Psychovisual Assessment of Tone-Mapping Operators for Global Appearance and Colour Reproduction

Villa Céline; Labayrade Raphaël; Université de Lyon, Lyon, F-69003, France; Ecole Nationale des Travaux Publics de l'Etat, CNRS, FRE 3237, Département Génie Civil et Bâtiment, Vaulx-en-Velin, F-69120, France

Abstract

To identify people's preference, psychovisual tests are carried out in virtual environment. Images of scenes used for psychovisual tests are natively High Dynamic Range (HDR) images. However, in order to allow a Low Dynamic Range (LDR) display device to project HDR pictures, their dynamic range must be compressed with a tone-mapping operator (TMO). Thus, before tests are carried out, selection of the most suitable TMO, perceptually speaking, is required to present images faithfully, depending on the aim of the application.

In this paper, three different experimental protocols are proposed for assessing the applications of a display device used for psychovisual tests and selecting the most suitable TMO from six candidates. An additional goal is to give a first idea of the applications that can be addressed by this device. Two subjective experiments were conducted and are presented in the paper. The first one is divided in two steps. The results of three different protocols (two protocols in the first test and the third in the second test) for the identification of the preferred TMO for five different scenes are compared and LDR image defects noticed by observers are highlighted in the paper.

Introduction

There are several fields in which psychovisual tests are carried out in order to identify people preference. To avoid real-world tests which are expensive and difficult to implement, psychovisual experiments can be performed in virtual environment for additional flexibility. The critical point here is to ensure that preferred solutions obtained in the virtual environment are also preferred in the real environment.

In the Building Science Laboratory, a virtual reality simulator is employed. With this device it may be possible to conduct experiments regarding colour calibration, luminance calibration, visual comfort, colour preference, etc. Currently work is in progress in order to identify the different types of experiments that can be performed with the simulator. During this explorative stage, it is important to point out possible defects and the applications that can be affected.

The simulator displays Low Dynamic Range (LDR) images in two or three dimensions. However, the dynamic range of the simulator is lower than the actual luminance range of the images to be displayed: photographs or synthetic images computed with light transport simulation programs (as V-Ray[1], Mental Ray[2] or Maxwell Render[3]), feature High Dynamic Range (HDR) luminances. In order to allow a LDR display device to project HDR pictures, their dynamic should be compressed with a Tone-Mapping Operator (TMO). TMO's compress dynamic range to present images that try to faithfully reproduce the captured scene. There are many TMO's available. Thus, in order to present images to the simulator that most faithfully represent reality, selection of the most suitable TMO, perceptually speaking, is required. However, this selection may vary depending on the aim of the application.

Two subjective experimental tests were carried out for that purpose. The first test was divided in two steps related to two different experimental protocols. Initially observers were in front of the real-world scene (Multi-stimuli Rating protocol (MuRt)) and secondly in front of the simulator screen (Multi-stimuli Ranking protocol (MuRk)). In the second experiment, a third experimental protocol is used in front of the real-world scene (Mono-stimuli Rating protocol (MoRt)). During those experiments, observers were asked to judge fidelity of the images to the real-world scene.

Thus, in this paper, three different experimental protocols are presented for assessing the application of display device used for psychovisual tests and selecting the most suitable TMO. An additional goal is to give a first idea of the applications that can be addressed by the simulator. Moreover, in this paper, the results of the three different protocols for the identification of the most suitable TMO are compared and LDR image defects noticed by observers are highlighted.

The paper is presented as follows. Section 1 is an overview of the TMO algorithms selected for the assessment. Then, previous work is summarised in section 2. Section 3 deals with HDR image capture and white balance selection. Section 4 and 5 are respectively dedicated to the presentation of the first and the second test conducted. The first two different protocols are tested during the first experiment, and the third one during the second experiment. Then, comparison of results obtained with the three different experimental protocols is detailed in section 6. Finally, conclusions and future work are presented in section 7.

Tone-Mapping Operators overview

In order to display HDR images with LDR display devices, the luminance range must be reduced with a tone-mapping operator. Two kinds of algorithms were proposed in the literature: global and local operators. Global operators work uniformly on the whole image, the same processing is applied to all the pixels. On the contrary, local operators adapt their action for each pixel taking into account its surroundings.

[1] www.chaosgroup.com

[2] www.mentalimages.com

[3] www.maxwellrender.com

Two global operators [1,5] and four local operators [2,3,4,6] will be assessed in our study. **Drago**'s operator [1] is based on a logarithmic compression function applied to every pixel. The tone-mapped average image luminosity can be adjusted through the so-called "bias parameter" which defines the logarithmic basis. Reinhard et al. suggested, in 2004 [2], an operator based on the adaptation of human photoreceptor cells to the luminous intensity. The compression is performed taking into account pixel intensity and the average intensity of the whole scene. For the remainder of the paper, this TMO will be denoted as **Reinhard 04**. **Durand** et al. proposed, in [3], a local operator using a bilateral filtering. The image is decomposed in a base layer, where the main information (such as strong contrast) is preserved, and a detail layer. Briefly, the base layer is compressed, and then the two layers are merged to obtain the final LDR image. **Fattal** et al. introduced the "gradient domain compression" operator in [4]. The algorithm computes the gradient field of the luminance. During the compression, the magnitude of large gradients is attenuated more than lower gradients. Then, a Poisson equation on the compressed gradient field is solved to obtain the tone-mapped image. An operator derived from photographic tone reproduction was introduced by Reinhard et al. in 2002 in [5]. The algorithm is based on the so-called "dodging and burning" method. The image is divided in zones depending on their luminosity and each zone is compressed independently. For the remainder of the paper, this TMO will be denoted as **Reinhard 02**. Finally, **Ashikhmin** describes in [6] a local operator that takes into account the characteristics of the Human Visual System (HVS). The algorithm is based on the linear approximation of the HVS Threshold vs Intensity function.

Previous work

In recent years, psychovisual studies have already been carried out to compare tone-mapping operator performances [7-11]. Researchers conducted different subjective experiments in order to assess the suitability of a TMO to reproduce HDR image content. The judgment protocol and evaluation method were specific to each study. Moreover, in some cases questioning was about observer's preference and in another cases it was about image fidelity to the reality.

Firstly, the use of a real-world reference has been discussed by some authors. Ashikhmin and Goyal, in 2006 [7], performed comparisons of two experimental protocols, with and without real-world reference. Results obtained were different according to which protocol was used. Thus, they conclude TMO comparisons tests should be carried out with a real-world scene reference. On the contrary, Cadik et al. [8] conducted similar experiments in 2006, but did not notice significant difference between the two populations of results. However, observers were asked to judge their global preference in Cadik et al.'s experiment, whereas in Ashikhmin and Goyal's work, they had to express their feeling about the fidelity of tone-mapped images with respect to reality. In light of this previous work and its alignment with the objective of the present paper, it was decided to perform subjective tests with real-world scene reference.

Secondly, different evaluation methods can be employed to judge images: ranking-based experiment, rating-based experiment on a predefined scale or paired-comparison experiment. During subjective tests, images can be judged one after another, two by two or all together. Cadik et al. [8] carried out two tests in which 14 TMO's were compared. In the first test, participants judged tone-mapped images on their similarity to the real-world scene; they were asked to rate them, one after another, on a scale from 1 to 10. In the second test, the ranking-based procedure was used and observers had not viewed the real-world scene, but rather were required to imagine it. According to Cadik et al., the methodology employed had only minor influence on the results. Kuang et al. [9], in 2004, presented a strong correlation between results obtained with pairwise comparisons and ratings. However, no comparison between the same evaluation method and different experimental protocols has ever been carried out. Such a comparison will be made in this paper. Two different rating schemes will be employed during tests with the same questioning and the same tone-mapped images.

In previous studies, tone-mapped images were displayed on computer screen (either LCD or CRT). Cadik et al.'s experiment [8] is the only one where participants had to rank tone-mapped images printed on photographic paper. This last method is potentially more convenient to compare a large number of images all together.

Finally, in previous work, researchers speculated about which questioning is the best to evaluate image characteristics and consequently the performance of TMO's. Cadik et al. [8] defined five image attributes on which observers had to judge the tone-mapped images: luminosity, contrast rendering, colour rendering, details and artefacts. According to his statistical analysis, attributes interact with each other and cannot be evaluated separately. Kuang et al. [9] also studied the value of judging images with their attributes but he asked observers to judge different images attributes to Cadik. According to their correlations studies, a single attribute is enough to judge an image. Yoshida et al. [10] in 2005 disagreed. He concluded it is not possible to compare realism of LDR images by judging only one attribute. In the present paper, the study is directed towards a global image judgment that includes implicit colour judgment. However, in order to highlight attributes that observers are paying attention to, they will be asked to make comment on tone-mapped images defects.

Experimental design

In order to compare the psychovisual performance of six TMO's, two experiments have been conducted with two different observer panels of 49 and 50 persons. The first test was conducted in two steps. First, observers were taken in front of the real-world scene and asked to compare all printed tone-mapped images to the real-world scene (MuRt). The comments expressed by the observers were recorded. Then, in the second step, the same tone-mapped images were displayed on the simulator-screen and compared by observers to their memory of the real scene (MuRk). The second test was similar to the first step of the first test. Observers were taken in front of the real-world scene asked to compare tone-mapped images. Then unlike the first experiment, images were judged one after another (MoRt). Thus, these two experiments will allow comparisons of three experimental protocols (multi-stimuli rating (MuRt), multi-stimuli ranking (MuRk) and mono-stimuli rating (MoRt)) and of two reproduction techniques (printed images and displayed image with the simulator).

In [9] and [11], it was recommended to use various scenes to determine the weaknesses and strengths of each TMO. Thus, five different scenes were assessed: one corridor, one classroom and one bathroom lit by artificial lighting and an office room lit by daylighting that was presented in two configurations; with blinds open and closed. It should be noted in Table 1, the dynamic range was higher in the natural lighting scene than in the other scenes.

The tone-mapped images were obtained from photographs taken with a Nixon Coolpix 5400 camera set on a tripod using manual settings[4]. As can be seen in Figure 1, the colour aspect of the image highly depends on the white balance used. In order to select a suitable white balance for each scene, preliminary tests were conducted with three observers.

Firstly, they compared photographs taken with eight different white balance settings. After that, two white balance settings were selected for each scene. Secondly, the same observers compared LDR images tone-mapped with Reinhard 02 corresponding to one of the two white balance settings selected before. Finally, fluorescent white balance was chosen for artificial lighting scenes. Depending on the meteorological conditions, sunny or cloudy white balances were chosen for natural lighting scenes.

For each scene, we measured the maximum and the minimum luminance with a luminance-meter from the point of view of the camera position. The camera has been previously calibrated by Dumortier [12]. Exposure Value (EV), corresponding to settings of exposure time and relative aperture, are recommended in a calibration table. From the calibration, the EV and then the calibrated settings (Exposure time and aperture) were deduced. The previous three observers also compared LDR images obtained from different series of digital images (obtained with different settings of the camera). Digital images settings finally used are summarised in Table 2.

Then, HDR software was used to create the HDR image. In order to select the most appropriate software of the available four, we measured 15 points of luminance in the bathroom to cover the entire range with the luminance-meter. Then, four different programs (Photomatix[5], Picturenaut[6], Qtpfsgui[7] and HDR Shop[8]) were employed to obtain a HDR image. Pixel luminance values of the same points were read on each HDR image. Pixel luminance should be related to the measured luminance in real scene by a constant scale factor. Thus, by comparing the ratio between the measured luminance and the pixel luminance, Photomatix was selected.

Scene	White Balance	Luminance (cd/m²)	
		Maxi	Mini
Classroom	Fluorescent	255,6	3,6
Bathroom	Fluorescent	6400	0,3
Corridor	Fluorescent	5610	1,4
Office room – Blind open	Sunny/Cloudy	5000-13000	6-9
Office room – Blind closed	Sunny/Cloudy	5000-12000	4-6

Table 1: *Luminance dynamic range of each scene and white balance used for HDR image generation*

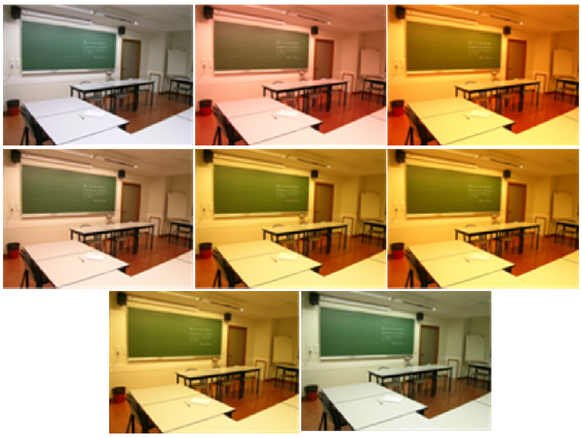

Figure 1: *Classroom scene photographs with eight difference white balance settings*

		Class	Bath	Corridor	Office Open	Office Closed
1	EV	**4,6**	**3**	**2**	**6**	**4**
	Exp.T	1/2	2	2	1/2	1
	F	3,5	4	2,8	5,6	4
2	EV	**7**	**4,6**	**6**	**7,9**	**6**
	Exp.T	1/4	1/2	1/2	1/15	1/2
	F	5,6	3,5	5,6	4	5,6
3	EV	**8,9**	**7**	**7,9**	**9,9**	**7,9**
	Exp.T	1/30	1/4	1/15	1/60	1/15
	F	4	5 ,6	4	4	4
4	EV	**11**	**8,9**	**9,9**	**11,9**	**9,9**
	Exp.T	1/125	1/30	1/60	1/125	1/60
	F	4	4	4	5,6	4
5	EV	**12,9**	**11**	**11,9**	**13,9**	**11,9**
	Exp.T	1/250	1/125	1/125	1/250	1/125
	F	5,6	4	5,6	7,9	5,6
6	EV		**12,9**	**13,9**	**15,9**	**13,9**
	Exp.T		1/250	1/250	1/1000	1/250
	F		5,6	7,9	7,9	7,9
7	EV		**14,9**	**15,9**	**17,9**	**15,9**
	Exp.T		1/500	1/1000	1/4000	1/1000
	F		7,9	7,9	7,9	7,9

Table 2: *Photographs settings*

[4] Settings of camera : sensibility 100 ISO, normal lens, Normal image correction, Normal saturation, normal image quality, image size 2592x1944

[5] www.hdrsoft.com

[6] www.hdrlabs.com/picturenaut

[7] qtpfsgui.sourceforge.net

[8] www.hdrshop.com

Once HDR images had been obtained with Photomatix using suitable white balance and image settings, they were tone-mapped with six different TMO's. Then, the LDR images obtained were printed by a professional photographer with a digital printer Minilab Fujifilm Frontier 550 on glossy silver photographic paper. To allow good visual appraisal of the images to make easier judgments, all the images were printed on 15x21 cm² paper in 300 dpi (size and printing resolution matched the tone-mapped images resolution). Moreover, images were adhered onto black drawing 160g/m² paper jutting out 2 cm from each side to avoid spoiling the photographs and make handling easier. With the images production process discussed previously reproducible results were able to be achieved in producing images.

Experience 1

Panel

49 observers from ENTPE were involved in the first experiment, including 30 students and 19 persons from the staff (either from administration or research labs). All the participants were aged between 20 and 60 years old with 47 % of the participants wore glasses. All the participants could be regarded as inexperienced on concepts relating to tone-mapping and subjective perceptual studies.

Subjective test protocol

Six tone-mapping operators (TMO): Ashikhmin, Drago, Durand, Fattal, Reinhard 02, Reinhard 04, were assessed by the 49 observers to identify the most suitable operator for visual comfort tests and, implicitly, for colour reproduction. The five scenes, described in section 3, were judged for assessment. For illustration purposes, Figure 2 shows tone-mapped images of the bathroom scene judged during the tests. The detail of the subjective test protocol is discussed as follows.

Pilot test

A pilot test was conducted prior to the experiment itself explaining to participants how to answer questions. Then, the subjective experiment was conducted in two steps.

First step: Multi-stimuli rating protocol

During the first step, participants compared printed tone-mapped images to real-world scenes in front of them. They judged each image after having observed all of them. To rate the tone-mapped images, each participant had to place printed images on a 1,15 m long scale which ranged from "not at all" to "extremely", in answering the following question: Does the image represent the real-world scene? The rating scale was a black line drawn on a 80x120 cm² table. The line was not graduated, except at 5 cm from the beginning (resp. end) of the line corresponding to the term "not at all" (resp. "extremely"). The rating scale and the experimental protocol are shown in Figure 3. Taking into account the width of the photographs, the range of available score is between 0 and 107.

It should be noted that this question is not explicitly dedicated to colour reproduction, but it is included. Observers were not given too precise clues about the purpose of the study in order to help them express their spontaneous feelings, and to collect comments about the defects of the images afterwards (described below).

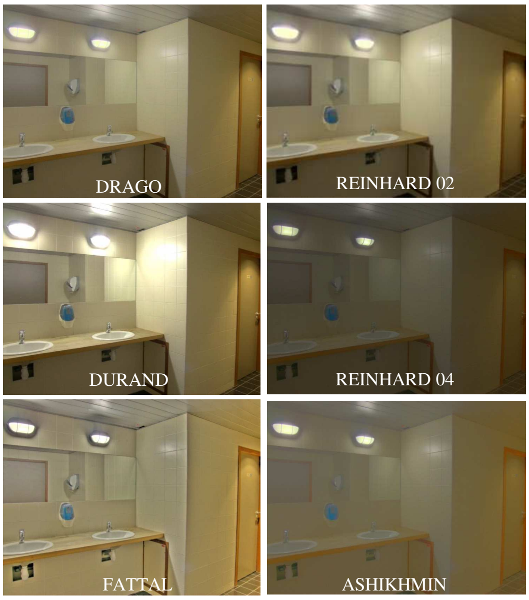

***Figure 2**: The six tone-mapped images of the bathroom scene judged during the subjective test – First test (for both first and second step)*

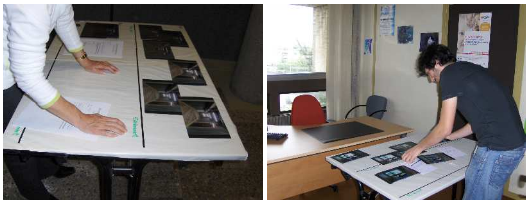

***Figure 3**: Photographs of subjective test – First test – First step*

	Average Illuminance (lux)	Uniformity	Lamp	Color Temp. (K)
Class.	643	0,74	Fluo.	2846
Bath.	39	0,72	CFL	3364
Corridor	158	0,8	Fluo.	3738
Office – Blind open	172-996	0,73	DayL.	n/a
Office – Blind closed	44-226	0,78	DayL.	n/a

***Table 3:** Photometric features on the rating table*

The use of printed photographs was chosen for LDR image presentation to make participation convenient for the observers. This method made the simultaneous judgment of images and the comparison with reality easier, since participants could observe the real-world scene from the same point of view. The photometric features of the table used in the test are summarised in Table 3.

After the judgments were made, participants had to express spontaneous comments on the defects of the two best images they identified by answering the following question: Which are spontaneously the defects in the two best images? In order to avoid bias, the five scenes were presented randomly to each person.

Second step: Multi-stimuli ranking protocol

The second step of the experiment consisted of presenting the same tone-mapped images to the same observers, but displayed on the BARCO Graphics 6300 virtual reality simulator screen and without reference. This step aimed at assessing the influence of both display device and real-world reference. For the five scenes, observers were asked to identify, from memories both the tone-mapped image that most resembled the real-world scene and the one that least resembled. The participant was able to freely observe all the images. Observers sat at 2.63 meters in front of the screen. The size of the images was 1.69x1.26 m² and their resolution was 1024x768. The dynamic range of the simulator was 169.4 cd/m² (ranging from 170,0 – 0,6 cd/m²).

Result of the first step: Rating

Firstly, the average and standard deviations of rating were computed. Drago and Reinhard 02 obtain the best global average respectively with 73.52 and 70.42 on the 0 – 107 scale. For each scene, the results are similar. Images tone-mapped with Ashikhmin obtain the lowest score whereas for Fattal and Reinhard 04 the performance changed depending on the scene and feature high standard deviations. Thus, participants did not share the same opinion about the similarity of these images to the reality. Finally, tone-mapped images using Durand obtain an average score of the same order of magnitude, between 45 and 55, for different scenes, except for the classroom which obtained a lower score.

Then, a Kruskal-Wallis test and a multiple average comparison test were performed from the collected data. According to the Kruskal-Wallis test, there is a statistically significant difference between TMO's and four groups of TMO's were identified. Table 4 presents the groups obtained. Group A represents the most preferred TMO's and group D the least. The test results indicate, whatever the scene, tone-mapped images with Drago and Reinhard 02 operators are perceptually the closest to the real-world reference with respect to the other TMO's judged here.

Results of the second step

In the second step of the subjective study, participants were asked to select which displayed tone-mapped images were the most and least similar to the real-world reference. In order to perform the statistical analysis, rank 6 (resp. 1) was assigned to the image chosen as the most (resp. the least) similar.

Firstly, for a given TMO, the ratio of the number of times that the tone-mapped image is chosen as the most similar to the reality, over the total number of observations was computed. The ratios, expressed as a percentage, are recorded in Table 5. Two groups appear: Drago, Reinhard 02 and Durand obtain the best results with a ratio higher than 88 % whereas the other TMO's obtain lower results (ratio is lower than 28%).

Secondly, in order to analyse judgments, the Kruskal-Wallis test and multiple average comparisons were employed. The results for the data obtained when all scenes are taken together, are presented in Table 6. Drago, Reinhard 02 and Durand are together in the best class (A). Based on the number of observations used in the Kruskal-Wallis test, the results obtained for Reinhard 02 and Ashikhmin are the most reliable.

	DRA	R02	DUR	R04	FAT	ASH
Global	A	A	B	C	C	D
Bathroom	A	A	A	C	B	C
Corridor	A	A	A-B	C	B	C
Classroom	A	A	B-C	B	D	C
Office room – Blind open	A	A-B	B-C	B-C-D	D	C-D
Office room – Blind close	A	A	A-B	B-C	D	C-D

***Table 4**: Multiple average comparison test results – First test – First step*

TMO	The most similar to the reality (in %)
Drago	98.2
Reinhard 02	92.7
Durand	88.9
Fattal	27.9
Reinhard 04	25
Ashikhmin	0.7

***Table 5**: Number of observations and multiple average comparisons test results – First test – Second step*

	Observations number	Classes
Drago	55	A
Reinhard 02	100	A
Durand	70	A
Fattal	83	B
Reinhard 04	45	B-C
Ashikhmin	136	C

***Table 6**: Number of observations and multiple average comparisons test results – First test – Second step*

Results of the first step: Defects

After rating images, participants were asked to express their feeling about image defects. Various categories of defects could be identified (listed below) and are detailed in Table 7:

- Luminosity and contrast: too dark/light;
- Inappropriate colours;
- Details : too much/ not enough accuracy, bad representation of textures, too much/not enough relief;
- Glare: too much/not enough;
- Artifacts (halos, artificial image, too much/not enough sharpness)

Colours reproduction was inappropriate if participants judged colours "too strong" or "not strong enough". They were judged "too bright", "too warm" or on the contrary "too drab". Comments on colours are related to different image features. Participants reported global (for the entire scene) or local (a part of the scene) colours defects linked to light colours or object colours. Moreover, several images presented "uniform" and "pastel" colours linked with not enough contrast. In addition, relief problems could be associated to both colour and luminosity defects. As can be seen in Table 7, the darkness of an image and inappropriate colours are the defects the most cited.

		ASH	R02	DUR	DRA	R04	FAT
Luminosity	Too dark	2	35	19	27	8	5
	Too light		6	3	6		
Inappropriate colours		6	39	11	53	12	8
Glare			14	9	12		1
Contrast	Too much		2		1		10
	Not enough		8	6	12	2	
Reflects		1	5	2	7	2	3
Sharpness			8	4	18	3	5
Details			10	7	13		1
Halos			1	3	3		
Artificial		1	1		4	1	4
Total		10	129	64	156	28	37

Table 7*: Number of comments on defects recorded – First test – First step*

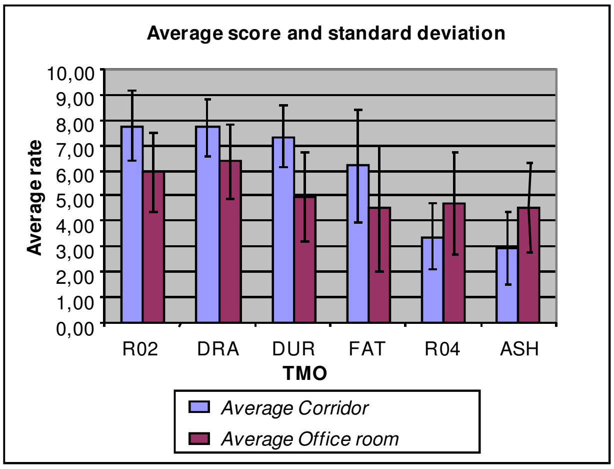

Figure 4*: Average score and standard deviation – Second test*

	DRA	R02	DUR	FAT	R04	ASH
Global	A	A-B	B-C	C	D	D
Corridor	A	A	A-B	B	C	C
Office room	A	A-B	B-C	C	C	C

Table 8*: Multiple average comparison test results – Second test*

Experience 2

Panel

The second experiment involved 50 observers from ENTPE: 22 students and 25 staff (16 from administration, 9 from research labs). Of those involved 42% had taken part in the first perceptual study. Moreover, 84% of participants could be regarded as inexperienced on concepts relating to tone-mapping and subjective perceptual studies and 48 % of the participants wore glasses.

Subjective test protocol: Mono-stimuli rating

The same TMO’s were assessed by observers in order to identify the most suitable operator for preference tests, including luminance field and colour reproduction. Two scenes already used in the first experiment were judged: the corridor lit by artificial lighting and the office room lit by daylighting presented with blinds open.

Participants compared printed tone-mapped images to real-world scenes in front of them. They judged each image one after another and were not allowed to see again those already rated. The observers were asked to answer the question: “Does the image represent the real-world scene?” by rating the image on a 1 (“not at all”) to 10 (“extremely”) integer scale. In order to avoid any bias, the two scenes and the images were presented randomly to each person. The test used the same printed images as in the first experiment.

Results of the second experiment

Firstly, the averages and standard deviations were computed. The results are shown Figure 4. Drago and Reinhard 02 obtain the best average scores: 7,68/10 and 7,76/10 (resp. 6,34/10 and 5,94/10) for the corridor scene (resp. for the office room). Average scores vary more between TMO’s for the corridor scene than for the office room. Similarly the standard deviations are higher for the office room.

Then, a Kruskal-Wallis test and a multiple average comparison test were performed from the collected data. According to the Kruskal-Wallis test, there is a statistically significant difference between TMO rating data. From the multiple average comparison test, four groups were identified. Each group brings together TMO’s with no statistically significant difference in performance. Table 8 shows the results for each scene and globally (the two scenes together). The test results indicate tone-mapped images with Drago and Reinhard 02 operators are perceptually the closest to the five real-world scenes studied in this work.

Protocol comparison

In order to compare results obtained with the three different experimental protocols (MuRt, MuRk, MoRt), the Kruskal-Wallis test was employed. Conclusions are as follows.

Firstly, there is no statistically significant difference between data obtained in both steps of the first experiment, i.e. in front of the real-world scene and with the images displayed on the simulator-screen. Thus, in this work, the perceived performances of TMO’s do not depend on the reproduction technique (either printed images or displayed on the simulator).

Secondly, during the first step of the first experiment and during the second experiment, participants judged images by rating them. However, the image presentation process and the rating scale were different. The average scores were compared. The scores obtained between 0 and 107 during the first experiment were scaled to lie between 1 and 10. As shown in Figure 5, TMO ranking orders are close in both tests. Except for the Drago image of the office room, average values are smaller for the first test where participants were more severe in their judgments. This discrepancy is probably due to the difference in rating scale. In the first test, observers saw the continuous drawn scale whereas in the second test observers gave discrete marks.

Moreover, in the first case, participants were alone whereas in the second case, the test organizer was with them.

The two datasets were subjected to a Kruskal-Wallis test in order to highlight similarities and differences in the results. Kruskal-Wallis test conclusions are shown in Table 9 for each scene common in both tests. There is no significant difference between marks obtained by images tone-mapped with Reinhard 02 for both scenes and those obtained by Drago images for the corridor scene. Therefore, the previous findings, section 4 and 5, on performance of Reinhard 02 and Drago for the corridor are confirmed.

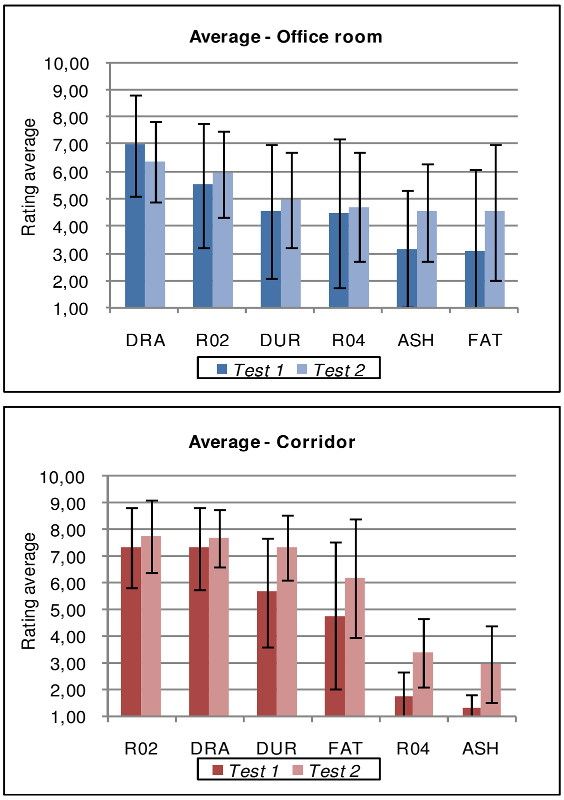

***Figure 5**: Average values. Comparisons of First test – First step and Second test*

Office room						
	DRA	R02	DUR	FAT	R04	ASH
Observed value	4,494	0,454	1,028	7,899	0,731	15,40
Theoretical value	3,841	3,841	3,841	3,841	3,841	3,841
Significant difference	yes	no	no	yes	no	yes
Corridor						
	DRA	R02	DUR	FAT	R04	ASH
Observed value	1,069	2,963	17,62	6,533	31,94	30,50
Theoretical value	3,841	3,841	3,841	3,841	3,841	3,841
Significant difference	no	no	yes	yes	yes	yes

***Table 9**: Kruskal-Wallis test results. Comparison of First test – First step and Second test*

Conclusions

In this paper, six tone-mapping operators were compared through two subjective experiments and three experimental protocols (Multi-stimuli Rating, Multi-stimuli Ranking, Mono-stimuli Rating). Observers first compared printed tone-mapped images to the real-world scenes. One panel judged all printed images together, placing them on a continuous scale and identified defects, whereas a second panel judged each image one after another giving a score on a discrete scale. Then, observers compared the same images displayed on the simulator screen. During this step, observers had seen the real-world scene before and were required to judge images from memory.

According to the two subjective experiments conducted in this paper, Reinhard 02 and Drago are the tone-mapping operators that obtain LDR images which most represent reality (see images Figure 6) with respect to the others TMO's of the five tested scenes. This result was demonstrated with the three different experimental protocols. According to subjective judgment, Ashikhmin's images are the least similar to the five real-world scenes. Results from others TMO's vary depending on the scene and the experimental protocol. However, despite the experimental protocol used, the ranking order is kept the same. Moreover, the results obtained with the two different reproduction techniques lead to similar findings on the preferred TMO's. This is an argument that the perceived performances of TMO's does not depend on the reproduction technique (either printed images or virtual reality simulator display device).

Our findings on the good performance of Reinhard 02 agree with most previous work [8-11]. The same results for Ashikhmin's operator were obtained in [8]. However, some results disagree with other studies. Drago's operator performance does not agree with that of Ledda's [11] with respect to the obtained results of image fidelity of the reality.

However, it is not that easy to compare with other studies. Each image obtained in this work is subjected to the same transforms, the only difference being the application of the different tone-mapping operators. Thus, our results relate to the employed camera, to its settings and calibration, to all regarding printing, and to all regarding the display device. Moreover, our results are specific to the five scenes tested in this study.

Before conducting our experiments, preliminary tests were organised in order to capture HDR images, from which tone-mapped images were obtained. White balance was selected for each scene: "fluorescent" for scenes lit by artificial lighting and "daylight" for the office room. Moreover, different series of digital images were captured and tested to determine the most appropriate series to use for the image production. The preliminary tests were conducted with a small observer panel and images mostly tone-mapped with Reinhard 02 were compared. This may introduce some bias into our experiments.

According to the comments into images, image darkness and the presence of inappropriate colours are the defects most identified by observers. In future work, in order to improve tone-mapped images, it is planed to vary parameters of the TMO. Moreover, according to [9,10], paired-comparison experiments obtain robust results and are easier for observers to judge than other evaluation methods. Thus, Thurstone's judgment will be used in future tests and results will be compared to the one obtained here.

According to our results, psychovisual tests related to luminance field and lighting configuration preference can be addressed by the simulator. However these applications require an accurate calibration of the device. The suitability of the simulator for colorimetric applications is more uncertain. Further investigations are needed in order to identify the potential advantages and limitations in using this technology.

References

[1] F. Drago, K. Myszkowski, T. Annen & N. Chiba, Adaptive logarithmic mapping for displaying high contrast scenes. Eurographics 2003, Computer Graphics Forum, 22, 3, pg 419-426. (2003).

[2] E. Reinhard & K. Devlin, Dynamic range reproduction inspired by photoreceptor physiology. IEEE Transactions on visualization and computers graphics, 12pg. (2004).

[3] F. Durand & J. Dorsey, Fast bilateral filtering for the display of High-Dynamic-Range images, ACM SIGGRAPH 2002, Proc. Computer Graphics, Proc. Annual Conference, pg. 257-266. (2002).

[4] R. Fattal, D. Lischinski & M. Werman, Gradient domain high dynamic range compression. 29th Proc. Computer Graphics and Interactive Techniques, Proc. Annual Conference, ACM Press, pg. 249-256. (2002).

[5] E. Reinhard, M. Stark, P. Shirley & J. Ferwerda, Photographic tone reproduction for digital images. ACM Transactions on Graphics, 21, 3, pg .267-276. (2002).

[6] M. Ashikhmin, A tone mapping algorithm for high contrast images. Eurographics Workshop on Rendering, pg. 1-11. (2002).

[7] 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, 30pg. (2008).

[8] M. Cadik, M. Wimmer, L. Neumann & A. Artusi, Evaluation of HDR tone mapping methods using essential perceptual attributes. Computers and Graphics, Elsevier, ISSN 0097- 8493, 32, 3, pg. 330-349. (2008).

[9] J. Kuang, H. Yamaguchi, C. Liu, G.M. Johnson & M.D. Fairchild, Evaluating HDR rendering algorithms. ACM Transactions on Applied Perception, 4, 2, 30pg. (2007).

[10] A. Yoshida A., V. Blanz, K. Myszkowski & H. Seidel, Perceptual evaluation of tone mapping operators with real-world scenes. Stereoscopic Displays and Virtual Reality Systems XII, Proc. SPIE, 5666, pg.192-203. (2005).

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

[12] D. Dumortier, B. Coutelier, T. Faulcon, F. Van Roy, Photolux: a new luminance system based on Nikon Coolpix digital cameras. Lux Europa 2005, Berlin, Germany, pg. 308-311. (2005).

Author Biography

Céline Villa received the Engineer Diploma from the ENTPE and the MSc in Buildings Engineering from the University of Lyon in 2009. She is a PhD student at the DGCB-LASH laboratory since October 2009. Her work focuses on the integration of users' preferences in lighting design and optimization.

Raphaël Labayrade received his Msc in Digital Imaging from the University of Saint Etienne and was graduated from the ENTPE engineer school in 2000. He received the Ph.D. degree in 2004 from the University of Paris 6 – Jussieu From 2000 to 2007, he was in the perception team of the LIVIC (INRETS) department. His main work dealt with artificial vision, for applications such as obstacles detection, road lane recognition, and data fusion. Since then he has worked in the DGCB-LASH (ENTPE). His work has focused on visual appearance, light transport simulation, and multi-criteria optimization of building parameters.

Acknowledgements

The authors would like to thanks test participants and Enora Parent and Michael Hirning for their help.

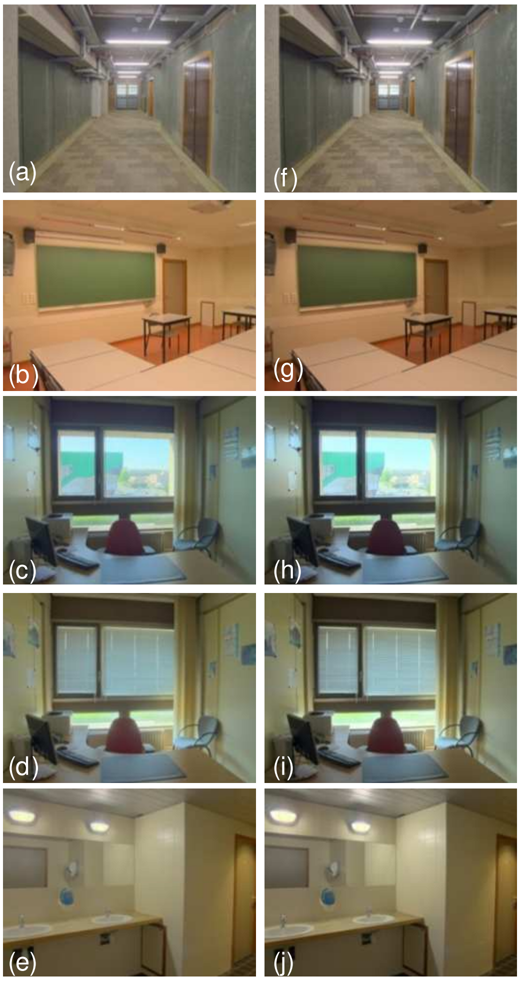

***Figure 6**: Best tone-mapped images – First test – First step LDR images tone-mapped with Drago (a-e) and Reinhard 02 (f-j)*