Which tone-mapping operator is the best? A comparative study of perceptual quality

XIM CERDA-COMPANY[1,2,*], C. ALEJANDRO PARRAGA[1,2], AND XAVIER OTAZU[1,2]

[1] Computer Vision Center, Edifici 'O', Campus UAB (Bellaterra) Barcelona, Spain
[2] Computer Science Department, Universitat Autonoma de Barcelona, Spain
[*] Corresponding author: ximcer@cvc.uab.es

Compiled February 8, 2018

Tone-mapping operators (TMO) are designed to generate perceptually similar low-dynamic range images from high-dynamic range ones. We studied the performance of fifteen TMOs in two psychophysical experiments where observers compared the digitally-generated tone-mapped images to their corresponding physical scenes. All experiments were performed in a controlled environment and the setups were designed to emphasize different image properties: in the first experiment we evaluated the local relationships among intensity-levels, and in the second one we evaluated global visual appearance among physical scenes and tone-mapped images, which were presented side by side. We ranked the TMOs according to how well they reproduced the results obtained in the physical scene. Our results show that ranking position clearly depends on the adopted evaluation criteria, which implies that, in general, these tone-mapping algorithms consider either local or global image attributes but rarely both. Regarding the question of which TMO is the best, *KimKautz* [1] and *Krawczyk* [2] obtained the better results across the different experiments. We conclude that a more thorough and standardized evaluation criteria is needed to study all the characteristics of TMOs, as there is ample room for improvement in future developments.
© 2018 Optical Society of America

OCIS codes: (100.2000) Digital image processing; (100.3010) Image reconstruction techniques; (330.5510) Psychophysics.

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

In almost all naturalistic viewing situations, we are immersed in scenes that could be described as High Dynamic Range (HDR), in other words, the intensity difference between the brightest and the darkest patch is much higher than the difference both imaging and capturing devices can faithfully capture. For instance, the energy ratio between sunlight and starlight is approximately about 100,000,000:1 [3]. If the Human Visual System (HVS) was to linearly represent these extreme differences in its normal daylight operation, it would require a much larger sensitivity range for its retinal sensors (cones) and neural pathways than is achievable within biological limitations. Instead, millions of years of evolution have solved this problem by adapting the sensorial and neural machinery, allowing it to non-linearly convert the large natural intensity range into a much smaller range of about 10,000:1 [4, 5].

A. Historical Context

The problem of translating the HDR world into Low Dynamic Range (LDR) depictions is very old. Renaissance painters such as da Vinci and Caravaggio tried to solve it by creating an artistic technique called *Chiaroscuro*, which pays attention to strong contrasts in different painted areas, creating very strong effects. This and the need to overcome the limitations of physical materials (oil paints and substrates) inspired later artists to produce remarkable paintings. Perhaps the most dramatic were created by depicting a single artificial source of light (such as a candle), making the details of the central subject very bright, while other subjects are slightly darker. It can be argued that some of the works by Rembrandt and Constable are no different from today's HDR photography [6–8].

The arrival of photography implied a new set of challenges given the strong limitations of early light-sensitive material [9]. Examples of the first characterization of silver halide films as plots of density vs exposure was made by Hurter and Driffield in 1890 [8]. In particular, outdoor scenes were very difficult to capture and early photographers experimented with multiple exposures to overcome dynamic range problems. When photographs involved human subjects, these had to remain still during the whole process so that several exposures could be combined into a single image.

B. Electronic HDR imaging

Analog HDR imaging allowed only limited manipulations (via exposure time, chemical reactions or the combination of several exposures, etc.) but the arrival of electronic digital imaging made possible for long-range interactions of pixels located in different parts of the image. This opened the field to multiple possibilities including mimicking the operation of the Human Visual System (HVS) and the work of chiaroscuro artists.

Physiological and psychophysical research has shown that photopic human vision is the result of highly nonlinear processing of the information captured by retinal cones. This processing includes the inhibition of the output of a neuron by the output of surrounding neurons in its field of view [10], which results in higher sensitivity for edges and spots than for uniform light. Other processing includes the combination of visual information in the retina into a series of post-receptoral chromatically-opponent channels to transmit it to the visual cortex via the optic nerve [11]. In the cortex, visual information is mostly processed in terms of its spatial frequency and visual orientation [12]. In the 1960's a series of psychophysical experiments with achromatic Mondrians run by Edwin Land demonstrated that patches reflecting light with exactly the same physical properties appear completely different to observers [13]. This implies that a digital image (where these patches produce exactly the same pixel values) cannot be modified using a pixel-wise transformation to simulate the appearance reported by observers. In other words, the information contained in individual pixels is not enough to mimic human vision. A comprehensive review of these experiments can be found in [14–17]. Other effects to consider are related to how the visual cortex processes local brightness interactions [18, 19]. More detailed experiments have shown the effects of edges in illumination perception by matching the appearance of painted wooden facets to that of a painted test target ("ground truth") [20], a paradigm very similar to ours (see below).

In order to mimic the response of the HVS, electronic imaging systems set out to use information not only from single pixels but from the entire scene. This allowed them a much larger flexibility to calculate appearances and to apply them to electronic displays or prints. Ironically, later HDR algorithms reverted to the old "multiple exposures" and "pixel-wise" processing techniques of analog photography for the same task (see below).

C. Tone-Mapping Operator (TMO)

Mapping the HDR dynamic range of the world into LDR media presents an important challenge for visual representation technologies mainly because most imaging devices (cameras and monitors) are only able to obtain/display images within a small range of about 100:1 [4] which can be increased up to 1000:1 for specialized HDR led-based displays [21]. To solve this problem, an assortment of non-linear image processing techniques were defined to display HDR scenes in LDR devices. To construct the HDR image, many LDR images of the same scene are usually taken at different exposure values, capturing a much larger dynamic range. This HDR image is generated by extracting from each LDR image the information corresponding to its region of interest (where it is neither over- or under-exposed) and combining them. Since this new HDR image cannot be displayed on a standard LDR monitor, an algorithm is needed to reduce its dynamic range to match that of the monitor. A common solution is to use a Tone-Mapping Operator (TMO) to reduce the dynamic range while keeping the perceptual characteristics of the original HDR image approximately constant. The performance of these TMOs depends of several factors including lighting and viewing conditions, aesthetic/realistic preferences, local/global assumptions, etc. and are usually evaluated using computational ([22, 23]) and psychophysical ([24–34]) methods.

Although HDR images are able to reproduce a wider range of luminance highlights and a shadows than LDR ones, the presence of veiling glare both in the camera and the eye limits the possible range of accurate luminance measurements [8]. Since HDR are perceptually closer to the original scene, there must be other reasons than simply obtaining a larger range of luminances for this perceived improvement. It has been hypothesized [8] that the improvement comes from a better preservation of relative spatial information that comes from digital quantization (spatial differences in highlights and shadows are preserved) and TMOs use this to replicate the HVS processing.

In this work we present a new set of experiments and analysis to psychophysically evaluate the performance of 15 state-of-the-art TMOs. This allowed us to rank the TMOs according to how well they represent the original scene as human observers perceive it. Unlike previous studies, all the experiments were performed in a controlled environment and tone-mapped images were presented side by side with the physical scene.

D. "Global" vs "Local" analysis

At this point we believe it is important to clarify the terminology used throughout this work. The term "Tone" is traditionally used to describe pixel data (as in "Tone Mapping") and was introduced by Mees [9] in 1920 to explain how exposure was related to photographic print density (silver halide response). Indeed *"Tone Scale"* is the name given to a look-up table that transforms data in an input space to a desired output space.

The term *"Global TMO"*, which is also used by several authors [35–37] generally refers to an algorithm that applies the same pixel-wise adjustment to all pixels in the image (although, in fact, it uses the most local information: a single pixel). In contrast, the term *"Local TMO"* generally defines an algorithm that applies a combination of pixel-wise processing and spatial transformations to improve the image. Although confusing we will follow the traditional terminology here, calling Global TMOs to algorithms that apply pixel-wise processing and Local TMOs to algorithms that apply a combination of pixel-wise and spatial image processing.

We will refer to our psychophysical experiments (see below) as "*Scene Reproduction*" when observers judge images by freely comparing them, and "*Segment Matching*" when they match the luminances of specific points in the scene to those of a reference table in the same scene.

2. STATE-OF-THE-ART

A. Previous TMO Psychophysics

Although the idea of using algorithms to match the brightness of real scenes to that of imaging devices is not new [35, 38], TMOs did not become popular until the turn of the century, when affordable digital cameras became available [1, 2, 32, 37, 39–49]. To date, many different psychophysical experiments have been performed and they can be classified as follows:

A.1. Experiments without a reference HDR scene

One of the first psychophysical experiments to evaluate TMOs compared the performance of 6 TMOs on 4 different (synthetic and photographic) scenes by asking subjects to make pairwise

perceptual evaluations and by rating stimuli with respect to three attributes: apparent image contrast, apparent level of detail, and apparent naturalness [24]. The results showed that preferred operators produced detailed images with moderate contrast.

Kuang et al. [25] performed pairwise comparisons on 8 different tone-mapping operators using 10 different scenes and two conditions (color and grey-level) where subjects had to choose the preferred image considering general rendering performance (including tone compression performance, color saturation, natural appearance, image contrast and image sharpness). Their results showed that the grey-scale tone-mapping performances are consistent with those in the overall rendering results, if not the same.

A.2. Experiments with a reference HDR scene

Yoshida et al. [26] conducted a psychophysical experiment based on a direct comparison between the appearance of real-world scenes and TMO images of these scenes displayed on a LDR monitor. In their experiment, they differentiate between global and local operators, and introduced, for the first time, the comparison between tone-mapped image and real scene, selecting two different indoor architectural scenes. Fourteen subjects were asked to give ratings according to several criteria like realism (image naturalness in terms of reproducing the overall appearance of the real world views) and image appearance (brightness, contrast, detail reproduction in dark regions, and in bright ones). They found that none of these image appearance attributes had a strong influence on the perception of naturalness by itself. This work was extended to find out which attributes of image appearance accounted for the differences between tone-mapped images and the real scene [30]. They observed a clear distinction between global and local operators. However, they concluded again that none of the evaluated image attributes had a strong influence on the perception of naturalness by itself which suggested that naturalness depends on a combination of the other attributes with different weights.

In an another work, Ashikhmin and Goyal [28] performed three different experiments. Subjects ranked different tone-mapped images depending on the task. In the first experiment, the authors asked which image they liked more without having the reference scene. In the second one, the authors asked which image seemed more real without viewing the reference scene and, in the third one, they asked which image was the closest to the real scene viewing the reference scene. They observed that rankings were totally different when subjects could compare the tone-mapped image to the reference scene.

In a subsequent study, Kuang et al. [31] performed three different experiments they named *preference evaluation*, *image-preference modelling* and *accuracy evaluation*. In the *preference evaluation* experiment, pairwise comparisons between tone-mapped images were performed. Here they used only color images and the aim was to evaluate the general rendering performance by instructing observers to consider perceptual attributes such as overall impression on image contrast, colorfulness, image sharpness, and natural appearance. In contrast, in the *image-preference modelling* experiment, they rated grey-scale images (which were grey-scale versions of the first experiment color images). Here, observers considered perceptual attributes such as highlight details, shadow details, overall contrast, sharpness, colorfulness and appearance of artifacts, comparing the TMO's visual rendering "to their internal representation of a 'perfect' image in their minds" [31]. In the *accuracy evaluation*, both pairwise comparison and rating techniques were used in order to evaluate the perceptual accuracy of the rendering algorithms. The pairwise comparison of TMOs was performed without viewing the real scene and subjects were asked to compare the overall impression on image contrast, colorfulness, image sharpness, and overall natural appearance. An additional rating evaluation was performed using the real scenes set up in the adjoining room as references. Here, subjects had to rate image attributes like highlight contrast, shadow contrast, highlight colorfulness, shadow colorfulness, overall contrast and the overall rendering accuracy comparing to the overall appearance of the real-world scenes. In both experiments, observers did not have immediate access to the real scene and had to rely on their memories (either short- or long-term) to perform the tasks.

To validate the iCAM06 operator [32], its authors performed two psychophysical experiments similar to the previous ones [31]. The first experiment was a pairwise comparison without viewing the reference scene. Observers had to choose the tone-mapped image that they preferred based on overall impression on image quality (considering contrast, colorfulness, image sharpness, and overall natural appearance). In the second experiment, observers were also asked to evaluate overall rendering accuracy by comparing the overall appearance of the rendered images to their corresponding real-world scenes, which were set up in an adjoining room.

While looking for a definition of an overall image quality measure, Cadík et al. [29] studied the relationships between some image attributes such as brightness, contrast, reproduction of colors and reproduction of details. They performed two psychophysical experiments, using 14 TMO, in order to propose a scheme of relationships between these attributes, being aware that some special attributes, which were not evaluated (e.g. glare stimulation, visual acuity and artifacts), can influence their relationships. In the first one, 10 subjects were asked to perform ratings using five criteria: overall image quality and the four basic attributes (brightness, contrast, reproduction of detail and colors). These evaluations were performed using a real scene as a reference (a typical real indoor HDR scene). In the second experiment, subjects did not have access to the real scene and they had to rank image printouts according to the overall image quality and the four basic attributes.

In a new study, Cadík et al. [34] performed exactly the same type of experiments adding two new scenes, that is, they had a total of three scenes, i.e. a real indoor HDR scene, a HDR outdoor scene and a night urban HDR scene. In the first experiment, subjects were asked to rate overall image quality and the quality of reproduction of five attributes by comparing samples to the real scene. These attributes were the same four basic ones of their previous work and the lack of disturbing image artifacts (which was one of the non-evaluated special attributes in [29]). These experiments were set-up in an uncontrolled natural environment, so subjects had to perform the experiments at the same time of the day as the HDR image was acquired. In the second experiment, subjects had no possibility of directly comparing to the real scene and had to rank the image printouts according to the overall image quality, and the quality of basic attributes.

A.3. Experiments using an HDR monitor

In 2005, Ledda et al. [27] performed two different psychophysical experiments comparing 6 different tone-mapping operators to linearly mapped HDR scenes displayed on a HDR device. They used 23 different color and grey-scale HDR scenes showing 3 different images per comparison: the HDR and two tone-mapped images. In the first experiment, subjects were asked

to select the TMO image more similar to the HDR reference by judging its global appearance. In the second one, they were asked to make their judgment based on reproduction detail.

In a later work, Akyüz et al. [33] asked subjects to rank six images (1 HDR image, 3 tone-mapped images, 1 objectively good LDR exposure value and 1 subjectively good LDR exposure value) according to their subjective preferences. They found that participants did not systematically prefer tone-mapped HDR images over the best single LDR exposures.

All the previous studies have been focused on subjective comparisons of global and local image appearance attributes such as contrast, colourfulness, sharpness, reproduction artifacts, etc. either within TMOs or against the real scene. While this is no doubt extremely important, we believe a good TMO should output a scene that produces the same visual sensation as the physical scene, in particular the interrelations between objects and their perceived attributes. For instance, no study has been conducted (as far as we know) to evaluate whether objects represented within a TMO image maintain the same perceived visual differences as the real scene. This is the main objective of our work.

B. Tone-Mapping Operators

As mentioned before, TMOs can be classified according to their processing in *global* and *local*. Global operators perform the same computation in all pixels, regardless of spatial position, which make them more computationally efficient at the cost of losing contrast and image detail. Some examples of global TMO are [1, 39, 43, 45]. On the other hand, local operators, which take into account surrounding pixels, produce images with more contrast and higher detail level, but they may show problems with halos around high contrast edges. Local operators are inspired on the local adaptation process present at the early processing stages of the human visual system. Some examples of local operators are [2, 32, 40–42, 44, 47, 49]. There are some tone-mapping operators which could be global or local depending on their setup configuration parameters. One example is [37] and another one is [48], which is developed in two stages, the first global and the second local. A brief summary of the properties of each tone-mapping operator used in our experiments is given in Table 1. The first column shows the names that we will use to refer to each operator throughout this work. The characteristics of each TMO are detailed below:

-*Ashikhmin* [40]. This local tone-mapping operator is inspired by the processing mechanisms present at the first stages of the Human Visual System. Intensity range is compressed by a local luminance adaptation function and, in a last step, detail information is added.

-*Drago* [43]. This global tone-mapping operator is based on luminance logarithmic compression that, depending on scene content, uses a predetermined logarithmic basis to preserve contrast and details.

-*Durand* [41]. This local tone-mapping operator decomposes the image in two layers: the base and the detail. Large-scale variations of the base layer are encoded, while the magnitudes of the detail layer are preserved.

-*Fattal* [42]. This local tone-mapping operator manipulates the gradient fields of the luminance image. Its idea is to identify high gradients in different scales and attenuate their magnitudes, while maintaining their directions.

-*Ferradans* [48]. This tone-mapping operator can be executed as global or local because it is divided in two stages. In the first stage, it applies a global method that implements the visual adaptation, trying to mimic human cones' saturation. In the second stage, it enhances local contrast using a variational model inspired by color vision phenomenology. In our work, this operator was run as local.

-*Ferwerda* [39]. This global tone-mapping operator is based on computational model of visual adaptation that was adjusted to fit psychophysical results on threshold visibility, color appearance, visual acuity, and sensitivity over the time.

-*iCAM06* [32]. This local tone-mapping operator is based on the iCAM06 color appearance model, which gives the perceptual attributes of each pixel, like lightness, chromaticity, hue, contrast and sharpness. It includes an inverse model which considers viewing conditions to generate the result.

-*KimKautz* [1]. This global tone-mapping operator is based on the assumption that human vision sensitivity is adapted to the average log luminance of the scene and that it follows a Gaussian distribution.

-*Krawczyk* [2]. This local tone-mapping operator is inspired on the anchoring theory [50]. It decomposes the image into patches of consistent luminance (frameworks) and calculates, locally, the lightness values.

-*Li* [44]. This local tone-mapping operator is based on multiscale image decomposition that uses a symmetrical analysis-synthesis filter bank to reconstruct the signal, and applies local gain control to the subbands to reduce the dynamic range.

-*Mertens* [46]. This technique fuses original LDR images of different exposure values (exposure fusion) to obtain the final "tone-mapped" image, which avoids the generation of an HDR image. Guided by simple quality measures like saturation and contrast, it selects "good" pixels of the sequence and combines them to create the resulting image. Thus, for this method instead

Table 1. Summary of used TMO's characteristics. Second column shows whether the TMO is global (G) or local (L). Third column shows whether it is inspired by the Human Visual System, and following columns show whether it processes luminace and color information.

TMO	Global/Local	HVS	Luminance	Color
Ashikhmin	L	✓	✓	
Drago	G		✓	
Durand	L		✓	✓
Fattal	L		✓	
Ferradans	L	✓	✓	✓
Ferwerda	G	✓	✓	✓
iCAM06	L	✓	✓	✓
KimKautz	G		✓	
Krawczyk	L		✓	
Li	L		✓	
Mertens	-			
Meylan	L	✓	✓	✓
Otazu	L	✓	✓	
Reinhard	G		✓	
Reinhard-Devlin	G		✓	

of an HDR image we used a stack of LDR images.

-*Meylan* [47]. This local tone-mapping operator is derived from a model of retinal processing. In a first step, a basic tone-mapping algorithm is applied on the mosaic image captured by the sensors. In a second step, it introduces a variation of the center/surround spatial opponency.

-*Otazu* [49]. This local tone-mapping operator is based on a multipurpose human colour perception algorithm. It decomposes the intensity channel in a multiresolution contrast decomposition and applies a non-linear saturation model of visual cortex neurons.

-*Reinhard* [37]. This tone-mapping operator can be executed as global or local. It performs a global scaling of the dynamic range followed by a dodging and burning (local) processes. In our work, this operator was run as global which is its default value in the toolbox.

-*Reinhard-Devlin* [45]. This global tone-mapping operator uses a model of photoreceptors adaptation which can be automatically adjusted to the general light level.

3. METHODS

In order to compare TMOs, we performed two different experiments called *Segment Matching* and *Scene Reproduction* experiments. The aim of the first experiment was to study the internal relationships among grey-levels in the tone-mapping image and in the real scene (i.e. a segment matching experiment similar to [51]). The aim of the second experiment was to evaluate TMOs according to how similar their results were perceived to be with respect to the real scene. In both cases, we obtained a ranking of the different TMOs. Behind these experiments is the idea that a good TMO is one whose output is perceptually similar to the real scene and, to do that, a good reproduction of the objects' relationships is needed.

A. Materials

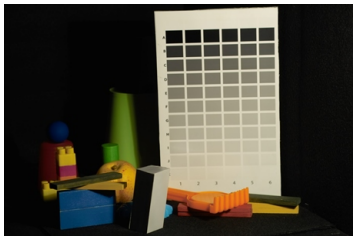
Our experiments were performed in a controlled environment where the only sources of light were a lamp, which illuminated the real scene and a CRT screen. We used a ViSaGe MKI Stimulus Generator and a Mitsubishi Diamond-Pro®2045u CRT monitor side-by-side with a handmade real HDR scene. The monitor was calibrated via a customary Cambridge Research Systems Ltd. software for ViSaGe MKI Stimulus Generator (Rochester, England) and a ColorCal (Minolta sensor) suction-cup colorimeter. Both the monitor and the real scene were setup so that the objects in both scenes subtended approximately the same angle (18.13°x 13.81°) and looked similarly positioned to the observer.

We built three different HDR scenes, each including a grey-level reference table and two solid parallelepipeds (cuboids). The reference table was built by printing a series of 65 grey squares (2.8 x 2.2 cm) arranged in a flat 11x6 distribution. The arrangement of rows and columns was labelled A,B,C,...,K for the rows and 1,2,3...,6 for the columns. The lightness of these patches decreased monotonically from the top (patch A1 - #1) to the bottom (patch K5 - #65), as measured by our PR-655 SpectraScan®Spectroradiometer. The printed values were selected so that their CIE L* (lightness) value was equally spaced, meaning that their distribution was approximately uniform in terms of perceived lightness (see Table 2). The cuboids consisted of pieces of wood (3.6 x 3.6 x variable length between 9.4 and 10 cm), whose sides (facets) were covered with arbitrary samples of the same printed paper as the reference table. There were two cuboids in each scene (one under direct illumination and the other in the shade). The third column of Table 3 shows the patch of the reference table that the cuboid's facet corresponded to, the fourth column indicates its position with respect to the illumination and the last column indicates its luminance (when placed within its scene). Table 2 also shows the luminance values for these patches once lit by our light source. The chromaticity of all printed material was CIE xy= (0.3652, 0.3817). The rest of the scenes consisted of many plastic and wooden objects of different colours and shapes (see Figure 1).

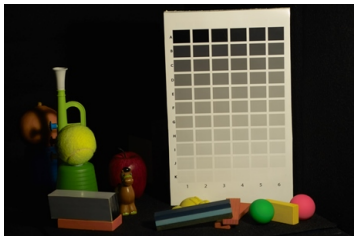
Two facets of one cuboid and three of the other were always visible from the subjects' location, resulting in 15 different grey

Table 2. In this table we show the L* CIELab colour space values and the luminance values (cd/m^2) of each patch in the reference table. The lightness values have been measured by our PR-655 SpectraScan®Spectroradiometer under a uniform illumination. In contrast, the luminance values were measured in the scene by the same Spectroradiometer. The lightness of the patches monotonically increases from patch #1 (A1) to patch #65 (K5). Middle gray which is universally defined as 18% reflectance on a white surround. In this table, 18% max is patch #34 (F4).

Lightness (L*) of patches in the reference table						
Coordinate	1	2	3	4	5	6
A	1.50	3.64	5.76	7.88	9.97	12.06
B	14.13	16.18	18.22	20.25	22.26	24.26
C	26.24	28.20	30.15	32.08	34.00	35.90
D	37.78	39.65	41.50	43.33	45.15	46.95
E	48.73	50.49	52.24	53.96	55.67	57.36
F	59.03	60.68	62.31	63.91	65.50	67.07
G	68.61	70.14	71.64	73.11	74.57	76.00
H	77.40	78.78	80.13	81.46	82.76	84.03
I	85.27	86.49	87.67	88.82	89.93	91.01
J	92.05	93.06	94.02	94.94	95.81	96.63
K	97.40	98.11	98.74	99.30	99.73	
Luminance (cd/m^2) of patches in the reference table						
Coordinate	1	2	3	4	5	6
A	0.559	0.565	0.617	0.692	0.853	0.815
B	0.901	0.957	1.007	1.360	1.484	1.622
C	1.575	1.734	1.992	2.218	2.624	2.978
D	3.031	3.243	3.651	4.133	4.634	5.076
E	5.215	5.718	6.439	7.199	7.827	10.39
F	12.95	15.85	16.94	18.69	20.63	21.32
G	22.53	23.67	25.79	28.34	30.40	31.47
H	32.66	34.2	37.69	38.40	42.75	44.63
I	46.28	48.39	52.27	55.57	58.1	60.61
J	61.76	66.17	67.75	69.82	74.71	78.24
K	78.96	84.67	90.37	96.38	104.0	



(a) Scene 1



(b) Scene 2

(c) Scene 3

Fig. 1. To show the general appearance of the physical scenes, here we show a single LDR exposure (chosen by simple visual inspection by the authors) from the set of LDR exposures used to create the HDR images. Since they are a single LDR exposure, the cuboids in the dark regions are not completely visible in these pictures.

Table 3. Photometric assessment of the scene facets used in our matching experiments. Column 3 indicates the corresponding gray patch (same material) from the reference table. Column 4 indicates whether the facet was directly illuminated or in the shade and column 5 shows its luminance from the observer's point of view.

Luminance values (cd/m^2) of the scene facets				
Scene	Facet	Original Patch	Light/Shade	Luminance (cd/m^2)
1	1	B4	light	0.66
1	2	H4	light	59.22
1	3	F4	light	33.38
1	4	B4	shade	5.34
1	5	F4	shade	7.48
2	6	C4	light	3.44
2	7	H1	light	32.13
2	8	C1	shade	5.93
2	9	J1	shade	12.00
2	10	E4	shade	0.80
3	11	A4	light	1.73
3	12	I1	light	25.93
3	13	J1	light	27.46
3	14	G4	shade	31.58
3	15	B1	shade	9.06

facets in total (see Table 3). The incandescent lamp (100W) had its bulb painted blue to simulate D65 illumination and was set up so that the luminance of the brightest object was about the same as the maximum luminance the monitor was capable of producing (about 100 cd/m^2).

We photographed the real scene using a Sigma Foveon SD10 camera placed in the exact same position as the subjects' heads during the psychophysical experiments. The same camera was calibrated for use in other measurements [52] and because of this, we have a fairly good idea of the linearity and spectral sensitivity of its sensors. The setup was arranged so that the images presented on the monitor looked geometrically the same as the real ones shown beside it. Since the walls were covered in black felt, reflections from all other objects were minimized. The dynamic range of the scenes as measured by multiple exposures using the camera were approximately 10^5 for scene 1 and 10^6 for scenes 2 and 3. The dynamic range of the reference table as measured by the PR-655 was 104.0 - 0.559 cd/m^2.

Although it has been shown that because of glare, is not possible to achieve an accurate representation of scene luminance distribution from a combination of many LDR images, this technique can still provide a good enough approximation [8]. In consequence, a set of 25 photographs were taken at different exposure values (from 15 sec to 1/6000 sec) using the same aperture, focal distance, zoom settings and visual field. Individual images were stored in RAW format and transformed into 16 bits sRGB (using the camera manufacturer's software).

To avoid any bias regarding the operators, all experiments started with a 1-minute subject adaptation to the ambient light. Most TMO implementations were obtained from the popular HDR Toolbox for MATLAB [53] while others (*Ferradans*, *iCAM06*, *Li*, *Meylan* and *Otazu*, were obtained from their corresponding authors' web pages). In order to avoid benefiting any of the TMOs, we ran all of them with their default settings. In *Ferradans*' case, we had to chose between two different parameters and we selected the default values specified in their paper ($\rho = 0$ and $\alpha^{-1} = 3$). Other cases required that the TMO's author was asked to perform the best tone-mapping, but we discarded this option because of its impracticality (we could not ask all authors the same) and besides, this practice impairs the reproducibility of the results.

4. EXPERIMENTS

A. Experiment 1: Segment Matching

A.1. Procedure

The Segment Matching experiment consisted on two different tasks:

Task 1. After adaptation, subjects were asked to match, in the real scenes (i.e. with monitor turned off), the brightness of the 5 cuboids' facets to the brightness of the patches in the reference table in each scene (see Figure 2a). Although there were no time constraints to perform the tasks, subjects were advised to take no more than 30 seconds per match.

Task 2. Here the real scene was not visible and the observers only saw digital (tone-mapped) versions presented on the monitor. Their task was similar to *Task 1*, except that all matchings were conducted entirely between the facets and patches shown on the screen (see Figure 2b).

There were three conditions for Experiment 1, corresponding to the three different scenes created (see Figure 1). Observers performed 240 matchings in total (5 facets x 15 different tone-mapped images x 3 three scenes plus 15 matchings in the real scenes). In practice, all matchings were conducted by writing for

each facet, the coordinates of the matching reference table patch on a piece of paper. The presentation order of the tone-mapped images was randomized.

A.2. Experimental Design

In Experiment 1, the independent variables (IVs) were the cuboid's facets and the reference table patches. The dependent variables (DVs) were the subjects' segment matches in the tone-mapped images (*Task 2*) and the control variable (CV) were the subjects' segment matches in the real scene (*Task 1*). Our null hypothesis was that there was not significant difference between the segments matched in the real scene (CV - *Task 1*) and the matches in the tone-mapped images (DVs - *Task2*) because the TMOs perfectly reproduce the perceptual relationships among the objects present in the real scene.

A.3. Participants

Task 1 was completed by a group of 12 observers with normal or corrected-to-normal vision, recruited from our lab academic/research community. This group (8 males and 4 females) was comprised by people aged between 17 and 54. Nine of them were completely naïve to the aims of the experiment. *Task 2* was completed by 10 of the previous observers (8 males and 2 females).

A.4. Results

Figure 3 shows a plot of the segments matches obtained in *Task 2* against the segments matches in *Task 1*. We fitted a linear model to the results obtained by each TMO. If a TMO reproduced well the interrelations among the grey facets, the fitting should be very similar to the fitting for the real scene (i.e. points should lay about the diagonal).

We performed two different analyses to evaluate to what extent the local interrelations perceived by the observers in the tone-mapped versions corresponded to those perceived in the real scene. In the first analysis, we studied the slopes of the different fitted linear models w.r.t. the slope obtained in the real scene. The smaller the difference, the better the reproduction of the interrelations (it means that the TMO maintained the relationships among the facets and patches). Figure 3 shows the offset between the lines fitted to the TMOs and the line fitted to the real scene. In the second analysis, we studied this displacement by computing the root mean squared error (RMSE) between them.

All results are shown in Table 4, where *iCAM06* has the smallest distance to the real scene in both analyses. Since its slope difference and RMSE are very small, we can assume that the pixel interrelations in its tone-mapped image perceptually mimic the real scene. Given that *iCAM06* is based on a color appearance model that considers perceptual attributes such as lightness, chromaticity, hue, contrast and sharpness, its results are expected to be in line with observers' perception.

We calculated the Spearman's rank correlation coefficient between the rankings obtained from both Segment Matching analyses (see Table 4) and obtained a value of 0.59 ($p < 0.05$). Since both rankings are quite similar, it is worth paying attention to some interesting cases such as *Ferradans*, whose slope is very close to that of the real scene, but the fitted model lays systematically under the real scene's line (i.e. its RMSE is very big). An opposite example is *Mertens* which has a different slope, but its RMSE is the second smallest.

Another interesting observation from Figure 3 is that, at the lowest and highest brightness values, the agreement between subjects is higher than at middle values (both horizontal and vertical dispersion lines are smaller). This suggests that the TMOs are more accurate at reproducing both the brightest and the darkest parts of the image. To analyze this effect in more detail, we studied the subjects' results for each facet. In Figure 4, the abscissa shows the segments matched in the real scene ordered from darkest to brightest and the ordinate represents the RMSE in the tone-mapped images with respect to the real scene. We defined RMSE as: $RMSE_{scene} = \sqrt{\frac{1}{n}\sum_{\forall i}(x_i - y_i)^2}$, where x_i is the i-th subject segment matched in the real scene, y_i is the i-th subject segment matched in the tone-mapped image and n is the number of subjects. Again, in almost all TMOs, the RMSE value is smaller for darkest and brightest facets than for mid-grey facets. Thus, not only the agreement between subjects but also the error ($RMSE_{scene}$) is lower for both brightest and darkest values.

Table 4. Performance of all TMOs in the Segment Matching experiment. The second and third columns show the analysis' results and the last the type of the TMO. In both metrics (i.e. slope difference and root mean squared error -RMSE- between the diagonal and the TMO fitted line), the smaller (indicated in bold), the more similar to the real scene, and thus, the better.

TMO	Slope Difference	RMSE	Type
Ashikhmin	0.29	11.19	Local
Drago	0.15	7.02	Global
Durand	0.12	5.01	Local
Fattal	**0.02**	4.66	Local
Ferradans	**0.00**	5.12	Local
Ferwerda	0.09	7.09	Global
iCAM06	**0.01**	**0.42**	Local
KimKautz	0.09	5.51	Global
Krawczyk	0.09	5.46	Local
Li	**0.01**	4.02	Local
Mertens	0.15	3.84	-
Meylan	0.16	5.93	Local
Otazu	0.24	5.30	Local
Reinhard	0.15	6.72	Global
Reinhard-Devlin	0.16	6.82	Global

B. Experiment 2: Scene Reproduction

B.1. Procedure

Experiment 2 consisted of a pairwise comparison of tone-mapped images obtained using different TMOs in the presence of the original scene (side by side). After 1-minute adaptation in front of the physical scene, a pair of tone-mapped images of the same physical scene was randomly selected and presented sequentially to the observer on the CRT screen besides the real scene. Subjects could press a gamepad button to toggle which image of the tone-mapped pair was presented on the monitor (only one image was displayed at a time). For this task, they were asked to 'select the image that was more similar to the real

(a) Task 1 **(b)** Task 2

Fig. 2. In Experiment 1, observers performed two tasks. In *Task 1* (Figure 2a), observers had to match the brightnesses of the 5 cuboids' facets to the brightnesses of 5 patches in the reference table. In *Task 2* (Figure 2b), observers had to perform the same task on the TMO image displayed on the calibrated monitor. (Red arrows are randomly drawed just for illustrative purposes).

scene.' As before, there was no time limit but subjects were advised to complete a trial in less than 30 seconds. After an image was chosen, a grey background was shown for two seconds, and a different random pair was selected for the next trial. Every subject performed 105 comparisons per scene taking around 25 minutes in total. There were three experimental conditions, corresponding to the three different physical scenes created (see Figure 1). Between conditions, subjects were forced take 5 to 10 minutes breaks outside while the physical scene was replaced.

B.2. Experimental Design

In this experiment the IVs were the different TMOs, the DVs were the subjects' evaluations (i.e. the preference matrix), and the CV was the real scene. Our null hypothesis was that there were no differences in the TMOs performances since all of them perceptually reproduce the real scene.

B.3. Participants

A group of 10 people with normal or corrected-to-normal vision, 7 males and 3 females recruited from our lab academic and research community, completed this experiment. This group was comprised by people aged between 17 and 54 y.o. Seven of them were naïve to the aims of the experiment.

B.4. Results

From the pairwise comparison results, we defined a preference matrix for each subject and each scene. We constructed a directed graph where the nodes were the evaluated TMOs and the arrows pointed from a preferred TMO to a non-preferred TMO, e.g. if the TMO_i is preferred over the TMO_j (tone-mapped image from TMO_i is more similar to the real scene than the one from TMO_j), we drew an arrow from node_i to node_j, for $i \neq j$.

From this graph, we were able to analyse intra-subject consistency coefficient ζ for each scene. The consistency coefficient for each subject and scene is defined by

$$\zeta_{st} = \begin{cases} 1 - \frac{24 d_{st}}{n^3 - n}, \text{if } n \text{ is odd.} \\ 1 - \frac{24 d_{st}}{n^3 - 4n}, \text{if } n \text{ is even.} \end{cases} , with$$
$$d_{st} = \frac{n(n-1)(2n-1)}{12} - \frac{1}{2}\sum_{i=1}^{n} a_{ist}^2 \quad (1)$$

where s is the scene number ($s \in [1,3]$), t is the subject number ($t \in [1,m]$), n is the number of evaluated TMOs, and a_i is the number of arrows which leave the node_i. The maximum ζ value is 1 (perfect consistency within-subject).

The consistency between subjects, i.e. inter-subject agreement, is measured by the Kendall Coefficient of Agreement [27, 54]. This measure is defined by

$$u_s = \frac{2\sum_{i \neq j} \binom{p_{ij}}{2}}{\binom{m}{2}\binom{n}{2}} - 1 \quad (2)$$

where p_{ij} is the number of times TMO_i is preferred over TMO_j and m is the number of subjects. Since the number of subjects is even ($m = 10$), the possible minimum value of u, given by Equation 2, is $u = -\frac{1}{m-1}$ and its possible maximum value is $u = 1$.

In order to study if u_s values are significant, we used the chi-squared test (χ^2). The χ_s^2 values are defined by

$$\chi_s^2 = \frac{n(n-1)(1 + u_s(m-1))}{2} \quad (3)$$

The number of degrees of freedom of the chi-squared test is given by $\frac{n(n-1)}{2}$.

Table 5. Summary of all statistical analysis from section B.4. We computed the intra-subject evaluation (consistency coefficient ζ), the inter-subjects evaluation (Kendall Agreement Coefficient u) and calculated chi-squared test to see if u values were significant.

Scene	$\bar{\zeta}$	u	χ^2	p, 105 df
1	0.91	0.61	681	< 0.001
2	0.95	0.65	719	< 0.001
3	0.93	0.55	624	< 0.001

In Table 5, we show all statistical measures for each scene, where we can see that intra- and inter-subject consistency values are very high and statistically significant. Then, in Figure 5, we show the results of the overall paired comparison evaluations for every scene (obtained from Thurstone's Law of Comparative Judgment, Case V [55]) with 95% confidence limits. Spearman's correlation between these rankings shows that TMOs have similar behaviour across different scenes (their coefficients are equal or higher than 0.90, with $p < 0.05$). We computed the mean value along all the scenes (Table 6) and observed that the best ranked TMOs were *KimKautz*, *Krawczyk* and *Reinhard*, which are completely different from the rankings obtained in the previous experiment.

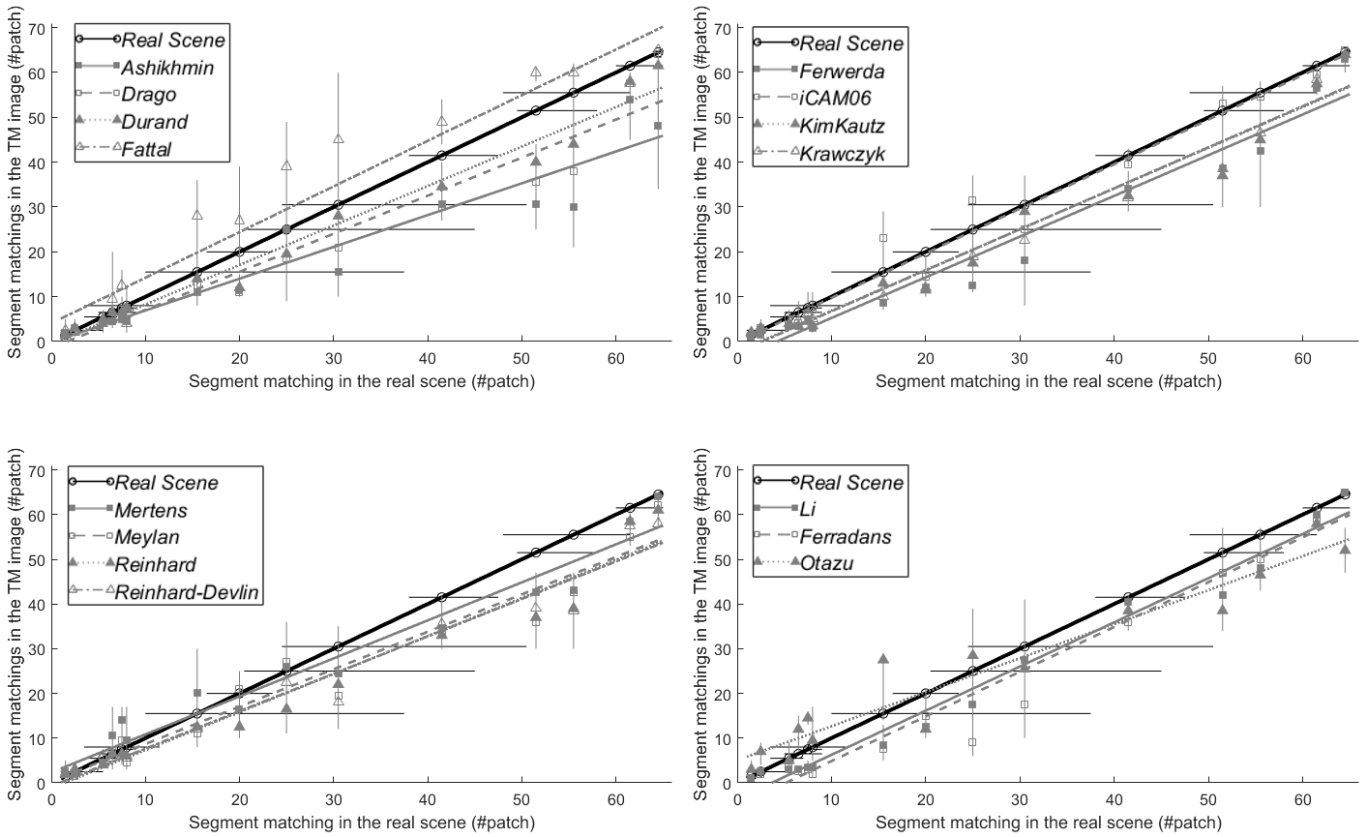


Fig. 3. Results of Experiment 1. Segments matches in the tone-mapped images are plotted against segment matches in the real scene. Markers and lines identify each TMOs. Since not all the data had a normal distribution, the markers show the median of the subjects' observations. Horizontal lines indicate the first and the third quartiles of *Task 1* and vertical lines indicate the first and the third quartiles of *Task 2*. For each operator, we fitted a linear model using the median of the subjects' observations. The figure is divided in four panels for clarity. The real scene is plotted against itself in all panels to provide a fixed reference ($y = x$). In summary, the better the TMO, the closer its fit to the solid black line.

5. DISCUSSION

Comparing the results of our two experiments, we observe that in Section A (Segment Matching experiment - see Table 4), local TMOs are significantly better than global ones. On the contrary, in Section B (Scene Reproduction experiment - see Table 6), global TMOs are significantly better than local ones. We computed Spearman's correlation coefficient between both experiments rankings and verified that there is no correlation.

An interesting example of this lack of correlation is *iCAM06*. It is clearly at the top of the rankings in the Segment Matching experiment, but it is in the middle position in the Scene Reproduction experiment. This means that it correctly reproduces relationships among grey-levels, but overall features are not maintained. An extreme example is *Fattal*, which is in the fourth position in the Segment Matching rankings, but is the last in the Scene Reproduction ranking. This can be explained because *Fattal* is based on local (or spatial) features, e.g. luminance gradients, but it does not enforce global features (such as global brightness and contrast). In fact, from Table 4 (RMSE results) we can conclude that *Fattal* produces a tone-mapped image which is systematically brighter than the real scene. Since *Fattal*'s fitted line has almost the same slope as the real scene (see Figure 3) removing this offset could improve its performance in the Scene Reproduction experiment.

From the previous results, we infer that overall appearance does not only depend on the correct reproduction of intensity relationships, but it might depend on many other weighted local attributes, such as the reproduction of grey-level and color relationships, contrast, brightness, artifacts, level of detail, etc. This is in agreement with other authors [26, 29, 30, 34]. Furthermore, our results show that overall attributes should also be considered to correctly reproduce the appearance.

Regarding the question of which is the best TMO, *KimKautz* and *Krawczyk* are very close in all rankings, hence both can be considered equally good.

A. Comparison to other Studies

In Section 4 A (Segment Matching) we took into account a particular criterion which, up to our knowledge, has never been studied in this kind of TMO ranking experiments. Moreover, we compare our Segment Matching results to the results obtained by other works that study TMOs applied to grey-level images (given that our analysis has been performed on grey-level facets).

Many works perform overall appearance comparisons, either with (as in our work) or without the real scene. Although Kuang et al. [25] performed an experiment without a real scene reference, our scene reproduction results agree with theirs in that *Fattal* is the worst ranked operator and *Reinhard* is one of the best ranked. Contrary to our results, Kuang et al. [25] conclude that *Durand* is better than *Reinhard*. The reason could be that

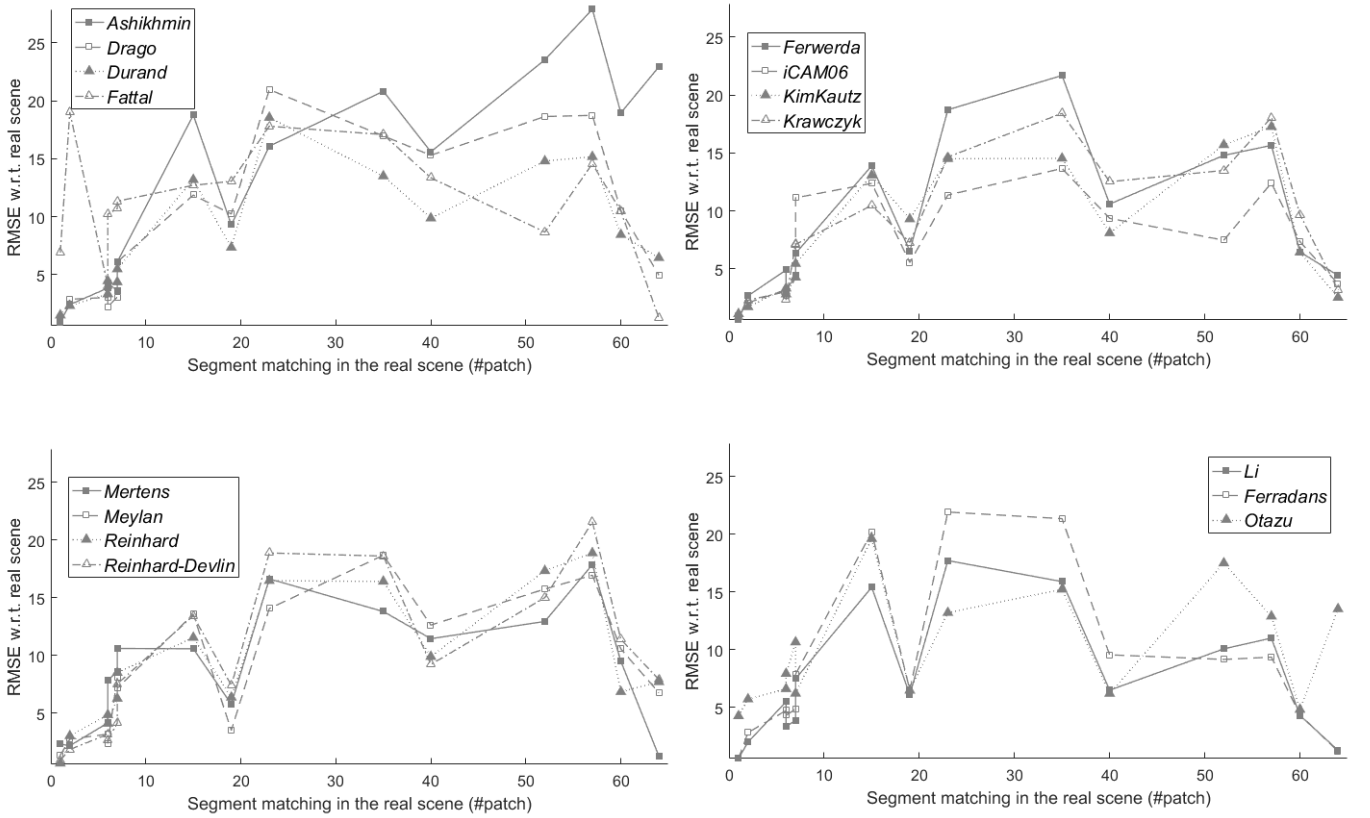


Fig. 4. The RMSE w.r.t. the real scene ($RMSE_{scene}$) is the difference between segments matched in the tone-mapped image and in the real scene. Different types of lines and markers represent different TMOs. Abscissa represents the segments matched in the real scene ordered from darkest to brightest. According to this metric, the smaller the value, the better the TMO.

they might have run *Reinhard* in local operator mode, which we did not. Furthermore, they performed a study with grey-scales images and their results showed that *Durand* was better than *Reinhard*, but iCAM was worse than *Reinhard*, which is approximately similar to our Segment Matching experiment's results. They differ in iCAM's result, but they used iCAM [56] instead of *iCAM06*, as in our case.

Yoshida et al. [26, 30] performed experiments with architectural indoor HDR scenes and they concluded that *Reinhard* and *Drago* were good in terms of naturalness and *Durand* was not ranked as highly as in [25] (in an experiment without the reference scene). Our results agree with Yoshida et al. [26, 30]. Moreover, Yoshida et al. [30] showed that global and local operators obtain different results, but global TMO results are more similar among themselves than local TMO. As pointed out in the previous section, this relationship is also present in our study (Tables 4 and 6).

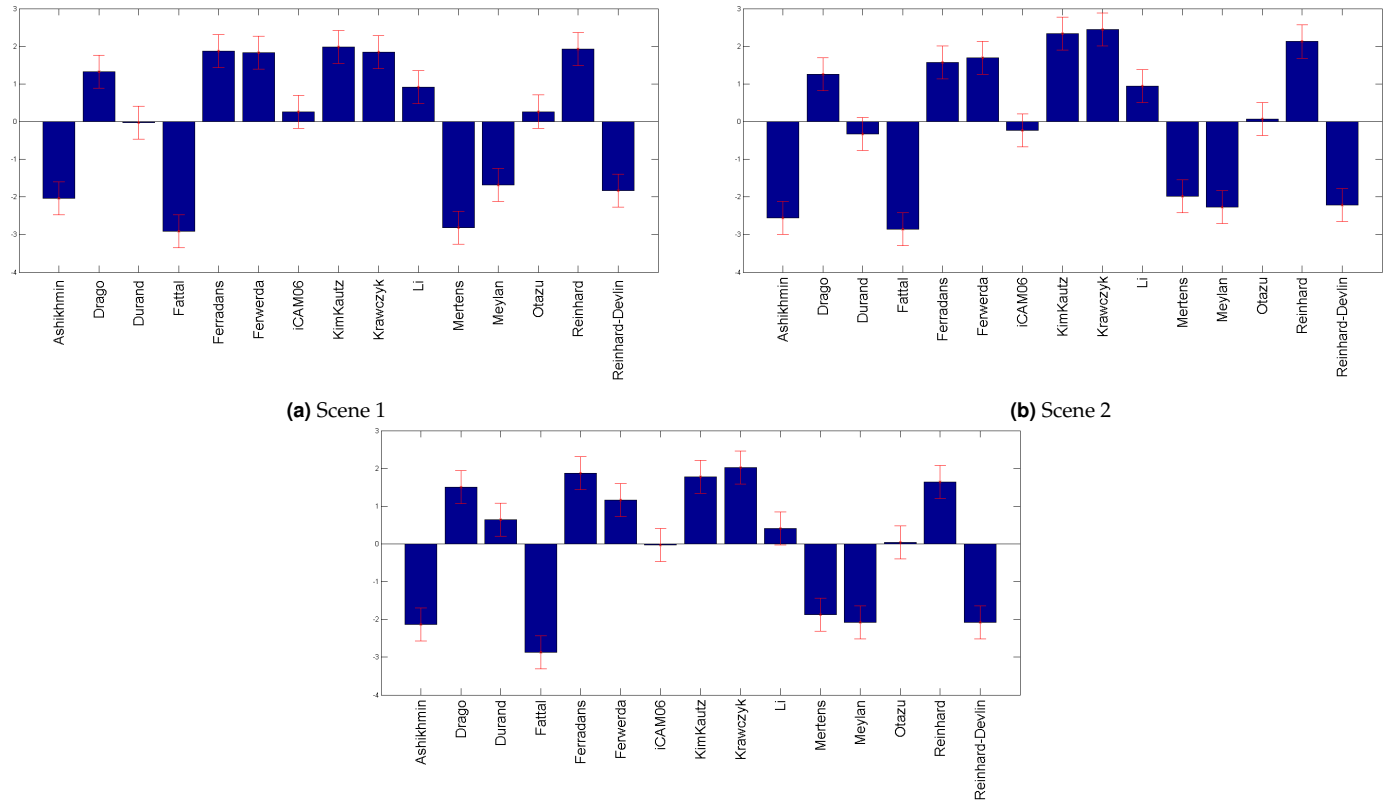
Ledda et al. [27] used a High Dynamic Range display and obtained a ranking according to the overall similarity of TMO images. In this ranking, iCAM was the first one, which does not agree with our results. In addition, their ranking shows the following TMO's order: *Reinhard*, *Drago* and *Durand*, which match to our results. These authors also performed experiments in grey-scales obtaining *Reinhard* as the best ranked, which does not agree with our results.

Cadík et al. [29, 34] performed a very exhaustive study of perceptual attributes. We agree with some of their results like the good ranking of *Reinhard* (close to the best) and the unnaturalness of *Fattal*. Moreover, we strongly agree with them in that the best overall quality is generally observed in images produced by global tone-mapping operators. Nevertheless, we want to point out that there was some conflict between these two studies. In the first one ([29]), *Durand* was the worst ranked operator, ranked even lower than *Fattal*, but in the second one ([34]), *Fattal* was the worst ranked and *Durand* was in a middle position. Our results are in line with Cadík et al. [34].

We do not agree with Kuang et al. [31] in that *Durand* is always the best ranked operator (with and without a reference scene). Furthermore, in contrast with our results, *Reinhard* is in a middle position of their ranking.

Kuang et al. [32] suggested, again, that *Durand* was better than *Reinhard* and *iCAM06* was even better than *Durand*. In our results, *Durand* and *iCAM06* are quite close, but *Reinhard* is much better than them. Again, *Reinhard* could have been run in local TMO mode.

In a similar study as Kuang et al. [25], Ashikhmin and Goyal [28] concluded that, comparing to the real scene, *Fattal* and *Drago* were two of their overall best performers. We do not agree that *Fattal* is one of the best performers, but we have to point out that, in their work, they tuned the TMO's parameters, which implies that *Fattal* could be a good TMO when a fine tuning of the parameters is performed. Furthermore, in their work, *Drago* obtained more or less the same results as *Fattal*, but *Reinhard* obtained worse results than them. They do not specify how they run *Reinhard*, but it is possible that they run it in the local mode. They obtained that the trilateral filtering [57], which



(a) Scene 1 **(b)** Scene 2

(c) Scene 3

Fig. 5. Case V Thurstone Law's scores for each evaluated TMO for each different scene. Thurstone scores are an arbitrary measure that shows how many times a particular TMO is better than the other ones. Thus, in that case, the higher score, the better TMO. Vertical lines show the 95% confidence limits.

is an improvement of *Durand*, was the worst ranked TMO, so it makes sense that, in our work, *Durand* has obtained worse results than *Drago* and *Reinhard*.

In [33], the outputs of the most internally sophisticated TMO are statistically worse than the best single LDR exposure. Since a global operator is generally less sophisticated than a local, we could expect that global TMO results are better than local TMO results. Contrary to this theory, *Mertens* (which cannot be considered a sophisticated TMO because it uses single exposure values) is on middle positions in the Segment Matching experiments but it is one of the worst ranked in the Scene Reproduction experiments.

Some authors emphasize the creation and use of particular metrics to compare tone-mapped images. For example, Ferradans et al. [48] performed an evaluation of several TMOs using the metric of Aydin et al. [22]. Although it is not the purpose of our work, we performed a very preliminary analysis comparing our results to those of Aydin et al.'s [22] as shown in [48]. We agree that *Fattal* was the operator with highest total error percentages, but disagree with the general overall TMOs ranking. A detailed analysis comparing numerical metrics and psychophysical results is scheduled for future work.

It is possible to identify several shortcomings in our study that need to be addressed before a more definitive conclusion is achieved. Firstly, we have assumed that the software provided by the Sigma camera manufacturers is accurate enough to convert the scene luminance array to the sRGB digital file used as input to all TMO algorithms. This assumption hides possible inaccuracies because of glare effects, lens aberrations and possible tone/chroma enhancements. In the past, we calibrated this camera and measured the linearity and spectral sensitivity of its sensors for use in daylight settings [52] and verified that tone/chroma enhancements are kept to a minimum at least for its raw image settings. For this work we did not employ our own calibration (which is valid within a fairly limited dynamic range) but decided to rely on the manufacturer's algorithm instead. All these limits the reproducibility of our experiments (unless of course the same camera is used). We are also aware that the absence of an accurate radiometric description of our scenes also limits the reproducibility of our experiments. To this end we provide photometric information at least of the patches and facets used in the matching comparisons (see Tables 2 and 3) and the dynamic range of the both the monitor and the scenes (see Section 3 A).

6. CONCLUSIONS

Our results show that TMO quality rankings strongly depend on the criteria used for the psychophysical evaluation. Not surprisingly, on one hand, local TMOs are better than global TMOs on our Segment Matching experiment because these operators do not consider just a pixel, but also a region of pixels (i.e. spatial information). On the other hand, global TMOs are better than local ones in our Scene Reproduction experiment. We have found no significant correlation between Segment Matching and

Table 6. Ranking obtained by averaging the scores given by Case V Thurstone Law in the three different scenes. In this ranking, the higher, the more similar to the real scene.

Averaged Thurstone Law's Scores		
TMO	Score	Type
Krawczyk	2.10	Local
KimKautz	2.03	Global
Reinhard	1.90	Global
Ferwerda	1.57	Global
Ferradans	1.48	Local
Drago	1.36	Global
Li	0.75	Local
Otazu	0.13	Local
Durand	0.10	Local
iCAM06	0.00	Local
Meylan	-2.01	Local
Reinhard-Devlin	-2.04	Global
Mertens	-2.22	-
Ashikhmin	-2.25	Local
Fattal	-2.89	Local

Scene Reproduction rankings, showing that observers are using several visual attributes to perform their tasks and some of these attributes are not considered by TMOs. We conclude that TMOs should take into account both local and global characteristics of the image, which implies that there is ample room for improvement in the future development of TMO algorithms. Furthermore, we suggest that an agreed standard criteria should be defined for a proper and fair comparison among them.

Our rankings also show there is no TMO that is clearly better than all the others across our experiments, but *KimKautz* and *Krawczyk* are perhaps the best ranked since they do not underperform in any of the metrics.

As a general conclusion, since none of the tested TMOs satisfies all the testing criteria ("Segment Matching", "Scene Reproduction" and their respective analyses), operators have to be selected depending on each particular task. This is a consequence of the lack of coherent understanding of the goals of a TMO, which is reflected in the wide variety of evaluation methods and results present in the literature. From a scientific point of view, a TMO should aim to perceptually reproduce the real scene instead of modifying image appearance according to aesthetics (for which we already have a wide selection of image tools). Having said so, it is also important to consider that these operators are widely used in digital cameras and mobile phone's cameras and TMO users often prefer aesthetic improvements over accurate scene reproduction.

FUNDING INFORMATION

This work is partially supported by:

Spanish Ministry of Economy, Industry and Competitiveness (DPI2017-89867-C2-1-R)

Agencia de Gestio d'Ajuts Universitaris i de Recerca (AGAUR) (2017-SGR-649).

CERCA Programme / Generalitat de Catalunya.

ACKNOWLEDGMENTS

We would like to thank Carlo Gatta for his useful comments on the psychophysical experiments design, Javier Retana for his useful comments on statistical analysis procedures and the reviewers who did very interesting and useful comments about the paper and the work.

Thanks to all subjects who have participated in the psychophysical experiments, and all authors who publicly share their code.

REFERENCES

1. M. Kim and J. Kautz, "Consistent tone reproduction," in "Proceedings of Computer Graphics and Imaging," (2008).
2. G. Krawczyk, K. Myszkowski, and H. Seidel, "Lightness perception in tone reproduction for high dynamic range images," in "Proceedings of Eurographics," (2005), 3.
3. J. Ferwerda and S. Luka, "A high resolution, high dynamic range display system for vision research," Journal of Vision **9** (2009).
4. E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec, *High Dynamic Range Imaging: Acquisition, Display and Image-Based Lighting* (Morgan Kaufmann Publishers Inc., 2005), chap. 6, pp. 187–221, 1st ed.
5. R. Snowden, P. Thompson, and T. Troscianko, *Basic Vision an Introduction to Visual Perception* (Oxford University Press, 2006).
6. J. McCann, "Art, science, and appearance in hdr images," Journal of the Society for Information Display **15**, 709–719 (2007).
7. C. Parraman, "The drama of illumination: artist's approaches to the creation of hdr in paintings and prints," in "Proc.SPIE," , vol. 7527 (2010), vol. 7527, pp. 7527 – 7527 – 12.
8. J. McCann and A. Rizzi, *The Art and Science of HDR Imaging* (WILEY, 2012), chap. 13, pp. 119–121, 1st ed.
9. C. Mees, *The fundamentals of photography* (Eastman Kodak Company, Rochester, N.Y., 1921), 2nd ed.
10. H. Barlow, "Summation and inhibition in the frogs retina," The Journal of Physiology **119**, 69–88 (1953).
11. A. Derrington, J. Krauskopf, and P. Lennie, "Chromatic mechanisms in lateral geniculate-nucleus of macaque," The Journal of Physiology **357**, 241–265 (1984).
12. C. Blakemore and F. Campbell, "On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images," The Journal of Physiology **203**, 237–260 (1969).
13. E. Land, "The retinex," American Scientist **52**, 247–253,255–264 (1964).
14. E. Land and J. McCann, "Lightness and retinex theory," Journal of the Optical Society of America **61**, 1–11 (1971).
15. J. McCann, "Capturing a black cat in shade: past and present of retinex color appearance models," Journal of Electronic Imaging **13**, 36–47 (2004).
16. J. McCann, "Retinex at 50: color theory and spatial algorithms, a review," Journal of Electronic Imaging **26** (2017).
17. J. McCann, "Lessons learned from mondrians applied to real images and color gamuts," in "7th Color Imaging Conference: Color Science, Systems and Applications," (1999), pp. 1–8.
18. X. Otazu, M. Vanrell, and C. Parraga, "Multiresolution wavelet framework models brightness induction effects," Vision Research **48**, 733–751 (2008).
19. X. Otazu, C. A. Parraga, and M. Vanrell, "Toward a unified chromatic induction model," Journal of Vision **10** (2010).
20. J. McCann, C. Parraman, and A. Rizzi, "Reflectance, illumination, and appearance in color constancy," Frontiers in Psychology **5** (2014).
21. A. Ruppertsberg, , M. Bloj, F. Banterle, and A. Chalmers, "Displaying colourimetrically calibrated images on a high dynamic range display,"

Journal of Visual Communication and Image Representation **18**, 429–438 (2007).
22. T. Aydin, R. Mantiuk, K. Myszkowski, and H. Seidel, "Dynamic range independent image quality assessment," ACM Transactions on Graphics **27** (2008).
23. H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," IEEE Transactons on Image Processing **22**, 657–667 (2012).
24. F. Drago, W. Martens, K. Myszkowski, and H. Seidel, "Perceptual evaluation of tone mapping operators," in "ACM SIGGRAPH Conference Abstracts and Applications," (2003).
25. J. Kuang, H. Yamaguchi, G. Johnson, and M. Fairchild, "Testing hdr image rendering algortihms," in "IS&T/SID 12th Color Imaging Conference," (2004).
26. A. Yoshida, V. Blanz, K. Myszkowski, and H. Seidel, "Perceptual evaluation of tone mapping operators with real-world scenes," in "Human Vision & Electronic Imaging X," (SPIE, 2005).
27. P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, "Evaluation of tone mapping operators using a high dynamic range display," ACM Transactions on Graphics **24**, 640–648 (2005).
28. M. Ashikhmin and J. Goyal, "A reality check for tone-mapping operators," ACM Transactions on Applied Perception **3** (2006).
29. M. Cadík, M. Wimmer, L. Neumann, and A. Artusi, "Image attributes and quality for evaluation of tone mapping operators," in "14th Pacific Conference on Computer Graphics and Applications," (2006), pp. 35–44.
30. A. Yoshida, V. Blanz, K. Myszkowski, and H. Seidel, "Testing tone mapping operators with human-perceived reality," Journal of Electronic Imaging **16** (2007).
31. J. Kuang, H. Yamaguchi, C. Liu, G. Johnson, and M. Fairchild, "Evaluating hdr rendering algorithms," ACM Transactions on Applied Perception **4** (2007).
32. J. Kuang, G. Johnson, and M. Fairchild, "icam06: A refined image appearance model for hdr image rendering," Journal of Visual Communication And Image Representation **18**, 404–414 (2007).
33. A. Akyüz, R. Fleming, B. Riecke, E. Reinhard, and H. Bülthoff, "Do hdr displays support ldr content? a psychophysical evaluation," ACM Transactions on Graphics **26** (2007).
34. M. Cadík, M. Wimmer, L. Neumann, and A. Artusi, "Evaluation of hdr tone mapping methods using essential perceptual attributes," Computers & Graphics **32**, 330–349 (2008).
35. J. Tumblin and H. Rushmeier, "Tone reproduction for realistic images," IEEE Computer Graphics and Applications **13**, 42–48 (1993).
36. G. Ward, *A contrast-based scalefactor for luminance display* (Academic Press Professional, Inc, San Diego, CA, 1994), pp. 415–421.
37. E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," ACM Transactions on Graphics **21**, 267–276 (2002).
38. N. Miller, N. P.Y., and M. D.D., "The application of computer graphics in lighting design," Journal of the Illuminating Engineering Society **14**, 6–26 (1984).
39. J. Ferwerda, S. Pattanaik, P. Shirley, and D. Greenberg, "A model of visual adaptation for realistic image synthesis," in "Proceedings of ACM SIGGRAPH," (ACM Press, 1996), pp. 249–258.
40. M. Ashikhmin, "A tone mapping algorithm for high contrast images," in "13th Eurographics Workshop on Rendering," (2002).
41. F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high dynamic-range images," in "Proceedings of ACM SIGGRAPH," (ACM Press, 2002), pp. 257–266.
42. R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," in "Proceedings of ACM SIGGRAPH," (ACM Press, 2002), pp. 249–256.
43. F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive logarithmic mapping for displaying high contrast scenes," in "Proceedings of Eurographics," , vol. 22 (2003), vol. 22.
44. Y. Li, L. Sharan, and E. Adelson, "Compressing and companding high dynamic range images with subband architectures," ACM Transactions on Graphics **24**, 836–844 (2005).
45. E. Reinhard and K. Devlin, "Dynamic range reduction inspired by photoreceptor physiology," IEEE Transactions on Visualization and Computer Graphics **11**, 13–24 (2005).
46. T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion," in "15th Pacific Conference on Computer Graphics and Applications," (2007), pp. 382–390.
47. L. Meylan, D. Alleysson, and S. Süsstrunk, "Model of retinal local adaptation for the tone mapping color filter array images," Journal of the Optical Society of America A **24** (2007).
48. S. Ferradans, M. Bertalmío, E. Provenzi, and V. Caselles, "An analysis of visual adaptation and contrast perception for tone mapping," IEEE Transactions on Pattern Analysis and Machine Intelligence **33** (2011).
49. X. Otazu, "Perceptual tone-mapping operator based on multiresolution contrast decomposition," Perception 41 ECVP Abstract Supplement p. 86 (2012).
50. A. Gilchrist, C. Kossyfidis, F. Bonato, T. Agostini, J. Cataliotti, X. Li, B. Spehar, V. Annan, and E. Economou, "An anchoring theory of lightness perception," Psychological Review **106**, 795–834 (1999).
51. J. McCann, C. Parraman, and A. Rizzi, "Reflectance, illumination, and appearance in color constancy," Frontiers in Psychology **5**, 1–17 (2014).
52. "Camera calibration methods," www.cvc.uab.es/color_calibration/CameraCal2.htm. Accessed: 2018-02-06.
53. F. Banterle, A. Artusi, K. Debattista, and A. Chalmers, *Advanced High Dynamic Range Imaging: Theory And Practice* (AK Peters. CRC Press, 2011).
54. M. Kendall and B. Babington-Smith, "On the method of paired comparisons," Biometrika **31**, 324–345 (1940).
55. E. Montage, "Louis leon thurstone in monte carlo: Creating error bars for the method of paired comparison," in "Proceedings of SPIE-IS&T Electronic Imaging," , vol. 5294 (2004), vol. 5294, pp. 222–230.
56. M. Fairchild and G. Johnson, "Rendering hdr images," in "11th Color Imaging Conference," (IS&T/SID, 2000), pp. 108–111.
57. P. Choudhury and J. Tumblin, "The trilateral filter for high contrast images and meshes," in "Proceedings of the Eurographics Symposium on Rendering," (2003), pp. 186–196.