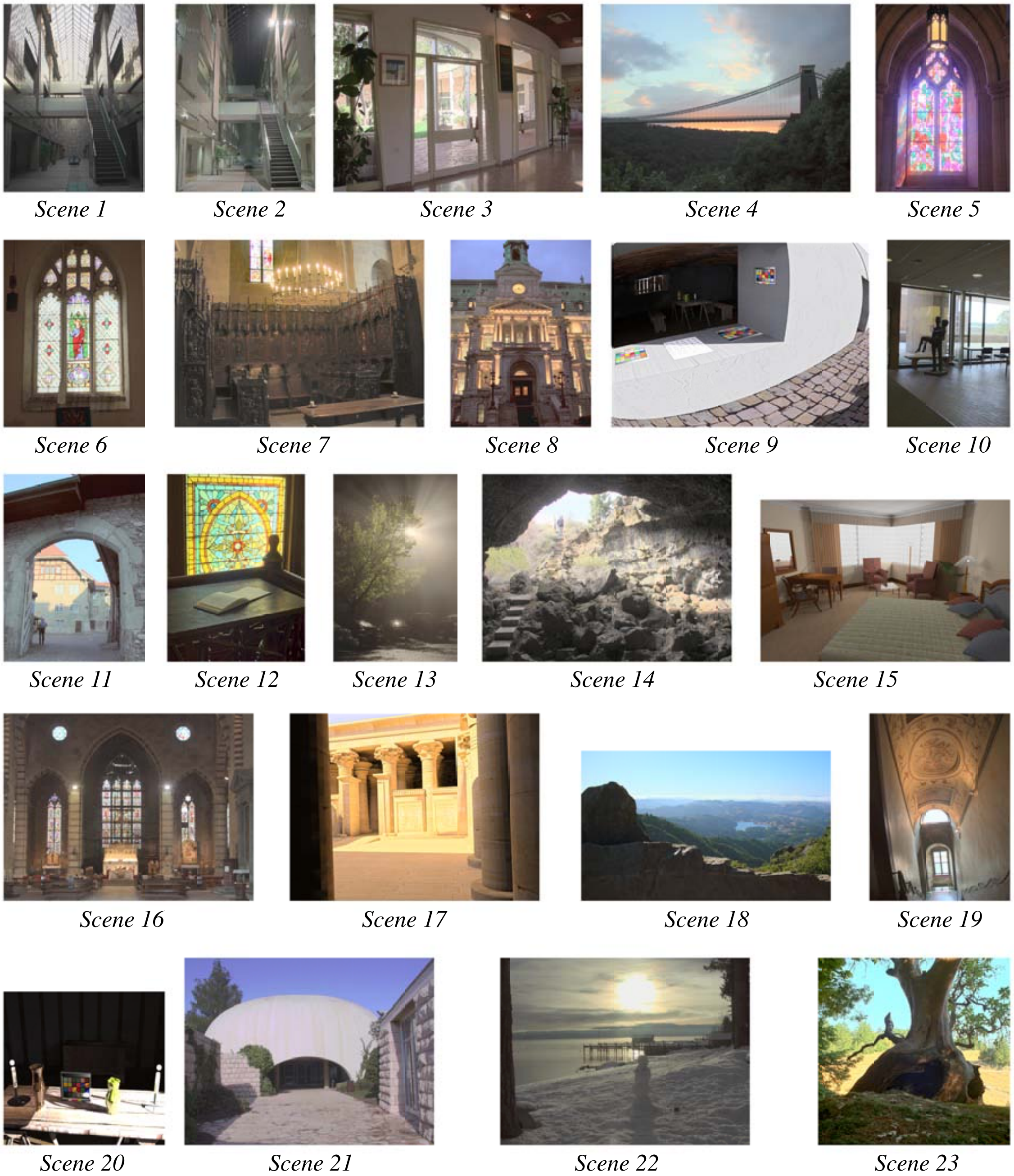
If u is statistically significant then we can say that there are differences between the operators although we do not know where these differences lie. Significance test of the *score differences* is performed in order to see whether the perceptual quality of any two algorithms from the test set is perceived as different. Otherwise, we may have to conclude that the perceived quality of the two operators is similar. In other words, we want to find R' such that the probability $P(R \geq R')$is less or equal to the significance level α (usually α=0.05). We declare the TMOs within each group (scores difference $< R$) to be not significantly different, while those from different perceptual groups are declared to be significantly different. The distribution of the range R is asymptotically the same as the distribution of variance-normalized range, W_t, of a set of normal random variables with variance = 1 and t samples [David 1969]. Therefore, we can use the following relation:

$$P\left(W_t \geq \frac{2R-1/2}{\sqrt{st}}\right) \tag{7}$$

where $W_{t,\alpha}$ is the value of the upper percentage point of W_t at significance point α. The values of $W_{t,\alpha}$ can be obtained from statistics books for example [Pearson and Hartley 1988]. Once the value of $W_{t,\alpha}$ is determined, we can solve for R':

$$R' = \frac{1}{2}W_{t,\alpha}\sqrt{st} + \frac{1}{4} \tag{8}$$

If the score difference for a given scene between two TMOs is larger than R^+ (the smallest integer greater than R'), then we can conclude that there is a statistically significant difference between the two operators indicating that one is closer to the ideal reference image than the other.



Scene 1 *Scene 2* *Scene 3* *Scene 4* *Scene 5*

Scene 6 *Scene 7* *Scene 8* *Scene 9* *Scene 10*

Scene 11 *Scene 12* *Scene 13* *Scene 14* *Scene 15*

Scene 16 *Scene 17* *Scene 18* *Scene 19*

Scene 20 *Scene 21* *Scene 22* *Scene 23*

Figure 6: The data set of 23 Scenes. *Scene 1 and 2* by Karol Myszkowski. *Scene 3, 5 and 21* by Dani Lischinski . *Scene 4, 6, 7, 8, 11, 14, 16, 18, 19 and 22* Greg Ward. *Scene 10* by Garrett Johnson. *Scene 12 and 23* by Industrial Light and Magic. *Scene 13* by Jack Tumblin. *Scene 15* by Simon Crone. *Scene 17* by Veronica Sundstedt.