J. Vis. Commun. Image R. 24 (2013) 1020–1030

Contents lists available at SciVerse ScienceDirect

J. Vis. Commun. Image R.

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

# Assessment of video tone-mapping: Are cameras' S-shaped tone-curves good enough?

Josselin Petit, Rafał K. Mantiuk *

*School of Computer Science, Bangor University, Dean Street, Bangor LL57 1UT, United Kingdom*

ARTICLE INFO

*Article history:*
Received 16 July 2012
Accepted 17 June 2013
Available online 3 July 2013

*Keywords:*
High dynamic range
HDR video
Tone mapping
Tone mapping evaluation
Quality assessment
Camera response curve
Fidelity
Preference

ABSTRACT

The performance of video tone-mapping operators is investigated in a rating experiment using two criteria: overall quality and fidelity to real-world experience. The study includes a tone-curve used in commercial cameras, rarely considered in tone-mapping evaluation studies. The quality is measured for a range of parameter settings, revealing the importance of parameter fine-tuning and often unsatisfactory results of the default operator parameters. In order to explain what makes best performing operators better, the results are analysed in relation to image statistics and the characteristics of the tone-mapping function. Our observations are: state-of-the-art tone mapping produces measurably better results than camera's S-shaped curve for high dynamic range scenes with important content spanned across a wide dynamic range; differences in colour reproduction strongly affect the results; fidelity and quality criteria produce similar results when no reference is present; and state-of-the-art operators produce the results of comparable quality when their parameters are well selected.



## 1. Introduction

Although tone-mapping is usually associated with high dynamic range (HDR) images, it is in fact a very common operation that can be found in most digital cameras. Whenever a photograph is taken or video is recorded, a camera applies its own tone and color processing in order to map sensor-captured light intensities to the color gamut of a storage format. This process shares similar goals and in many cases uses the same techniques as tone-mapping of HDR images. Yet, these two applications of tone-mapping are rarely considered together and the results of tone-mapping HDR images are rarely compared to the state-of-the-art techniques found in cameras.

Tone mapping operators for HDR images were thoroughly compared in numerous assessment studies, which we are reviewed in Section 2. However, none of these studies considered tone-mapping of video sequences, nor did they include the most prevalent operator, which is the tone curve of a camera. Although image tone-mapping has received much more attention than video tone-mapping, we believe that the latter is even more relevant as it greatly extends the applicability of tone-mapping, making it suitable for such applications as movie postprocessing, computer games, adaptive display devices or camcorders.

This work reports the result of a quality assessment experiment performed on video sequences processed by several tone-mapping operators, including a typical S-shaped tone-curve found in consumer cameras. Our intention is not to benchmark operators with the goal of finding a winner but to answer the following questions:

- Is tone-mapping task-specific? Is there a difference between tone-mapping algorithms that simulate the limitations of the human vision to achieve the best fidelity with memorized reality and those whose main goal is to reduce the dynamic range and produce the most pleasing images?
- Do sophisticated tone-mapping algorithms offer any measurable improvement over a camera tone-curve?
- Can the choice of tone-mapping parameters bias the result of the quality assessment experiment? Do the default parameters reflect actual performance of an operator?
- What makes the best performing operators better? How can quality assessment studies help design better operators?

The remainder of this paper is divided as follows: Section 2 gives an overview of the existing video tone-mapping operators and a summary of the tone-mapping assessment studies. Section 3 describes and justifies the experiment design. Section 4 discusses the results. Finally, Section 5 analyzes the relation between the measured quality and image statistics.

* Corresponding author.
*E-mail addresses:* josselin.petit@gmail.com (J. Petit), mantiuk@bangor.ac.uk (R.K. Mantiuk).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.06.014

## 2. Previous work

### 2.1. Tone mapping for video

In this section we focus solely on tone-mapping operators that are intended for video sequences, as they are the subject of our study. Excellent reviews of tone-mapping algorithms for static images can be found in [32,27].

Ferwerda et al. [12] were the first to introduce models of human light and dark adaptation to tone-mapping. Their method incorporates the model of temporal adaptation into Ward's contrast-threshold tone reproduction operator [38]. Durand and Dorsey [10] further extended this algorithm by simulating flares, after-images and the loss of acuity.

A similar adaptation model was used in the work of Pattanaik et al. [28], where a series of forward and inverse models of adaptation and appearance was used to tone-map video sequences. To reproduce the realistic appearance of input scenes, the forward model assumed scene adaptation while the inverse model assumed display adaptation. The operator inspired another work from [22] on the tone-mapping algorithm that is both locally and temporarily adaptive.

Operators that rely on the forward-and-inverse visual models cannot guarantee that the resulting scene is contained within the dynamic range of the output device. This problem is solved by the method of Irawan et al. [14], which employs Ward's histogram-based operator [39]. The operator distributes contrast compression throughout the image while ensuring that no contrast is enhanced beyond its visibility in the original scene. The temporal processing is made through a model of pigment bleaching and fast neural adaptation similar as in Pattanaik et al. [28], plus a third new component for slow neural adaptation. The operator also models mal-adaptation, predicting lower contrast sensitivity when the eye is not fully adapted.

Van Hateren [36] proposed a sophisticated model of human cones to compress contrast in HDR video sequences. The model is based on the neurophysiological measurements of cell responses to light and was validated in the wide range of light from 0.5 to 2000 td. This is an example of forward-only visual model, which may raise concerns whether an abstract (neural) response of cones is the right quantity to be displayed on a screen.

Focusing on the color reproduction of images, a color appearance model (iCAM) framework was developed and used for tone mapping. The most recent revision of the model (iCAM06) [17] improves the previous revisions by including a bilateral filter, as in [11], and using photoreceptor response functions to compress the base layer dynamic range, assuming local adaptation. Details are then added with the detail layer, which takes into account the Stevens effect. The base and detail are finally combined in the IPT uniform opponent color space, where the Hunt effect is also simulated. We received an updated version of the algorithm from the author with the extensions that make the algorithm capable of processing video sequences in a temporally coherent manner.

While the above discussed operators aim at perceptual realism and reproduce the appearance of the original scene, the display adaptive tone mapping from [23] attempts to minimize changes and introduce the least amount of contrast distortions to the reproduced scene. A global tone-curve is found that minimizes contrast distortion, which in turn is quantified by a visual model. Temporal coherence is achieved by low-pass temporal filtering.

### 2.2. Tone mapping evaluation

Evaluation of tone-mapping for images has received a substantial amount of attention and led to establishing the CIE TC8-08 committee with the goal of testing spatial appearance models, including tone-mapping operators [15]. Because of the large body of literature, we do not discuss each work separately, but rather summarize the most relevant facts about each study in Table 1. We refer to the table later in the text when we compare our results with those studies.

### 2.3. Subjective quality assessment methods

Three different ways of assessment for accuracy or preference were used for the previous experiments: ranking, rating, and paired comparison. Some papers compared the results of these methods: Čadík et al. compared how well the operators performed in rating (with reference) and ranking (without reference) experiments, for different image attributes [7]. They noticed a similar trend for both experimental methods with some minor differences for similarly performing operators. Kuang et al. compared the accuracy paired comparison vs. rating-scaling [20]. They found a very good correlation but also that while the paired comparison is more time-consuming, it also delivers more accurate and precise results.

In this work we employed a direct rating assessment, mainly because a full-pairwise comparison of 20 video sequences would require 190 comparisons for each video clip, making the study impractical.

### 2.4. Tone mapping operator tuning for assessment

Most tone-mapping algorithms have several parameters that can be tuned depending on the input image or the expected result. The selection of these parameters differed from one tone-mapping evaluation study to another: some used default parameters; others run a pilot study in which a set of the "best parameters" was selected; or the authors were asked to tune the parameters of their operators (refer to *Parameters* column in Table 1). In the case of comparison with a reference, some authors also have adjusted each tone mapping parameter to make the resulting images as close as possible to the reference scene [16]. In this work we explore how the variation of parameters affects the results by including those variations in the main experiment rather than in a pre-study.

### 2.5. Objective quality metrics

The major limitation of any subjective quality study is the limited number of scenes that can be tested as the budget of the time the participants can spend on experiments is always limited. For most experimental studies with complex images it is impossible to provide statistical evidence that generalizes beyond the tested set of images. The statistical testing ensures generalization of the results for the population of observers but not for the population of images. However, if a computational metric that could predict the quality of tone-mapped video was available, a much larger set of scenes could be tested.

Little work has been done in the area of automatic quality evaluation of tone-mapped images. There exist metrics for predicting visibility of distortions and their effect on quality in HDR images [24], but these cannot be used to compare an HDR image with its tone-mapped reproduction [4]. Smith et al. [35] proposed the first metric intended for prediction in quality degradation in tone-mapped images due to local and global contrast distortion. However, the metric was only used in the context of controlling counter-shading algorithm and was not validated against experimental data. Aydin et al. [4] proposed a metric for comparison of HDR and tone-mapped images that is robust to contrast changes. The metric was later extended to video [5]. However, both of these metrics provide only qualitative results in the form of the distortion maps and they do not provide aggregate mean-opinion-score, which

**Table 1**
Comparative summary of the studies on tone-mapping evaluation. *Reference* column tells what kind of reference stimulus was used. *Criteria* are the attributes that were investigated. *Parameters* is the choice of the parameters for the operators. *Procedure* is the experimental procedure employed. *Comments* summarize the most relevant findings.

| Study | *Reference* | *Criteria* | *Parameters* | *Procedure* | *Comments* |
|---|---|---|---|---|---|
| Drago et al. [9] | None | Overall difference | Adjusted by the TMO authors | Similarity judgements | MDS analysis led to two abstract dimensions of the biggest differences: "naturalness" and "detail" |
| Kuang et al. [19] | None | Preference | Default | Paired comparisons | Results correlated for gray-scale and color |
| Ledda et al. [21] | HDR display | Fidelity | Default | Paired comparisons | Results correlated for gray-scale and color |
| Yoshida et al. [40] | Real scene | Fidelity (Naturalness), brightness, contrast, details in dark and bright regions | Preselected in a pilot study | Rating | Brightness, contrast and detail attributes are not correlated with naturalness |
| Delahunt et al. [8] | None | Preference | Default | Paired comparisons | The most preferred face lightness is between 46–49 CIE $L^*$ units |
| Ashikhmin and Goyal [3] | None and real scene | Preference and fidelity | Adjusted by the TMO authors | Ranking | Found difference between the results with and without reference |
| Yoshida et al. [41] | None and real scene | Preference and fidelity | Adjusted in the experiment | Method of adjustment | Brightness and contrast preference is subjective. The preference for white-point is relatively consistent |
| Akyuz et al. [1] | None | Preference | Not revealed | Ranking | Tested tone mapping operators not statistically different from the manually selected best exposure |
| Grave and | Bremond [13] | DLP projector very close from a screen | Visibility (Landolt C discrimination) | Default | Sampling psychophysical function |
| Results are | scene dependant. The visibility conservation requires a specific algorithm | | | | |
| Čadík et al. [7] | None and real scene | Fidelity, brightness, contrast, color reproduction, details reproduction, artifacts | Preselected in a pilot study | Rating (reference) and ranking (no reference) | The results of the experiment with and without reference are not statistically different. The same for ranking and rating. Contrast, color and artifacts most strongly correlated with fidelity |
| Akyüz and | Reinhard [2] | LDR reference profiles | Contrast perception | Default | 2-AFC (alternative forced choice) procedure |
| Cornsweet | illusion reproduction vary between the operators, contrast is not reproduced equally well in all regions | | | | |
| Kuang et al. [16] | Real scene (but not seen simultaneously) and HDR display | Fidelity | Default, except two Photoshop TMOs that were adjusted by an expert | Paired comparisons | Results are strongly correlated for the reference shown on an HDR display and as a real-scene |
| Villa and | Labayrade [37] | Real scene | Fidelity | Default | Rating and ranking |
| No | statistical difference between rating and ranking, inappropriate color and exposure are often cited as image artifacts | | | | |
| Ours | None | Preference and fidelity | 5 parameter variations tested in the main experiment | Rating and Ranking | See text |

could be used to quantitatively compare the results of tone-mapping. These metrics have not been validated against the subjective quality assessment results and their reliability cannot be confirmed. The automatic quality assessment for tone-mapping requires further research before it can be used for quantitative evaluation of tone-mapping operators.

## 3. Experiment

To measure an overall quality of video tone-mapping, with regard to *quality* and *fidelity with memorised reality* criteria, we performed a subjective quality assessment study. The paragraphs below describe the experimental procedure and discuss our choices.

### 3.1. Viewing conditions

The experiment took place in a dark room with controlled lighting directed towards a wall behind the screen, which produced ambient illumination of 50 lux while avoiding glare. All the video clips were displayed on a colorimetrically calibrated 24" 1920×1200 display (HP LP2480zx). The color calibration assumed that the source material is encoded in the sRGB color space. All video sequences were generated in full HD resolution so that no resampling was needed to play the clips in full screen mode. The clips were generated at 60 Hz and video playback was optimized to ensure that no tearing artifacts or dropped frames appear. Black areas were shown around the video if the clip aspect ratio did not match that of the screen.

The relevance of a real-world reference scene in tone-mapping evaluation studies is still a disputable topic. Ashikhmin and Goyal [3] demonstrated that given no reference, there is no difference between preference and fidelity criteria, but such a difference is measured when a reference is shown. Contrary to that, Kuang et al. [18] did not report a difference between preference and fidelity regardless of whether the reference was shown or not. Čadík et al. [7] also did not report a statistical difference between the results collected with and without a reference. We decided not to show a real-world reference because the experimental setup for video sequences would be very difficult and its impact on the results still disputable.

Some studies used an HDR display as a reference [21,13,16]. However, we excluded that option as the currently available HDR displays cannot produce the range of luminance that can be found in our video sequences and thus they do not faithfully reproduce the real-world colors. Although some studies report no statistical difference when an HDR display is used instead of a real-scene as a reference [16], this could be also explained by the small impact of any kind of reference in such studies.

### 3.2. Video clips

The content used in the experiment were selected to span a wide range of illumination conditions and to ensure that the clips are high quality and free from artefacts. We decided not to use the content from an HDR video camera available to us because of its poor quality and high level of noise, which were likely to bias quality assessment experiment. Instead, we used sequences of panning over good quality HDR panoramic images and computer generated clips from a driving simulator.

The driving simulator [29] was designed to produce accurate physical illumination units that correspond to the real-life driving experience. The light sources were modeled according to the light manufacturer's specification [30]. The texture reflectance maps were generated from photographs (taken at daylight). The accuracy of the rendering has been validated by comparing luminance values between computer-generated HDR images and HDR images photographed in the actual streets. The relative luminance errors were found to be less than 1 $\log_{10}$ unit (≈3.3 f-stops) for each measured spot (streetlight bulbs, car lights, lighted streets, background and sky). Such differences have a negligible effect on the visual sensitivity and adaptation models used in some of the tested tone-mapping operators. Despite the attention paid to proper simulation of illumination, the computer generated sequences could not be considered photo-realistic or equivalent to similar sequences, captured by a camera. However, such sequences represent a very common application scenario, where tone-mapping is used for simulators or video games. The selected video clips intentionally mixed both camera and computer generated content, to test the performance of the operators in both scenarios.

All the clips, containing both day-light and night-time scenes of varying dynamic range, are listed in Table 2. The timeline in Fig. 1 illustrates the temporal dynamic range and luminance variations of each HDR clips, using median, the 3rd and 97th percentiles of pixel luminance. The percentiles were used rather than the maximum and minimum values, as they are more robust to camera noise, clipping and rendering errors. The timeline plot shows that the clips contain a good variety of illumination changes, that differ both in amplitude and temporal frequency. In all the sequences, the absolute HDR pixel values were adjusted to approximate the absolute luminance levels in the corresponding real-world environments. One of the driving simulator clips (*Rivoli broken light*) contains strong flickering due to a broken street light, which we introduced intentionally to test operators' response to high frequency illumination variations. The clip *Ocean* contains gradual but high change of illumination because of panning a camera directly in front of the sun.

### 3.3. Tone mapping operators

In this study we considered only tone mapping operators that have been specifically designed to process video sequences and produced consistently good results. The initial selection included the operators by Durand and Dorsey [10], Pattanaik et al. [28], Irawan et al. [14] and Mantiuk et al. [23]. We did not consider one of the first video tone-mapping operators by Ferwerda et al. [12] because similar models and methods were incorporated in more recent and advanced operators of Pattanaik et al. [28] and Irawan et al. [14]. We also had to exclude two of the selected operators because of evident problems in the generated sequences. The operator of Pattanaik et al. [28] produced substantially desaturated and unnatural colors, and the operator of Durand and Dorsey [10]

**Table 2**
Video clips used in the experiments. The dynamic ranges were determined using the average of the ratios between the 97th and 3rd percentiles over all the frames of each clip.

| Clip name | Illumination range | Mean dynamic range (percentiles, $log_{10}(cd/m^2)$) | Source |
|---|---|---|---|
| Fountain | Daylight | 2.3 | Panorama |
| Ocean | Daylight | 2.0 | Panorama |
| Flower | Daylight | 1.4 | Panorama |
| Bridge Studios | Night | 2.5 | Panorama |
| Rivoli empty street | Night | 1.9 | Driving simulator |
| Rivoli broken light | Night | 2.1 | Driving simulator |
| Rivoli car passing | Night | 3.0 | Driving simulator |

resulted in high temporal fluctuation of brightness. We decided not to test these two operators because they were too likely to be universally rejected by all the observers. We also did not test the operator by Ledda et al. [22] because our own implementation produced excessive flickering artifacts. Finally, we received from the author an updated version of iCAM06 operator [17], which performed very well in other tone-mapping evaluation studies. Although the operator was intended for static images, the extensions ensured temporal coherence for video sequences.

In addition to the existing video tone-mapping operators, the study included an operator that simulates the response function of a consumer camera (Canon 500D) and its temporal exposure control. This is the most frequently used form of tone-mapping, yet it has received very little attention in tone-mapping studies. Including this operator is also motivated by an unexpected result in [1], which did not find a statistically significant difference between the results of state-of-the-art tone-mapping and the best exposure selected from an exposure sequence. Such a single exposure is reproduced by our *Camera* tone mapping operator.

The response function of a camera was measured for each channel separately using the method by Robertson et al. [34]. Temporal coherence was enforced by anchoring the exposure setting to the geometric mean of luminance filtered over time using an exponential filter with the time constant $\tau = 0.85$ s. Such a filter approximates the observed behavior of a typical camera.

The tone mapping operators used in the study are summarized in Table 3. The table also shows the parameters that were varied for each operator. Because testing more than several parameter combinations in an experiment would be impractical, for each operator we preselected five (from 49 generated) combinations that produce acceptable and visually different results. An example of a scene tone-mapped for a single operator with 5 different parameter settings is shown in Fig. 2.

### 3.4. Participants

10 paid participants, 8 males and 2 females, completed the experiment. They were asked to wear their corrective lenses and confirmed to have normal color vision. Most participants completed the experiment twice, with at least one day break between the experimental sessions.

The participants were divided into two groups (5 participants each) and each group was given a different task. The first group was asked to rate the overall image quality (preference) and the second group to rate how the generated results corresponded to their impression of viewing these scenes in real-world conditions (naturalness or fidelity with reality). Both tasks involved an 11-point rating scale. None of the participants was aware of the task of the other group.

### 3.5. Experimental procedure

The participants were asked to read a briefing form which introduced them to the experiment. For the preference criteria, the

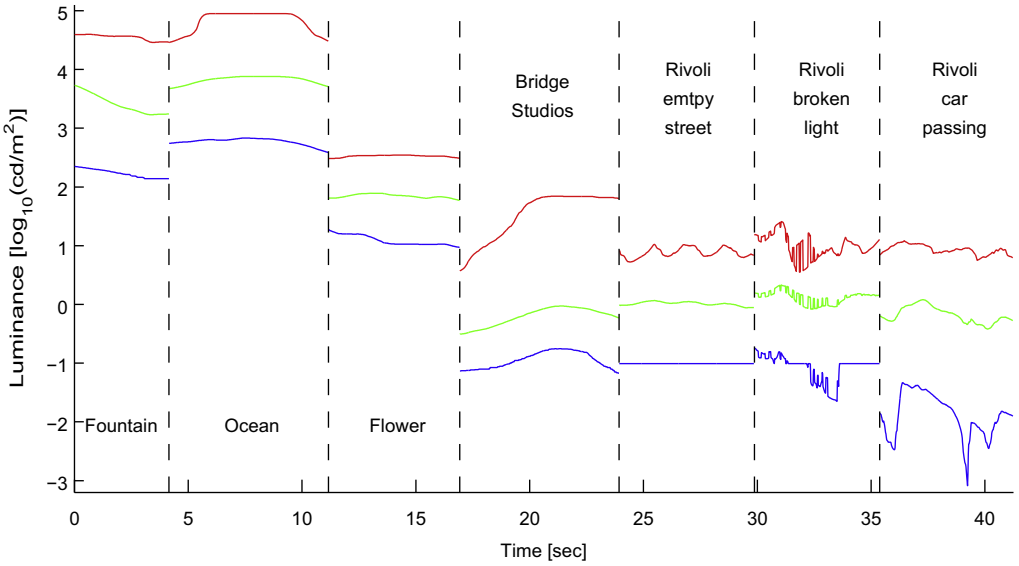


**Fig. 1.** Timeline of luminance variations of the median (in green), the 3rd (in blue) and 97th (in red) percentiles of each video clip. The 3rd and 97th percentiles are presented to illustrate the dynamic range variations of the HDR clips, and the median is plotted as it correlates with the overall brightness. The flat profile of the 3rd percentile in the *Rivoli empty street* and *Rivoli broken light* clips is due to the dark texture used for the night sky, which had the lowest luminance value for most frames. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 3**
Video tone mapping operators used in the experiment.

| Short name | Full name | Code | Goal | Parameters | Reference |
|---|---|---|---|---|---|
| iCAM06 | iCAM'06 | Provided by the author | Preserve color appearance | p – contrast; ct – adaptation time | [17] |
| DATMO | Display adaptive TMO | pfstools | Preserve contrast | e – contrast enhancement; c – color saturation | [23] |
| PTMO | Perceptual tone mapping for HDR streams | Own implementation | Preserver visibility | g – gamma; c – color saturation | [14] |
| Camera | Camera response curve | Own implementation | Best quality | x – exposure; c – color saturation | n/a |

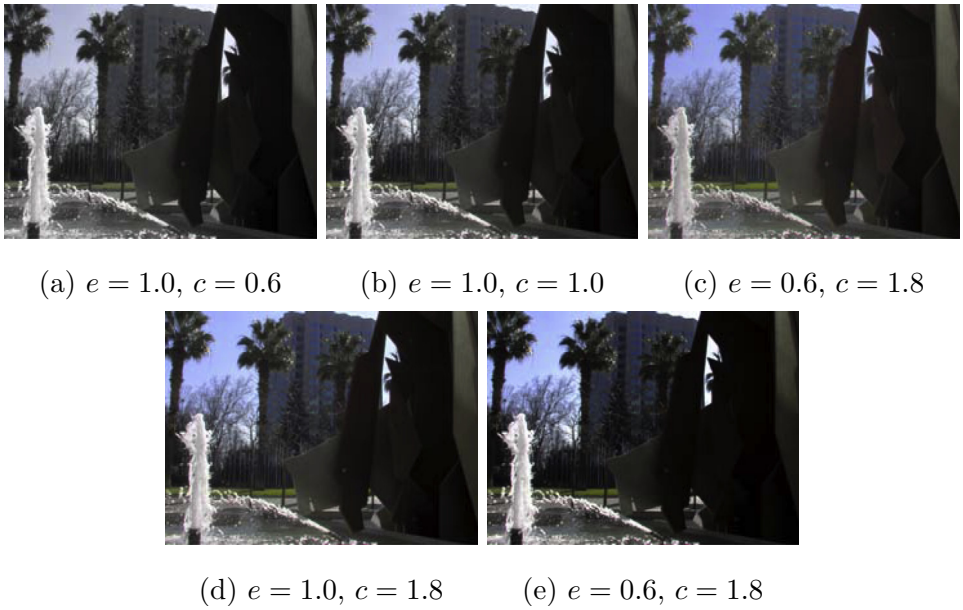


**Fig. 2.** Five different clips were produced for each operator using different parameter settings. The figure shows the parameters variation for the clip *Fountain* and the operator *DATMO*. The images are best seen when enlarged on a display.

detailed instruction was: "Rate the highest those video clips you prefer the most and you would select for your own video album". For the fidelity criteria the instruction was: "Rate how closely the displayed video clip matches the appearance of an actual scene you would expect to see in the real world". All the participants were asked to not look for a single feature, but to make each judgment based on their overall impression. To familiarize the participants with the experiment, a training session with 7 clips was used before the main experiment session.

For each video sequence 20 representative thumbnails (4 tone mapping operators $\times$ 5 parameter variations) were displayed in random order on the screen. A slider with an 11-point rating scale (from very poor to excellent) was shown above each thumbnail image. The video clips were played at full-screen resolution after clicking on a corresponding thumbnail image. To ensure that all the clips were viewed at least once, the rating slider for each clip was deactivated until the clip was played. The participants were allowed to see each video as many times as they wished and to change their previous ratings. Depending on the participant, the experiment took between 30 and 50 min to complete.

## 4. Results and discussion

### 4.1. Inter- and intra-observer correlation

It is desirable to know whether observers can perform the task reliably and whether the ratings from different participants agree with each other. To test that, we first compute the intraclass correlation coefficient (ICC) between the ratings of all observers: ICC $(2,1) = 0.874$ (95% confidence interval: 0.839–0.903). The relatively high value of the coefficient suggests that there is a substantial agreement between observers. The F-statistics confirms the statistically significant dependence ($F(139,1251) = 8.57, p < 0.001$). To check the reliability of the observers, we compute ICC $(2,1)$ between the repetitions of the experiment for each observer individually. The ICC values range from 0.38 to 0.89 with the mean 0.64. We decided not to exclude the observers with low ICC values because of the broad confidence intervals of these estimates. For only one observer the intra-observer correlation was significantly higher than the inter-observer correlation. This suggests that both the learning effect and individual bias were small for repeated measurements made on different days. For that reason we consider each repetition as an individual and independent sample in the following analysis.

### 4.2. Fidelity and quality criteria

Contrary to our expectations, there is no statistical significant difference between the group of observers that was asked to rate overall quality and the group that was asked to rate fidelity with the memorised reality. The three-way analysis of variance (7 scenes $\times$ 20 conditions $\times$ 2 criteria) did not indicate any statistical difference for the assessment criterion. The F-value was the largest for scene *Ocean* ($F(1,359) = 2.59, p = 0.11$) but the effect was still much above the 0.05 significance level. This is an unexpected result as the tested operators address very different criteria: operators PTMO and iCAM06 are meant to reproduce the appearance of real-world scenes, while operators DATMO and *Camera* are meant to minimize tone and contrast changes.

One possible explanation for no observable effect of criterion is that both fidelity and quality are strongly correlated. The reproduction that appears close to reality is also likely to be preferred. Another explanation is that observers are unable to determine the fidelity with reality given no real-world reference scene. The psychological evidence shows a very short time span (80 ms or less) when the visual system can retain much detail about the seen objects [33]. Exact color or tones of the real-world scenes are unlikely to be well remembered and we tend to rely on so-called *memory colors* [6], which are often significantly different from the real-world colors.

In addition, it could be argued that a perceptual match of the displayed content with reality is impossible, as scenes shown on a display will never appear as seen in real-world. This is because the displays are not capable of reproducing the exact light field of real-world scenes. Given that, the task of finding the highest fidelity operator is inherently difficult and prone to high measurement errors. Such errors could hide any potential difference between the fidelity and quality criteria.

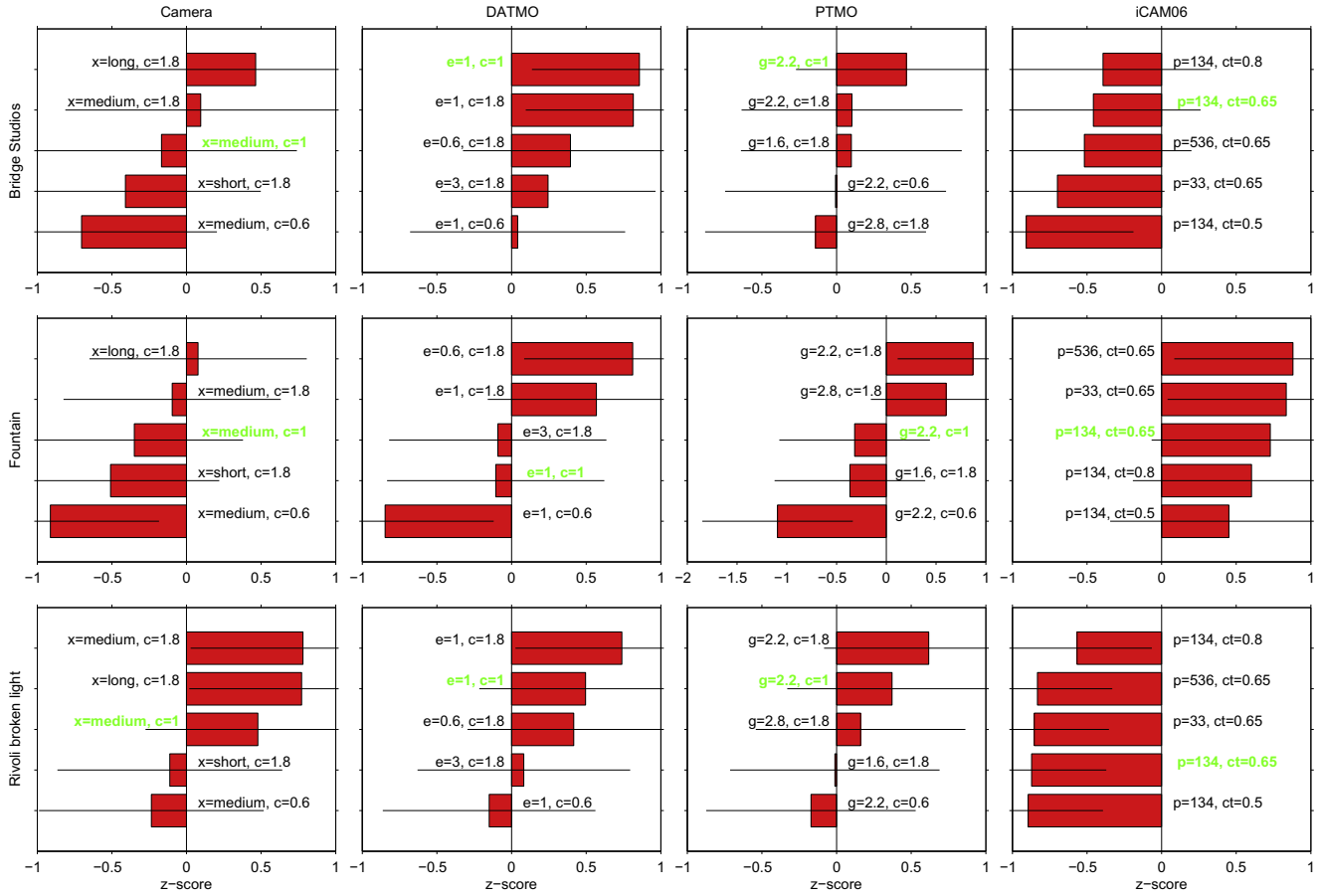


**Fig. 3.** The results for selected clips (rows), grouped individually for each operator (columns). The thin horizontal lines denote the statistical significance threshold at the significance level $\alpha = 0.05$. If the end of a wide bar overlaps with the thin line from another bar, there is no statistical significance for this pair of conditions. The significance thresholds are corrected for pair-wise comparisons. The bold-green labels indicate default parameter values.

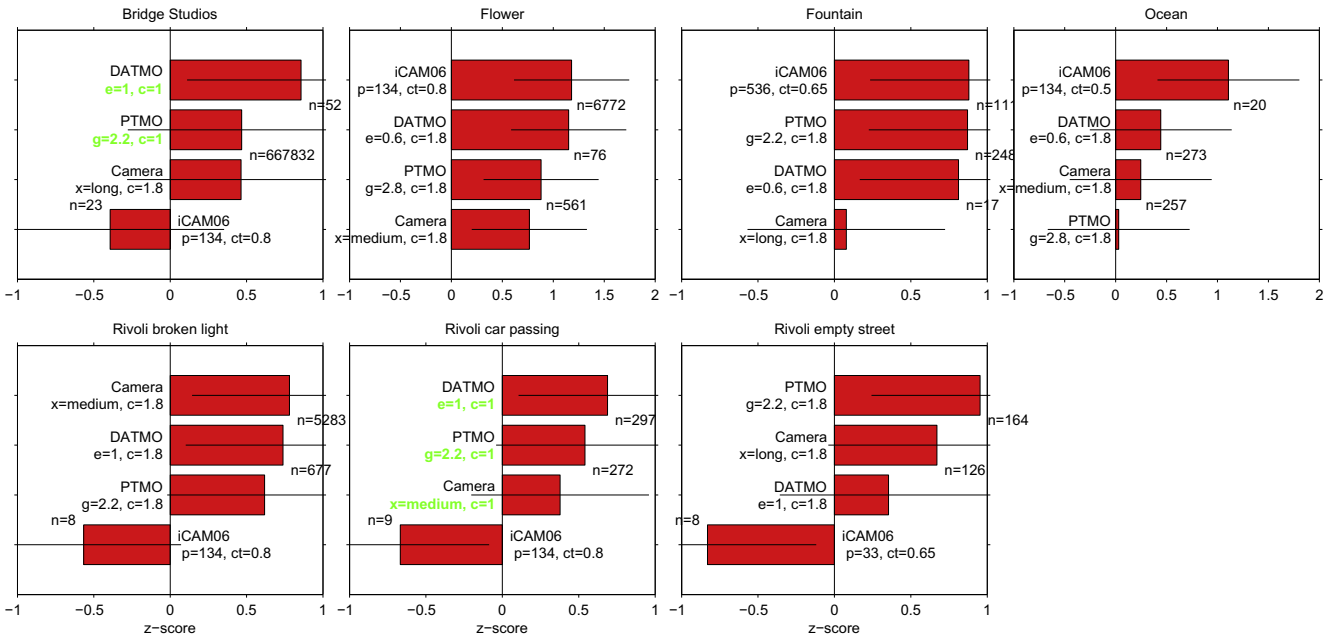


**Fig. 4.** The results for the best parameter settings for each operator. Refer to Fig. 3 for the description of notation.

### 4.3. Parameter selection

The default parameter settings, marked with bold green labels in Fig. 3, were in several cases not the ones that achieved the highest quality score. This suggest that fine-tuning of the parameter values is likely to change the tone-mapping ranking results.

Although most parameter variations did not lead to statistically significant differences (refer to thin black lines in Fig. 3), there were several settings that were universally rejected. In particular, most observers penalized low color-saturation ($c = 0.6$) and high contrast ($e = 3$ for DATMO and $g = 1.6$ for PTMO). The parameter variation may produce little measurable effect for some operators (iCAM06), but they may also render the results of other operators either unacceptable or exceptionally good.

The importance of parameter fine-tuning is reflected in other studies, which demonstrate strong subjective variations between parameter settings [41], or rank the manually adjusted operators as performing better than those that use default parameters [1,16].

### 4.4. Differences between operators

Fig. 4 shows the z-scores for the best performing parameter settings for each operator. The plots indicate that all four operators perform comparably for most scenes. In the case of a low dynamic range scene (*Flower*) there is no statistical difference between any pair of operators. In the case of the three driving simulator scenes (*Rivoli*), which are relatively similar to each other, we can observe different permutations of the three best performing operators. This further demonstrates the lack of significant differences between them.

The result of *iCAM06* was selected as the best reproduction of the *Ocean* scene. The differences was significant for all operators except the *DATMO*, but even that difference is likely to be significant for a larger sample size. However, the results of the same operator were unacceptable for all night scenes: *Bridge Studios* and all *Rivoli* clips. We discuss a possible reason for that in the next section.

Although *DATMO* and *PTMO* aim at very different goals, which are preserving contrast or visibility, we did not find a statistical difference between them for any of the scenes. This is reflected in our previous finding, which did not find any difference between quality and fidelity criteria. It is likely that the differences between different goals of tone-mapping are not measurable in rating experiments, especially when no reference is available.

### 4.5. Performance of the Camera operator

Interestingly, we did not find a statistically significant difference between the *Camera* operator and the best performing operator for all except the two scenes: *Fountain* and *Ocean*. For both these scenes there was another operator that performed significantly better than the *Camera* operator. To further investigate where this operator failed, we plot in Fig. 5 its results compared with the best performing operators. The results are shown for the two clips the operator struggled with (*Fountain* and *Ocean*), and one clip for which it performed comparably with the other operators (*Bridge studios*). The plots below the results show that the *Camera* operator struggles especially with the scenes that contain both a large dynamic range and a multi-modal distribution of luminance. The sigmoidal tone curve found the *Camera* operator is designed to preserve contrast in the central part of a tone-curve. If a scene contains several disjoint segments of luminance that need to be well preserved (several peaks in the histogram), a more flexible tone curve is likely to do a better job, as can be seen for the scene *Ocean* in Fig. 5. In the case of the scene *Fountain*, the stretched tone-curve that gave good visibility of dark and bright regions resulted in the most preferred reproduction. Since the shape of the camera tone curve is fixed, it cannot be adapted to scenes of a wider dynamic range.

An interesting case in shown for the scene *Bridge studios* in Fig. 5, where the most preferred tone curves do not have their slopes aligned with the peak of the histogram. If the image was processed by histogram equalization, the steepest slope of the tone curve would be mapped to the peak in the histogram. However, the peak is due to the background, which does not contain much useful information but occupies a large portion of the frame. The DATMO operator, which was the most preferred choice for this scene, has a mechanism that puts more emphasis on the luminance range that contains large contrast variation and thus can produce a tone-curve that is aligned with the content rather than image histogram. The *Camera* operator performed well in this case because the dynamic range of the image portion that contains relevant information is relatively small. It is important to note that the exposure for this operator was effectively manually adjusted in the experiment as several exposure variations were included in the data set. More advanced operators can find the best exposure automatically.

We can conclude that the fixed S-curves found in most cameras are not suitable for the scenes with important image parts that differ significantly in their luminance. Therefore, even though S-curves are commonly used for tone-mapping and color appearance modeling [31,17,26], they may not be the best choice for the scenes that contain very different luminance levels. This characteristic of the scene cannot be captured well by the standard measures of the dynamic range, such as the percentile ratio listed in Table 2. In fact the camera operator performed well for the scenes of seemingly high dynamic range, but whose relevant content occupied the limited portion of that range. Therefore, a successful tone operator needs to adaptively choose the shape of the tone curve by analyzing scene content.

### 4.6. Power analysis

Little statistical difference between operators could be surprising given other studies on static operators, which reported more significant differences. To test whether the sample size is sufficient for our experiment, we perform power analysis assuming that the measured variance is equal to the true z-score variance and the t-test is performed to detect the differences. The $n$ values reported between the bars in Fig. 4 indicate how many samples need to be collected to make the difference between a pair of conditions statistically significant, assuming a statistical power of 0.8.[1] The required sample size is very close to the number of samples we collected ($n \leqslant 19$) for the clips that were easily distinguishable. Only in two cases would doubling the number of samples allow detection of a potential difference. For the other clips, the large numbers on the plots suggest that significant differences between the corresponding pairs of conditions are unlikely to be detected in this experiment regardless of the number of observers. Moreover, a larger sample size could reduce confidence intervals but is unlikely to increase the differences in quality between operators and thus make the measured effect more relevant. For example, if 51 out of 100 observers choose one operator over another, the difference between the operators is most likely negligible even though it is possible to prove that it is statistically significant.

There are two reasons that can explain the small difference in quality between the tested operators. Firstly, we preselected only the properly performing operators, excluding those that would

[1] The statistical power is the probability of finding a difference when the difference exists.

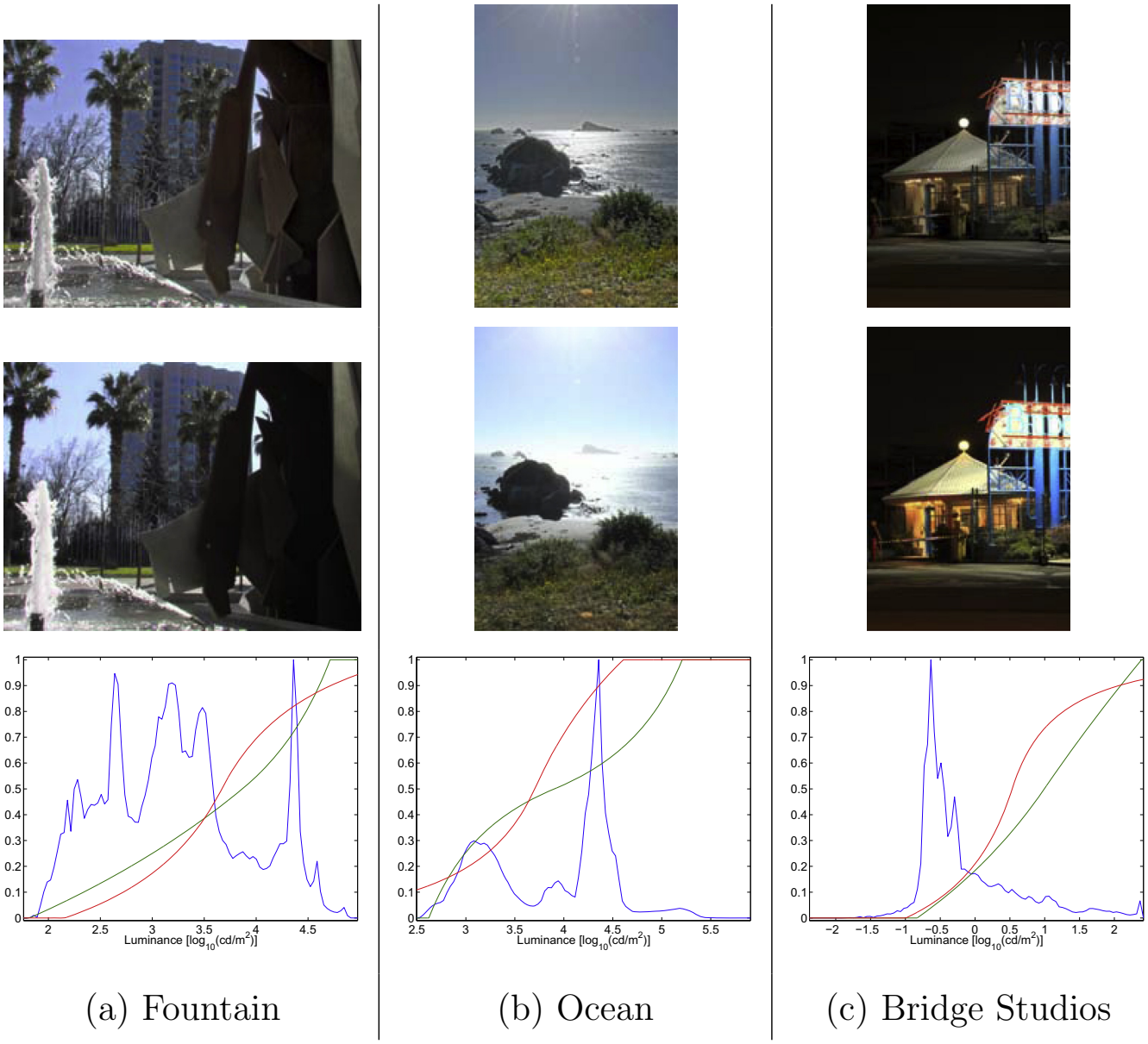


**Fig. 5.** Comparing the best rated tone mapped images (1st row) with the images from the *Camera* operator with the best rated parameters tuning (2nd row). The 3rd row contains the histogram of the original HDR image (in blue) and the tone curves of respectively the best operator (in green) and the *Camera* operator (in red). Best rated operators are iCAM06 for (a) and (b) and DATMO for (c). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

clearly fail to deliver high quality results. Secondly, we considered a wide range of parameters, giving a "fair chance" for each operator to perform well. If only the default of poorly selected parameters were used, many operators would perform significantly worse. The finding that state-of-the-art tone-mapping operators are difficult to differentiate in a quality experiment was also reported by Čadík et al. [7] and Kuang et al. [16], who did not find a statistical difference between the best performing operators.

This finding is important for two reasons: first there could be little room for appreciable improvement in the quality of the results produced by tone-mapping operators. Second, tone-mapping evaluation studies must explore the range of tone-mapping parameters, but also carefully select images, so that advantages of more sophisticated algorithms can be demonstrated in specific cases.

## 5. Quality and tone-mapped image statistics

In this section we analyze the relationship between the measured quality and image statistics. Our goal is to explain why some operators at a particular parameters setting perform better than others.

The four statistics that were selected for the analysis are the median CIE L* value, standard deviation of the same L*, averaged standard deviation for all $8 \times 8$ blocks, and mean CIE L*a*b* chroma, all computed for all pixels of tone-mapped frames. They were selected as a correlate of an overall image brightness (lightness), contrast, sharpness and colorfulness respectively. As shown in previous tone-mapping evaluation studies [40,7], image attributes such as brightness, contrast and sharpness (detail) influence the preference and fidelity of tone-mapped images. If the correlates

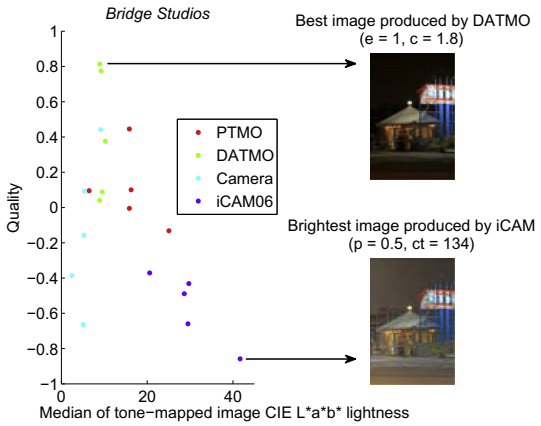


**Fig. 6.** Relation of brightness and quality for the night scene *Bridge Studios*.

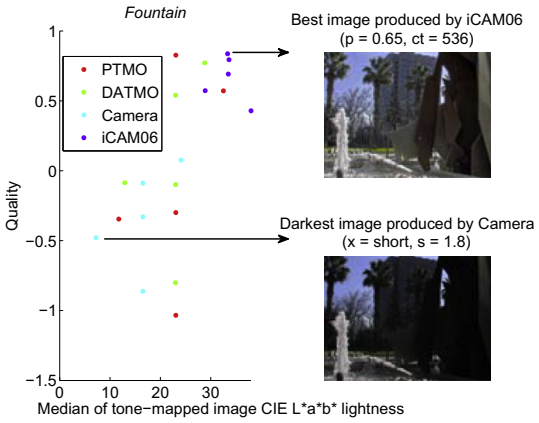


**Fig. 7.** Relation of brightness and quality for the day-light scene *Fountain*.

capture these attributes, they could be used to predict the quality of tone-mapped images.

Because of the large amount of possible relations, we found it necessary to create a visualization tool, which is included in the supplementary materials.[2] It allows us to explore all possible correlations and quickly navigate to relevant video frames. In the following paragraphs we summarize the most interesting findings after analyzing a large number of such relations.

### 5.1. Brightness

Brightness was reported as a decisive factor for quality assessment [3,41]. Using median CIE L* value as the correlate of brightness, we plot its relation to quality for a night scene in Fig. 6 and for the day-light scene in Fig. 7. The first observation is that the best rated night scenes exhibit much lower median L* values than the best-rated day-light scenes. Although it is not surprising that night-scenes should be relatively darker than day-light scenes, this basic rule was broken by iCAM06 operator, resulting in its very poor quality ratings for all night scenes. Therefore, adapting the tone-curve to the brightness of a particular scene is essential for all operators that must be robust to a range of scene content.

### 5.2. Chroma

Overall image colorfulness has a large impact on the quality ratings, with a strong preference for colorful images in case of day-light scenes, plotted in Figs. 8 and 9, and no such obvious correlation for night scenes (refer to the supplementary materials). We did not observe a typical inverted-U-shape relation between chroma and quality, where the quality decreases when chroma is excessively high. This could be due to the limited range of chroma values we used in this study. The results suggest that boosting image colorfulness improves quality, though this observation may apply only to images that do not contain faces and skin-colors.

Several studies reported no statistical difference between gray-scale and color tone-mapping results [19,21]. The correlation of the mean chroma value and quality, shown in Figs. 8 and 9, demonstrates the opposite. We suspect that the lack of such an effect reported in previous studies is the result of a minimal variation of tone-mapping parameters in those experiments; the generated images possibly did not vary much in overall chroma, making it difficult to differentiate between the quality scores of gray-scale and color images.

[2] Supplementary materials can be found at: http://pages.bangor.ac.uk/eesa0c/projects/video_tmo_cmp/

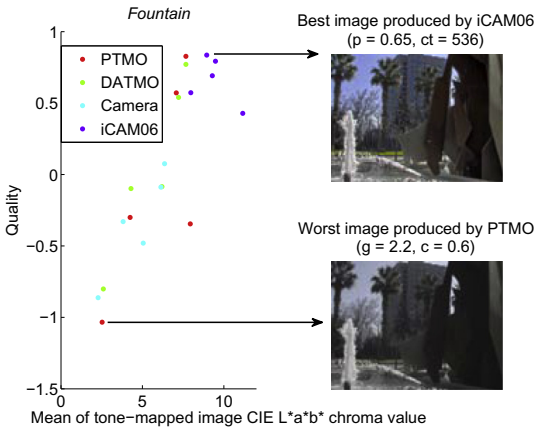


**Fig. 8.** Relation of chroma and quality for the day-light scene *Fountain*.

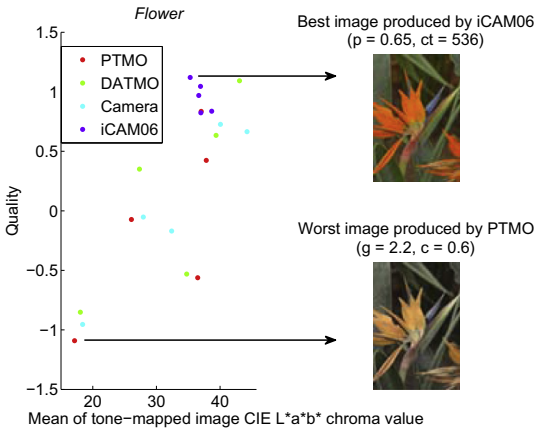


**Fig. 9.** Relation of chroma and quality for the day-light scene *Flower*.

### 5.3. Other correlates

No significant correlation was found between overall image contrast (standard deviation of L*), sharpness (standard deviation in 8 × 8 blocks) and quality. We also explored correlation between quality and tone-mapping parameters, which were unified across all operators using the generic tone mapping operator [25] (refer to the supplementary materials for the details). The correlations between these correlates and quality were much weaker than for chroma and lightness. These correlates, however, were useful to

better understand the differences between results produced by the operators. *Camera* operator on average produced the results of the highest overall contrast, but these results were not always the most preferred. *iCAM06* produced consistently the most preferred chroma for all day-light scenes, but it also made images too bright for night scenes. We encourage the reader to explore those differences in the full report included in the supplementary materials.

## 6. Conclusions

An experiment assessing the quality and fidelity of tone mapping operators has been conducted for HDR video sequences. The results did not indicate a statistically significant difference between the three best performing operators for most scenes. The only exception was the scene *Ocean*, for which iCAM06 performed significantly better. We attribute this result to a better color reproduction. The camera tone-curve did not perform significantly worse than the other operators for the scenes of moderate dynamic range, which confirms the findings of previous studies [1]. But we also found that the fixed S-shaped tone curves found in cameras cannot cope with the scenes where important content is spanned across a very wide dynamic range. This suggests that the cameras capable of capturing an extended dynamic range would benefit from more sophisticated tone-mapping.

We did not find evidence that there is a preference difference between tone-mapping algorithms that simulate the limitations of the human vision to achieve the best fidelity with reality, and those whose main goal is to reduce the dynamic range and produce the most pleasing images. We suspect that both tasks result in a similar expectation for the resulting video, especially if the observer is not able to see a visual reference with the original scene.

The analysis of tone-mapped image statistics revealed that the best operators enhanced colorfulness for day-light scenes. There was also a strong indication that day-light scenes need to be overall much brighter than night scenes. The iCAM06 operator, which produced much brighter images for night scenes, was rated significantly lower for those scenes. Therefore, it is important for operators to adapt to the scene content.

The last recurring observation is the importance of exploring the variation of tone-mapping parameter settings in tone-mapping evaluation studies. The lack of such variation can easily render the results of some operators unacceptable. It also shows that automatic parameter adjustment does not always produce the best results.

## References


[1] A.O. Akyüz, R. Fleming, B.E. Riecke, E. Reinhard, H.H. Bulthoff, Do HDR displays support LDR content? A psychophysical evaluation, ACM Transactions on Graphics 26 (3) (2006) (article no. 38).

[2] A.O. Akyüz, E. Reinhard, Perceptual evaluation of tone reproduction operators using the Cornsweet–Craik–O'Brien illusion, ACM Transactions on Applied Perception 4 (4) (2008) 1–29.

[3] M. Ashikhmin, J. Goyal, A reality check for tone mapping operators, ACM Transactions on Applied Perception 3 (4) (2006) 399–411.

[4] T.O. Aydin, R. Mantiuk, K. Myszkowski, H.P. Seidel, Dynamic range independent image quality assessment, ACM Transactions on Graphics 27 (3) (2008) 69.

[5] T.O. Aydin, M. Čadík, K. Myszkowski, H.P. Seidel, Video quality assessment for computer graphics applications, ACM Transactions on Graphics 29 (6) (2010) 1.

[6] P. Bodrogi, Investigation of Colour Memory, Ph.D. Thesis, University of Pannonia, 2007.

[7] M. Čadík, M. Wimmer, L. Neumann, A. Artusi, Evaluation of HDR tone mapping methods using essential perceptual attributes, Computers and Graphics 32 (3) (2008) 330–349.

[8] P.B. Delahunt, X. Zhang, D.H. Brainard, Perceptual image quality: effects of tone characteristics, Journal of Electronic Imaging 14 (2) (2005) 1–12.

[9] F. Drago, K. Myszkowski, T. Annen, N. Chiba, Adaptive logarithmic mapping for displaying high contrast scenes, in: Proc. of Eurographics, Granada, Spain, 2003, pp. 419–426.

[10] F. Durand, J. Dorsey, Interactive tone mapping, in: Proc. of Eurographics Symposium on Rendering, Brno, Czech Rep., 2000, pp. 219–230.

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

[12] J.A. Ferwerda, S.N. Pattanaik, S. Peter, D.P. Greenberg, A model of visual adaptation for realistic image synthesis, in: Proc. of SIGGRAPH, ACM, New Orleans, Lousiana, 1996, pp. 249–258.

[13] J. Grave, R. Brémond 5 (2) (2008) (article no. 12).

[14] P. Irawan, J.A. Ferwerda, S.R. Marschner, Perceptually based tone mapping of high dynamic range image streams, in: Proc. of Eurographics Symposium on Rendering, Konstanz, Germany, 2005, pp. 231–242.

[15] G.M. Johnson, Cares and concerns of CIE TC8-08: spatial appearance modeling and HDR rendering, in: Proc. of SPIE Image Quality and System Performance II, vol. 5668, 2005, pp. 148–156.

[16] J. Kuang, R. Heckaman, M.D. Fairchild, Evaluation of HDR tone-mapping algorithms using a high-dynamic-range display to emulate real scenes, Journal of the Society for Information Display 18 (7) (2010) 461–468.

[17] J. Kuang, G.M. Johnson, M.D. Fairchild, iCAM06: a refined image appearance model for HDR image rendering, Journal of Visual Communication and Image Representation 18 (5) (2007) 406–414.

[18] J. Kuang, C. Liu, G. Johnson, M. Fairchild, Evaluation of HDR image rendering algorithms using real-world scenes, in: Proc. International Congress of Imaging Science, Milwaukee, Wisconsin, USA, 2006, pp. 461–468.

[19] J. Kuang, H. Yamaguchi, G.M. Johnson, M.D. Fairchild, Testing HDR image rendering algorithms, in: Proc. IS&T/SID 12th Color Imaging Conference, Scotsdale, Arizona, 2004, pp. 315–320.

[20] J. Kuang, H. Yamaguchi, C. Liu, G.M. Johnson, M.D. Fairchild, Evaluating HDR rendering algorithms, ACM Transactions on Applied Perception 3 (3) (2007) 286–308.

[21] P. Ledda, A. Chalmers, T. Troscianko, H. Seetzen, Evaluation of tone mapping operators using a high dynamic range display, ACM Transactions on Graphics 24 (3) (2005) 640–648.

[22] P. Ledda, L.P. Santos, A. Chalmers, A local model of eye adaptation for high dynamic range images, in: Proc. AFRIGRAPH, ACM, Stellenbosch, South Africa, 2004, pp. 151–160.

[23] R. Mantiuk, S. Daly, L. Kerofsky, Display adaptive tone mapping, ACM Transactions on Graphics 27 (3) (2008) (article no. 68).

[24] R. Mantiuk, K.J. Kim, A.G. Rempel, W. Heidrich, HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions, ACM Transactions on Graphics 30 (4) (2011) 1.

[25] R. Mantiuk, H.P. Seidel, Modeling a generic tone-mapping operator, Computer Graphics Forum 27 (2) (2008) 699–708.

[26] N. Moroney, M.D. Fairchild, R.W.G. Hunt, C.J. Li, M.R. Luo, T. Newman, The CIECAM02 color appearance model, in: Proc. of IS&T-SID 10th Color Imaging Conference, 2002, pp. 23–27.

[27] K. Myszkowski, R. Mantiuk, G. Krawczyk, High Dynamic Range Video. Synthesis Digital Library of Engineering and Computer Science, Morgan & Claypool Publishers, San Rafael, USA, 2008.

[28] S.N. Pattanaik, J. Tumblin, H. Yee, D.P. Greenberg, Time-dependent visual adaptation for fast realistic image display, in: Proc. of SIGGRAPH, ACM, New Orleans, Lousiana, 2000, pp. 47–54.

[29] J. Petit, R. Brémond, A high dynamic range rendering pipeline for interactive applications: in search for perceptual realism, The Visual Computer 28 (6–8) (2010) 533–542 (Special issue CGI 2010).

[30] Philips, Philirama 1998/99, Philips, 1999.

[31] E. Reinhard, T. Pouli, T. Kunkel, B. Long, A. Ballestad, G. Damberg, Calibrated image appearance reproduction, ACM Transactions on Graphics 31 (6) (2012).

[32] E. Reinhard, G. Ward, S. Pattanaik, P. Debevec, W. Heidrich, K. Myszkowski, High Dynamic Range Imaging: Acquisition, Display, and Image-based Lighting, second ed., The Morgan Kaufmann Series in Computer Graphics, Elsevier (Morgan Kaufmann), Burlington, MA, 2010.

[33] R. Rensink, Seeing, sensing, and scrutinizing, Vision Research 40 (10–12) (2000) 1469–1487.

[34] M.A. Robertson, S. Borman, R.L. Stevenson, Dynamic range improvement through multiple exposures, in: Proc. IEEE International Conference on Image Processing. Brussels, Belgium, 1999, pp. 159–163.

[35] K. Smith, G. Krawczyk, K. Myszkowski, Beyond tone mapping: Enhanced depiction of tone mapped HDR images, Computer Graphics Forum 25 (3) (2006) 427–438.

[36] J.H. Van Hateren, Encoding of high dynamic range video with a model of human cones, ACM Transactions on Graphics 25 (4) (2006) 1380–1399.

[37] C. Villa, R. Labayrade, Psychovisual assessment of tone-mapping operators for global appearance and colour reproduction, in: Proc. of Colour in Graphics Imaging and Vision 2010, Joensuu, Finland, 2010, pp. 189–196.

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

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

[40] A. Yoshida, V. Blanz, K. Myszkowski, H.P. Seidel, Perceptual evaluation of tone mapping operators with real world scenes, in: Proc. of SPIE Human Vision and Electronic Imaging X, San Jose, CA, vol. 5666, 2005, pp. 192–203.

[41] A. Yoshida, R. Mantiuk, K. Myszkowski, H.P. Seidel, Analysis of reproducing real-world appearance on displays of varying dynamic range, Computer Graphics Forum 25 (3) (2006) 415–426.