A SUBJECTIVE EVALUATION OF TONE MAPPING AND EXPOSURE FUSION ALGORITHMS IN STANDARD AND SMALL SCREEN DISPLAY DEVICES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

MUSTAFA LEVENT EKSERT

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2013

Approval of the thesis:

A SUBJECTIVE EVALUATION OF TONE MAPPING AND EXPOSURE FUSION ALGORITHMS IN STANDARD AND SMALL SCREEN DISPLAY DEVICES

submitted by **MUSTAFA LEVENT EKSERT** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences** ________________

Prof. Dr. Adnan Yazıcı
Head of Department, **Computer Engineering** ________________

Assist. Prof. Dr. Ahmet Oğuz Akyüz
Supervisor, **Computer Engineering Department, METU** ________________

Examining Committee Members:

Prof. Dr. Veysi İşler
Computer Engineering Department, METU ________________

Assist. Prof. Dr. Ahmet Oğuz Akyüz
Computer Engineering Department, METU ________________

Assoc. Prof. Dr. Pınar Karagöz
Computer Engineering Department, METU ________________

Assist. Prof. Dr. Sinan Kalkan
Computer Engineering Department, METU ________________

Assist. Prof. Dr. Tolga Çapın
Computer Engineering Department, Bilkent University ________________

Date: ________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: MUSTAFA LEVENT EKSERT

Signature :

ABSTRACT

A SUBJECTIVE EVALUATION OF TONE MAPPING AND EXPOSURE FUSION ALGORITHMS IN STANDARD AND SMALL SCREEN DISPLAY DEVICES

Eksert, Mustafa Levent
M.S., Department of Computer Engineering
Supervisor : Assist. Prof. Dr. Ahmet Oğuz Akyüz

September 2013, 57 pages

Standard display devices are not capable of displaying real world scenes as they are perceived by humans. One of the reasons for this is their limited dynamic range. The field of high dynamic range (HDR) imaging has been developed to alleviate this problem. Among the techniques of HDR imaging, tone mapping operators (TMOs) and exposure fusion algorithms (EFAs) are commonly used to display HDR content on conventional low dynamic range (LDR) monitors. In this thesis, 4 TMOs and 3 EFAs are compared in terms of image quality reproduction on standard and small-screen display devices. The latter is motivated by the fact that small-screen monitors are becoming more widespread due to the advances in mobile devices and digital cameras. The quality reproduction is performed with respect to four criteria namely contrast, color, detail, and similarity. The results show that best TMOs outperform the best EFAs in terms of reproduction of contrast, detail, and similarity. However, EFAs are found to be better in reproduction of colors. Also, the differences between the algorithms are found to be minimized on a small-screen display device.

Keywords: high dynamic range imaging, tone mapping, exposure fusion, evaluation study

ÖZ

STANDART VE KÜÇÜK EKRANLI GÖRÜNTÜ ARAÇLARINDA TON EŞLEME VE POZ FÜZYONU ALGORİTMALARININ ÖZNEL DEĞERLENDİRMESİ

Eksert, Mustafa Levent
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Yrd. Doç. Dr. Ahmet Oğuz Akyüz

Eylül 2013 , 57 sayfa

Standart görüntüleme aygıtları, gerçek bir sahneyi insanlar tarafından algılandığı gibi göstermekte yetersiz kalmaktadır. Bunun sebeplerinden biri de standart görüntüleme cihazlarının sınırlı dinamik aralığıdır. Yüksek dinamik aralık (YDA) alanı bu problemin kısmen giderilmesi amacıyla ortaya atılmıştır. YDA görüntülemenin teknikleri arasında, ton eşleme operatörleri (TEO) ve poz füzyonu algoritmaları (PFA), geleneksel düşük dinamik aralıklı (DDA) monitorlerde YDA içeriğini görüntülerken sıklıkla kullanılır. Bu tezde 4 TEO ve 3 PFA, standart ve küçük ekranlı görüntüleme aygıtlarındaki resim kalitesi bazında karşılaştırılmıştır. Küçük ekranlı görüntüleme cihazları mobil cihazlar ve dijital aygıtların yaygın kullanılıyor olmasından ötürü çalışmaya dahil edilmiştir. Kalite değerlendirmesi kontrast, renk, ayrıntı ve benzerlik adında dört esasa göre yapılmıştır. Sonuçlar göstermiştir ki en iyi TEO'lar, en iyi PFA'lara kontrast, ayrıntı ve benzerlik ediniminde üstünlük sağlamışlardır. Buna karşın, PFA'ların renk ediniminde daha iyi oldukları gözlemlenmiştir. Ayrıca algoritmalar arasındaki farkların küçük ekranlı görüntüleme aygıtlarında azaldığı gözlemlenmiştir.

Anahtar Kelimeler: yüksek dinamik aralıklı görüntüleme, ton eşleme, poz füzyonu, değerlendirme çalışması

To Those Who Make This Possible

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisor Asst. Prof. Ahmet Oğuz Akyüz for his great support and guidance during the preparation of my master thesis. I am also grateful to my family for their endless support. I appreciate Melike Selin Aydın for her contributions to the study and all who participate in the experiments for their kindness and consideration. I am also thankful to MODSIMMER for providing the experiment room.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

HDR	High dynamic range
LDR	Low dynamic range
TM	Tone mapping
EF	Exposure fusion
TMO	Tone mapping operator
EFA	Exposure fusion algorithm
HVS	Human visual system

CHAPTER 1

INTRODUCTION

The real world as well as computer generated artificial scenes contain a vast range of luminances. Although the human eye can perceive the majority of this range, digital cameras as well as display devices fall short in capturing and presenting this high luminance range accurately. The techniques of HDR imaging have been developed to solve various problems that stem from this discrepancy. In this thesis, my goal is to provide a rigorous psychophysical evaluation of two major techniques of high dynamic range (HDR) imaging namely tone mapping (reproduction) operators and exposure fusion algorithms. Both of these methods aim to display a high contrast scene on a conventional display device as effectively as possible. However, no formal evaluation hitherto exists that compares these two groups of techniques against each other. In this chapter, I first provide a discussion of differences of HDR imaging from low dynamic range LDR imaging to clarify the concepts used in this thesis. This is followed by the contributions and the outline of the thesis.

1.1 HDR vs. LDR Imaging

Photography has emerged in the middle of the nineteenth century and since then several improvements have been achieved in the quality of both capturing, processing, and printing [3, 4, 5]. More recently, digital photography supported with improvements in digital storage and display technologies was born. As in the conventional one, digital photography has gone through several improvements in image capture, processing, display, and printing.

One of the improvement in image presentation is spatial resolution. By increasing spatial resolution more local details can be represented and the aliasing effect can be reduced [41]. Another quality improvement in digital imaging is the bit-depth. Higher bit-depth enables more accurate representation of colors and reduces banding artifacts. In digital imaging, typically 8-bits per color channel are used resulting in 2^{24} number of distinct colors. Although this seems like a large number, it is not enough to represent the range of luminances found in real world as well as computer generated scenes. For instance, from starlight to sunlight the real world contains more than 10 log units (orders of magnitude) of dynamic range which is defined as the ratio of the maximum to the minimum luminance. The luminance levels of several typical scenes are given in Table 1.1 for illustration purposes. To deal with these high range of luminances more than 8-bits per color channel are required in both capture, processing, and display of real world scenes.

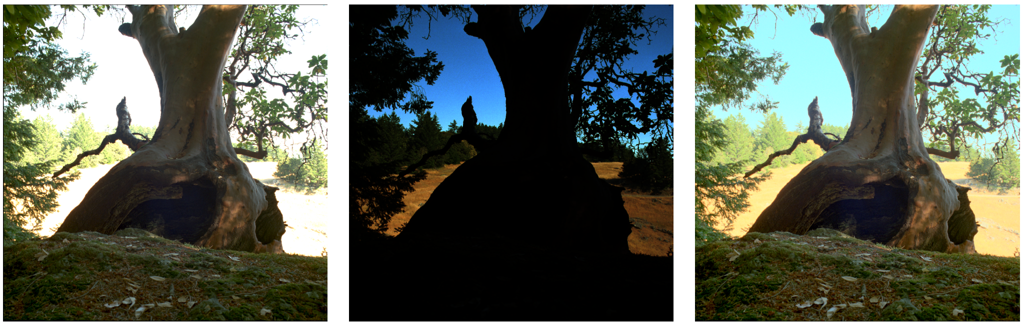
Table 1.1: Ambient luminance for some lighting conditions from [51]

Condition	**luminance ($\mathbf{cd/m^2}$)**
Starlight	10^{-3}
Moonlight	10^{-1}
Indoor lighting	10^{2}
Sunlight	10^{5}

The dynamic range in natural scenes may vary depending on several factors such as scene content and lighting conditions. However, on average, most real world scenes contain at least 4 log units of dynamic range (1:10000) [41]. The human visual system (HVS) is a very advanced and complex system that can cope with this dynamic range. In fact, the HVS is known to be responsive for 5 log units of dynamic range in a given environment and to 10 log units dynamic range over time through visual adaptation [21].

As traditional image capture devices are unable to capture this range directly, several techniques have been developed to capture HDR scenes using these devices (along with developments in hardware that can directly capture HDR data). Typically, multiple shots of the same scene each with a different exposure time are captured which are then combined into a single HDR image. Alternatively, these individual exposures are directly merged to obtain a new LDR image that contains details in both light and dark regions (known as exposure fusion).

Displaying an HDR image on a standard monitor requires further processing. If such an image is displayed by clamping high and low luminances, or by linearly scaling its dynamic range to that of the target display, undesired representations may occur. Using a proper tone mapping operator can yield a much more plausible representation as illustrated in Figure 1.1.



(a) Clamping high and low luminances.

(b) Linearly scaling.

(c) Proper tone mapping.

Figure 1.1: Different presentations of HDR data. 1.1(a) clamping high and low luminances results in detail loss in dark and bright regions, 1.1(b) linearly scaling result is relatively dark, 1.1(c) proper tone mapping result preserves details in shadows and sunny areas © ILM.

There are two mainstream approaches for representing HDR image content on LDR display devices: tone mapping and exposure fusion. In the former, first an HDR image is created from multiple exposures. Then the dynamic range of this HDR image is reduced to make displayable on an LDR monitor. Different tone mapping operators (TMOs) put emphasis on preserving different attributes during this process. For instance, while some operators aim to reproduce all details, others aim to preserve the global contrast in lieu of details. More details on these operators are given in the next chapter. The latter method, exposure fusion, aims to directly create a detail-rich LDR image from a set of bracketed exposures. That is, in this pipeline, the generation of the HDR image is bypassed. For instance, the HDR mode in iPhone 4 and above models utilizes a proprietary algorithm of this type [2]. Again, more information about exposure fusion algorithms are given in the following chapter.

1.2 Contributions

The proposed work contributes to the literature as follows:

- This study presents a rigorous subjective evaluation of several tone mapping and exposure fusion algorithms by means of reproduction of color, contrast, detail, and similarity in two types of display conditions (normal and small screen).
- Tested methods consist of diverse and recent techniques from TM and EF literature which have not been tested in previous evaluation studies as well as well-known and commonly used algorithms.
- For the first time, TMOs and EFAs are compared against each other in a rigorous psychophysical experiment.
- In addition to a normal sized display condition, a small camera LCD screen is used as a presentation medium to evaluate how well each algorithm preserves the similarity between the real world scene and its LDR depiction. This is motivated by the widespread adoption of mobile devices and digital cameras that can capture HDR images.

1.3 Outline

In Chapter 2, TMOs and EFAs which are used in the experiments are explained in detail, some fundamental EFAs and TMOs are presented, and the similar studies in the literature are introduced. In Chapter 3 experimental methodology and the statistical analysis used in the experimental data evaluation are presented. In Chapter 4 the experiment details are introduced and the collected results are analyzed and discussed. In Chapter 5 overall experiment results are discussed. In Chapter 6 the thesis is summarized and the possible future works are discussed. In Appendix A instructions for experiment one are provided. In Appendix B instructions for experiment two are provided.

CHAPTER 2

RELATED WORK

In this chapter, several important tone mapping operators and exposure fusion algorithms are described. More emphasis will be put on the methods that are evaluated in this thesis. Additionally, earlier work on subjective evaluation of tone mapping operators is reviewed.

2.1 Tone Mapping

The tone mapping problem is first introduced by Tumblin and Rushmeier [49]. In that algorithm, the purpose is to match the impression of the real world luminance in the human eye and the effect of the displayed image on a display device. To achieve this, the operator works on two separate models named the observer and the display models. The observer model uses brightness as a function of luminance computed in the log domain. On the other hand, the display model is influenced by the CRT encoding-decoding process taking into account the display parameters such as the maximum display luminance and the screen background luminance. The proposed tone reproduction operator concatenates the observer model and the inverse display model to match brightness of the original scene and the displayed image.

Ward et al. [28] propose a tone mapping operator that carries out contrast compression in both computer generated images such as ray tracing and radiosity results and photographic images. The method aims to preserve visibility and imposes glare, color sensitivity, and visual acuity models for simulating human vision imperfections. The algorithm first applies contrast reduction with a special type of histogram equalization, named histogram adjustment, in order to prevent contrast expansion in highly populated bins of the histogram. The first part of the algorithm also involves a model of the human contrast sensitivity. Then several human visual limitations such as the veiling glare around bright regions are added to improve the realism of the reproduced image. Moreover, color sensitivity based on luminance values and visual acuity, which account for resolution reduction in dark regions, is provided in the final result.

Different from the earlier work that directly use the luminance (or its logarithm), gradient domain techniques have been developed that work on the luminance gradients. A well-known example is the efficient and robust tone mapping operator introduced by Fattal et. al. [20]. The method reduces large gradients while expanding small gradients. The modified gradient map is converted back to the luminance domain by solving a Poisson equation. The tone reproduction process is performed on gradients of 2D logarithmic luminance. An attenuation function obtained with a multiscale decomposition of the gradient image is used during the reduction process. The method provides effective compression by preserving visible details without introducing halo artifacts.

In the remainder of this section, the four tone mapping operators that were used in this study are discussed in greater detail.

2.1.1 Photographic Tone Reproduction for Digital Images

This operator borrows from the well-known Zone system that has been used for decades in conventional photography [3, 4, 5]. For applying the zone system, the photographer first chooses a zone for the middle gray. The photographer then chooses the lightest and darkest region from the scene in order to determine the dynamic range of the scene by means of number of zones. If a scene has 9 zones, the scene can be printed properly, if not, darkest areas are mapped to pure black and the lightest areas are mapped to white; dodging and burning which withholds or adds extra light can be applied in case the clamped detail is important to express in the final print.

In the algorithm, first the log average luminance is computed as:

$$\overline{L}_w = \exp\left(\frac{1}{N}\sum_{x,y} log(\delta + L_w(x,y))\right), \tag{2.1}$$

where $L_w(x,y)$ is the world luminance of a pixel, N is the number of pixels and δ is the term that is used to avoid singularity. Then the scaled luminance is calculated with the following:

$$L(x,y) = \frac{\alpha}{\overline{L}_w} L_w(x,y). \tag{2.2}$$

For a typical normal-key scene, the key value α is set to 0.18 and this value can be set to different values between $[0,1]$.

Modern photography tends to compress high luminance regions because high luminance pixels are not frequent in a normal image data [37, 46]. Therefore, an equation with this characteristics is used in order to compress the luminance values by a factor of $1/L$ for high luminance and 1 for low luminance:

$$L_d = \frac{L(x,y)}{1 + L(x,y)}. \tag{2.3}$$

A different version of the equation which allows for burning beyond a threshold is given below:

$$L_d(x,y) = \frac{L(x,y)\left(1 + \frac{L(x,y)}{1+L_{white}^2}\right)}{1 + L(x,y)}, \tag{2.4}$$

where L_{white} is the lowest luminance value which is mapped to white.

The photographic operator also has a local component which simulates the dodging-and-burning process. However, in this study only the global version of this operator is used.

2.1.2 Compressing and Companding High Dynamic Range Images with Subband Architectures

Global tone mapping algorithms which use nonlinear tone curves [49, 52] or histogram [28] reduce the dynamic range but they are not successful in preserving details. Some of the early tone mapping algorithms address this problem and preserve visual details in image [45], but they cause halo artifacts (gradient reversals). This algorithm proposes a multiscale image processing technique for tone mapping which avoids visual artifacts caused by other multiscale dynamic range compression approaches.

In the method, the original signal is split into subbands, b_i, with different filters f_i. The original signal can be formed by summing all the subbands:

$$s(x) = \sum_i b_i(x). \tag{2.5}$$

A compression can be applied to subbands in order to reduce the dynamic range of the image. Applying a sigmoidal function directly to the subbands can compress the signal but it causes distortion in the synthesized image. Therefore, gain functions which are obtained from a sigmoidal compression are first blurred. The subbands are compressed by using these modified gain maps, G as:

$$b'(x) = b(x)G(x). \tag{2.6}$$

The modified subbands are then synthesized by synthesis filters to reconstruct the compressed and less distorted signal $s'(x)$.

A set of analysis-synthesis models are available for synthesizing and analyzing operations. In this method, nested filtering model with symmetric analysis and synthesis filter banks is used. In this model, 3 high-pass and 1 low-pass 2D zero padded filters namely hi_x, hi_y, hi_{xy}, and lo are used for constructing subbands of the image.

hi filters are directly used for constructing subband signals and *lo* filters are used for further subband signal constructions. The resulting *lo* signal is recursively split into subbands with these four synthesis filter banks with the same 4-way subband decomposition until a certain recursion depth n is reached. 2D filter banks are modified in each successive level, expanded in x and y direction by inserting 0 in the middle of the filter mask, (e.g $f_1 = [1, -1]$ transforms into $[1, 0, -1]$ in 1D filter mask). Therefore, a multiscale subband decomposition which avoids the aliasing problems resulting from the subsampling technique is achieved.

Haar filters are used in subband decomposition since they are easy to implement and produce plausible results with gain maps. After the gain map contrast reduction is carried out to the subbands, the corresponding synthesis filters are applied to the compressed subbands and the synthesized signals are summed in order to reconstruct the modified final image result.

For an n level subband composition model, there exist $3n + 1$ subbands where $B_{3n+1}(x, y)$ is the lowpass residue. The following gives the activity map of the subband:

$$A_i(x, y) = g(\sigma) * |B_i(x, y)|, \tag{2.7}$$

where $g(\sigma)$ is the Gaussian kernel. A nonlinear function $p()$ is applied to activity map in order to derive the gain control:

$$G_i(x, y) = p\left\{A_i(x, y)\right\} = \left(\frac{A_i(x, y) + \varepsilon}{\delta}\right)^{(\gamma-1)}, \tag{2.8}$$

where γ is the compressive factor between 0 and 1, ε is noise level related parameter and δ is the gain control stability level which modifies the activity value by increasing less than itself and decreasing the activity value greater than its value. δ is calculated with the following:

$$\delta_i = \alpha_i \frac{\sum_{(x,y)} (A_i(x, y))}{M \times N}, \tag{2.9}$$

where α_i is spatial frequency related function between 0.1 and 1.0 from the lowest to the highest frequency subband.

After the gain is calculated, the subband is compressed with the following equation:

$$B'_i(x, y) = G_i(x, y) \times B_i(x, y). \tag{2.10}$$

Finally, resulting subbands are convolved with synthesis filters and summed in order to obtain compressed image result.

Since the subbands of the same signal tend to have similar activity, an appropriate gain map can be used for all subbands. Therefore, an aggregated activity map that is calculated by summing up activity maps of all subbands is used instead of separate gain maps for each subband. δ is taken as one tenth of the average of A_{ag} in gain map calculation. This gain map is used to modify the subbands in the following equation:

$$B'_i(x, y) = m_i G_{ag}(x, y) \times B_i(x, y), \tag{2.11}$$

where m_i is a scale-related constant.

2.1.3 Display Adaptive Tone Mapping

In this work, Mantiuk et al. [34] propose a tone mapping operator which reduces contrast according to a human visual system model. This operator also takes into account the visual properties of the display device which is used for displaying the resulting image. Besides, a comparison is performed between the proposed work and the previous algorithms in a small case study.

In order to match the perception of the real scene and displayed image, the operator tries to solve an optimization problem. The optimization process essentially updates the tone mapping parameters in order to minimize the difference between the HVS model response of the original image and the displayed image.

Apart from the optimized parameters, there exist other parameters related to the display properties and viewing conditions. Display device is modeled with the formula below:

$$L_d(L') = (L')^\gamma \cdot (L_{max} - L_{black}) + L_{black} + L_{refl}, \tag{2.12}$$

where L_d is displayed luminance, L' is the pixel value in range $0 - 1$, γ is gamma value of the display (default 2.2), L_{max} and L_{min} are the peak and minimum luminance value of the display and L_{refl} is the ambient light reflected from the display surface which affects the minimum luminance value of the display. L_{refl} is calculated with the following formula:

$$L_{refl} = \frac{k}{\pi} E_{amb}, \tag{2.13}$$

where E_{amb} is the ambient luminance value in lux and k is the reflectivity of the display panel.

The HVS model is based on the factor of visual contrast distortions. A transducer function [53] is used to estimate the response of HVS: $R = T(W, S)$ where S is sensitivity and W is contrast. For a given sensitivity value, the response function has a detection threshold and invisible contrast noises are mapped to below the threshold value. Sensitivity is calculated with the contrast sensitivity function from [14] with the parameters frequency ρ (cycle per degree), adapting luminance L_a (cd/m^2), and viewing distance v_d, respectively:

$$S = CSF(\rho, L_a, v_d). \tag{2.14}$$

Contrast value W is calculated in a multiscale fashion, by taking differences of consecutive Gaussian pyramid levels in log domain. For efficiency, the background luminance and contrast values are separated into smaller blocks, named bins; and for a given background luminance, contrast value, and the level of the Gaussian pyramid, a conditional probability density function is introduced in order to represent similar contrast values. The conditional probability density function is formulated as:

$$c_{i,m,l} = P\Big(m\delta - \frac{\delta}{2} \leq G_l < m\delta + \frac{\delta}{2} \;\Big|\; x_i - \frac{\delta}{2} \leq I_{l+1} < x_i + \frac{\delta}{2}\Big), \tag{2.15}$$

where x_i is the background luminance bin, $\delta = x_{i+1} - x_i$ the distance between consecutive bins and $m = -M, \ldots, -1, 1, \ldots, M$ where $M\delta < 0.7$ which gives an appropriate interval for contrast bins.

Piecewise linear curve which illustrates the correlation between the original image luminance and displayed image luminance is formed with the parameters obtained by the conditional probability density function. x_i and δ are used as the background luminance and luminance interval respectively for the original image. The corresponding display image parameters; on the other hand, are y_i and d_i: background luminance bin for the corresponding x_i value and the distance is $d_i = y_{i+1} - y_i$ for $i = 1 \ldots N - 1$.

An objective function is used to linearize log display luminance vs. log image luminance curve. To achieve this, the difference between display luminance response and image luminance response is minimized by adjusting d_i parameters of the display luminance. The optimization problem is formulated as follows:

$$\arg\min_{d_1\ldots,d_N} = \sum_{l}\sum_{i=1}^{N-1}\sum_{m=-M,m\neq 0}\left[T\left(\sum_{k\in\phi} d_k, S_d\right) - T\left(e\sum_{k\in\phi}\delta, S_r\right)\right]^2 \cdot c_{i,m,l}, \tag{2.16}$$

such that

$$d_i \geq 0 \quad \text{for} \quad i = 1\ldots N-1, \tag{2.17}$$

$$\sum_{i=1}^{N-1} d_i \leq L_d(1) - L_d(0) \quad \text{for} \quad i = 1\ldots N-1, \tag{2.18}$$

where $\phi = i+m,\ldots,i-1$ for $m < 0$ and $\phi = i,\ldots,i+m-1$ for $m \geq 0$. The constant e is the constant enhancement factor used for enhancing contrast of the reference image ($e = 1.15$ by default). T is the transducer function. The minimization function is called iteratively in order to recompute current d values from the previous d vector until it approaches the minimal value in 3-7 iterations.

Finally, d values are used in the final tone curve reproduction equation:

$$y_i = L_d(0) + \sum_{k=1}^{i-1} d_k + \alpha(L_d(1) - L_d(0) - \sum_{k=1}^{N-1} d_k), \tag{2.19}$$

where α is an image brightness factor which adjusts displayed image brightness.

2.1.4 Globally Optimized Linear Windowed Tone Mapping

This paper proposes a recent tone mapping technique which applies local tone mapping operation on the small overlapping virtual windows on the image. The proposed algorithm categorizes transitions of the scene as smooth and sharp transition and imposes a local approach that preserves global property of the image. The proposed algorithm directly processes the image radiance instead of multiscale decomposition or image segmentation. Moreover, the algorithm does not bring out problems caused by layer decomposition such as halo artifacts. The algorithm operates on windows, compresses the strong edges by preserving the small details, and imposes an optimization problem which combines a set of local-based constraints.

The algorithm is defined as linear mapping of HDR radiance map to LDR within small radiance groups on the image called windows. The mapping is formulated with the following basic linear function:

$$I^l(j) = p_i I^h(j) + q_i, \qquad j \in \omega_i, \tag{2.20}$$

where p and q are linear function parameters, I^l is the compressed pixel value, I^h is the original pixel value, and ω_i is the window i. The problem is essentially defined as an objective function minimization:

$$f = \sum_i \left(\sum_{j\in\omega_i} \left(I^l(j) - p_i I^h(j) - q_i\right)^2 + \varepsilon c_i^{-2}\left(p_i - c_i\right)^2 \right). \tag{2.21}$$

The term $ci^{-2}\left(p_i - c_i\right)^2$ is squared relative error of guidance map c_i which is contributed to the objective function in order to avoid trivial solution to p_i and q_i: 1 and 0. ε is weight of guidance error term.

The minimization process is conducted in the following convex function:

$$\arg\min_{p,q,I^l} f = \arg\min_{I^l} \sum_i \arg\min_{p_i,q_i} f_i, \qquad \text{where}$$
$$f_i = \left(\sum_{j\in\omega_i} \left(I^l(j) - p_i I^h(j) - q_i\right)^2 + \varepsilon c_i^{-2}\left(p_i - c_i\right)^2 \right).$$

The problem is solved by first setting the partial derivatives of p_i and q_i to zero and calculating the optimal I_i^ls by solving the resulting linear system. The window size is set to 3×3 by default.

Setting the guidance map directly controls p_i value which controls the contrast compression in a local window. Compressing high contrast and enhancing low contrast proportional to the standard deviation of the window ω_i is a naive approach but it amplifies noise on the image. Moreover, light regions have more contrast values compared to dark regions [42]. Therefore, the guidance map is calculated by the Gaussian approach using the following equation:

$$c_i = \left(\mu_i^{\beta_1} \sigma_i^{\beta_2} I^h(i)^{\beta_3} + \kappa\right)^{-1}, \tag{2.22}$$

where κ is a small weight which is set to 0.05, μ_i, σ_i, $I^h(i)$, and βs are window mean, window standard deviation, radiance value and attenuation constants of pixel i respectively. The sum of βs influences the compression rate of the image. Default value of βs are $\beta_1 = 0.6, \beta_2 = 0.2$ and $\beta_3 = 0.1$. Different values of βs give different results by means of local contrast and global illumination.

2.2 Exposure Fusion

Exposure fusion is an alternative technique to HDR image tone mapping. In exposure fusion, instead of compressing the luminances of an HDR image, the individual exposures of a bracketed image sequence are directly combined in the LDR domain. The result is an LDR image that contains details from all exposures.

The pioneering work for exposure fusion can be considered as Goshtasby's method [24]. This method involves separating the input exposures into fixed sized blocks and selecting the block with the highest entropy for the final image. The smooth transition between the blocks is ensured by using rational Gaussians. A more detailed description of this method is given in the following subsection.

Jo and Vavilin [25] propose a cluster based exposure fusion algorithm which blends regions that have the best exposures in the bracketed sequence of the scene. Similar to Goshtasby's work [24], each image is divided into several regions; however, in this method, the regions are constructed with neighboring pixels with similar intensity values rather than simply separating image to the fixed size blocks. After decomposing exposure images into clusters, fused image is formed by the pixels that are selected from the best exposed clusters which are determined by the entropy and detail calculation. The resulting image is then blurred with bilateral filtering [48] for smoothing sharp transitions between the connected clusters. Therefore, high contrasts are attenuated while small details are recovered in the final image.

Another method which is a probabilistic model based fusion technique for multi exposure images proposed by Shen et al. [43]. Input exposure images are weighted with a weight map produced by a probability function. Generalized random walk [30] is used to solve a global optimization problem of two quality measures in the multi exposure image data. The method gives plausible results that are comparable to other exposure fusion methods [24, 36] and tone mapping operators [32, 40, 42]. The method provides color consistency of the pixels avoiding unnatural and saturated color defects and preserves local contrast in the final image.

In the remainder of this section, the three exposure fusion algorithms that were used in this study are discussed in greater detail.

2.2.1 Fusion of Multi Exposure Images

This method provides a combination process of multi exposure images into a single image by selecting fixed blocks from different exposures of the bracketed image sequence. The selection of the blocks is based on the amount of information in each block. The selected blocks are blended into one image with a monotonically decreasing function. Optimal block size and function width parameters are iteratively altered in order to maximize the information content of the final image. Information is measured by entropy which is calculated as:

$$E_g = \sum_{i=0}^{255} -p_i \log(p_i), \tag{2.23}$$

where p_i is the probability of intensity i ranging from 0 to 255. $p_i = n_i/n$ where n_i is the number of pixel which has intensity i and n is the total number of pixels. n_i values are estimated by histogram calculation. In color images, image data is converted to CIELab space [13] and 3D histogram is clustered to the most dominant 256 colors with centers c_i via Xiang's clustering algorithm [54]. Entropy of the image is then calculated by the following equation:

$$E_c = \sum_{i=0}^{255} -p_i^c \log(p_i^c), \tag{2.24}$$

where p_i^c is the probability of pixels within the cluster i.

Each input image is divided into $d \times d$ blocks, which have 16 pixels at minimum, and the block which has the highest entropy among the corresponding image blocks with the same index among the other exposure images are selected for the final image. The selected image blocks are blended together to avoid discontinuities along the block borders.

The following monotonically decreasing blending function is used for combining image blocks:

$$\mathbf{O}(x, y) = \sum_{j=1}^{n_r} \sum_{k=1}^{n_c} W_{jk}(x, y)\mathbf{I}_{jk}(x, y), \tag{2.25}$$

where $\mathbf{I}_{jk}(x, y)$ is a pixel intensity value with index of (x, y) within the image blocks of index (j, k) in exposure sequence and $W_{jk}(x, y)$ is the weight of (x, y) index within the image block (j, k). The intensity value is the weighted sum of the pixels with same index (x, y) in exposure images where the weight of selected block is approximately 0.8 and the sum of other block weights is 0.2.

The weight function assigns maximum weight to the center of the selected blocks, decreasing as distance to the center of the block is increasing. Rational Gaussian surface [23] is used to model the blending function:

$$W_{jk}(x, y) = \frac{G_{jk}(x, y)}{\sum_{m=1}^{n_r} \sum_{n=1}^{n_c} G_{mn}(x, y)}, \tag{2.26}$$

n_r and n_c are the number of image blocks vertically and horizontally. $G_{jk}(x, y)$ is defined as:

$$G_{jk}(x, y) = \exp\left\{ -\frac{\left(x - x_{jk}\right)^2 + \left(y - y_{jk}\right)^2}{2\sigma^2} \right\}, \tag{2.27}$$

where (x_{jk}, y_{jk}) are the center coordinate of jk^{th} block and σ is the standard deviation of the Gaussian or width of blending function.

Two parameters d and σ are modified to obtain optimal values using gradient-ascent algorithm: d and σ parameters are initially set to 160. The values are incremented or decremented by a constant $\triangle = 32$ until the resulting image reaches the highest entropy compared the other parameter combinations.

2.2.2 Mertens et al.'s Exposure Fusion

This method fuses the individual exposures into one LDR image by weighting the pixels in the same position using three criteria namely well exposedness, saturation, and contrast [36].

For well exposedness, low weights are assigned to over and under exposed regions and high weights are assigned to pixels with moderate exposures. This is accomplished by using a Gaussian:

$$g(i) = \exp\left(-\frac{(i - 0.5)^2}{2\sigma^2} \right), \tag{2.28}$$

where i is the intensity value of a color channel and $\sigma = 0.2$. This Gaussian is applied to each color channel separately and their product is taken to compute the final weight. Contrast measure is determined by applying a Laplacian filter to the grayscale version of each image and taking the absolute value of the filter response. Saturation is determined by taking the standard deviation within the R, G, B values of each pixel.

The quality measures are combined by the following multiplication:

$$W_{i,j,k} = \left(C_{i,j,k}\right)^{\omega_C} \times \left(S_{i,j,k}\right)^{\omega_S} \times \left(E_{i,j,k}\right)^{\omega_E}, \tag{2.29}$$

where C, S and E are contrast, saturation and well exposedness weights. ω values are weighting exponents and i, j, and k refers to the pixel of k$^{\text{th}}$ image in (i, j) position. By default $\omega_C = \omega_S = \omega_E = 1$.

The weights of a pixel in position (i, j) are normalized by the following formula:

$$\hat{W}_{i,j,k} = \left[\sum_{k'=1}^{N} W_{i,j,k'}\right]^{-1} W_{i,j,k}, \tag{2.30}$$

and the final image is obtained by weighted blending:

$$R_{i,j} = \sum_{k=1}^{N} \hat{W}_{i,j,k} I_{i,j,k}. \tag{2.31}$$

This approach results in seam artifacts in the final image since weight map unexpectedly varies due to the difference of exposure times. Smoothing the weight map via Gaussian filters impose halo artifact around the edges and using bilateral filtering has problems with choosing optimal parameters. Therefore, the method uses Laplacian pyramid decomposition [10] of each exposure image and Gaussian pyramid of normalized weight maps of the images:

$$\mathbf{L}\{R\}^{l}_{ij} = \sum_{k=1}^{N} \mathbf{G}\left\{\hat{W}\right\}^{l}_{ij,k} \mathbf{L}\{I\}^{l}_{ij,k}. \tag{2.32}$$

$L\{A\}$ is the Laplacian pyramid decomposition of image A, $G\{B\}$ is the Gaussian pyramid of image B, and l is the level of the pyramid. In this formula, each level of the resulting Laplacian pyramid is expressed as weighted average of Laplacian pyramid of exposure image. The resulting pyramid is combined to obtain the final image. The operation is applied in every level separately.

Multiresolution blending clears seam artifact since it works on the image features instead of intensities directly and by the use of Laplacian filter factor, the effect of sharp differences in weight function are attenuated in the flat regions and preserved in edges.

2.2.3 Gradient Directed Composition of Multi Exposure Images

This method proposes a gradient based exposure fusion method in which a high quality image from exposure sequences of an HDR scene is aimed to be obtained by combining exposure images by means of visibility and consistency assessments [56] .

In an exposure sequence of a scene, it is observed that gradient magnitude is high for well exposed regions and low for over or under exposed regions [56]. Moreover, if moving objects exist in the scene, gradient direction changes occur in the exposure images. Therefore, gradient magnitude and direction are used as metrics for combining exposure images in order to produce well exposed images avoiding the ghosting artifact. As a result, a method which eliminates complex camera calibration and tone mapping operation and attenuates ghosting artifacts is introduced.

Exposure images are combined by the general formula:

$$H(x,y) = \sum_{i=1}^{N} W^i(x,y) I^i(x,y), \tag{2.33}$$

where N is number of exposures, I and W are the intensity and weight functions respectively. The result depends on the weight term W which is calculated by the gradient based visibility quality measure. The resulting image is produced seamlessly with a multiresolution reconstruction process using Laplacian pyramid.

The gradients of the exposure images are calculated by first derivative of 2D Gaussian kernel. Eventually, the gradient result is refined by cross bilateral filtering [39]. The visibility of a pixel is expected to be high where the gradient magnitude becomes larger since gradient decreases gradually if a detail is over exposed or under exposed in an exposure image. Therefore, the visibility assessment is developed as follows:

$$V^i(x,y) = \frac{m^i(x,y)}{\sum\limits_{i=1}^{N} m^i(x,y) + \epsilon}, \tag{2.34}$$

where $m^i(x,y)$ is the gradient magnitude in the pixel location (x,y) of i^{th} exposure image and ϵ is a small value to avoid singularity. In static scenes $V^i = W^i$ and this setting produces pleasant results in preserving details in every regions in all exposure images.

Visibility assessment is sufficient to set pixel weights if the scene is static. Therefore, consistency assessment of the method is not used in this thesis since all of the used exposure sequences represent static scenes.

2.3 Subjective Evaluation

As HDR imaging has been developing, several tone mapping operators have been presented in recent years. As a result, many case studies have been conducted for the assessment of the resulting image quality with performance analysis. Experiment types used in these works depend on the objective of the studies and various instructions such as ranking [11], rating [55] and preference [7, 11, 16, 50] were used in order to test operator performances with respect to some basic image attributes such as contrast, color, detail, brightness. In some of the experiments which test naturalness or reality, the real scene of tone mapped image was provided for reference [7, 11, 55]. However, most of the experiments were conducted with LCD or CRT display and tone mapping evaluation on a small screen is rare [50]. Moreover, a case study which evaluates exposure fusion techniques have not been proposed yet since exposure fusion is a new concept in HDR imaging field. Recently, a literature survey was carried out in order to compare exposure fusion algorithms but it does not involve a subjective evaluation [22].

A visual perception case study was conducted by Drago et al. [16] with 6 tone mapping operators. 11 subjects evaluated 4 images consisting of photographic and synthetic scenes. The tone mapping results were displayed in pairs to the subjects which were expected to specify distances between two images and choose the natural and better-looking one in the displayed image pair. The results were analyzed by INDSCAL and the first and the second most salient dimensions were associated with the detail and naturalness attributes respectively via the multiple regression analysis. Pairwise preference rankings obtained by subject choices in image pairs were used in PREFMAP, leading to determine an ideal point in INDSCAL-derived stimulus space. Results show that a proper contrast reduction which preserves spatial details are most likely to be preferred. In stimulus space, tone mapping operators resided in 3 groups which are categorized by their performances. In the first group, which has the best evaluation score, there are 3 stimuli near the ideal point. In the second group, there are 2 stimuli which have relatively natural results. In third group there is only histogram adjustment method [28] which has the worst performance, resulted from the unnatural image reproduction. Although the study proposes a sound comparison method, subjects are not asked to evaluate basic attributes (color and contrast etc.) of the image and the real world scene is not used for photographic scenes in order to compare naturalness of the image pair.

A psychophysical experiment was developed by Kuang et al. [27] This study was aimed to observe the relation between tone mapping and overall rendering performances. Approximately 30 subjects compared the image pairs of 10 scenes with a variety of dynamic range. Colored tone mapping results were used in overall rendering performance evaluation and gray scale images obtained from luminance channel of the tone mapping results were used in tone mapping evaluation. 10 scenes were rendered with 8 different HDR image rendering algorithms involving well-known global and local operators with default parameters. Colored and gray scale results were separately displayed in pairs to the participants which are asked to choose the image they preferred between two images. Image rendering and tone mapping results are statistically very similar and Durand's bilateral filtering [17] and Reinhard's global photographic tone reproduction [40] have the best scores in both evaluation results. In this study, it is concluded that the tone mapping results are very effective in overall HDR image rendering quality.

Another psychophysical experiment in tone mapping evaluation was conducted by Yoshida et al. [55]. 7 tone mapping operator results of 2 HDR images were compared with their corresponding real-world scenes. Each subject rated all 14 images by means of a set of attributes: naturalness, overall contrast, overall brightness, detail reproduction in dark regions, and detail reproduction in bright regions. Before conducting the experiment, a pilot study was carried out in order to fine tune parameters of tone mapping operators. Statistical analysis on evaluation results was conducted with ANOVA [47] and no significant difference was observed between 2 scenes. The study shows that tone mapping operators are perceived differently when compared with real-world scenes. Image attributes are not correlated with each other and the most uncorrelated attributes are overall brightness and detail reproduction in bright regions since high overall brightness causes the lack of detail visibility in bright regions. Mahalanobis distances calculated with MANOVA [47] show that the local and global tone mapping operator results are correlated within themselves. The global operators have higher overall brightness and strong contrast while the local operators are good at preserving details. The work has a sound contribution to the field since it proposed the first real-scene-referenced tone mapping evaluation. General characteristics of global and local operators can be easily observed in the resulting attribute scores. However, two scenes are similar by means of content and dynamic range. In order to strengthen the validity of results, different types of real world scene could be used to test other scene variations.

Ashikhmin and Goyal [7] proposed a tone mapping evaluation study which compares the evaluation of operator results with and without real world scene reference. 5 tone mapping results were tested with 15 participants in 3 types of experiments. In the first and second experiment, the subjects were expected to rank 5 tone mapping results of 5 scenes with respect to the preference and realism criteria. Unlike the first and second experiments, in the third experiment the real world reference was used for the evaluation and 4 scenes were ranked according to their similarity to the corresponding real world scene. 3 scenes were common in all experiments. All tone mapping results of a scene were displayed in the screen at once. Default parameters and authors' implementations were used for creating the tone mapped images. The first two experiments results are highly scene-dependent, preventing to form a significant ranking among tone mapping performances. However, in the third experiment, overall operator performances do not show drastic changes in 4 scenes. For this experiment, gradient domain compression [20] and adaptive logarithmic mapping [16] have the best scores. Results of the first and second experiments are relatively correlated but the third experiment is completely different than the other two. This observation can be interpreted that subjects' decisions change when the real scene is provided. Therefore, contrary to the Kuang et al.'s work [27], results of this study suggest that the virtual realism, which is defined as the subjective depiction of a real scene composed in human mind, is unreliable for evaluation of realism. Performances of operators are different than the other studies such as [27] and [55]. Therefore, the reliability of subjective studies without real world reference are questionable. Determining which attributes are exactly effective in subjects' decision is a very hard task. However, low or high presence of detail reduces the naturalness of a tone mapping image result and overall brightness is essential for reality of the images.

Cadik et al. [11] conducted perceptual experiments by evaluating tone mapping results by means of a set of attributes namely brightness, color, detail, contrast, artifacts, and overall quality. 2 different experiments each of which have 10 distinct participants were used to assess 14 tone mapping results of 3 common scenes. In the first experiment, subjects were asked to rate images on a standard LCD screen by using the real scene of the images as reference while in the second experiment, subjects were asked to rank printed versions of the same tone mapping results by comparing them with the real scene predictions emerged in participants' minds without a real scene reference. All the attributes stated above is evaluated in both referenced and non-referenced experiments. The study aimed to determine the effect of basic attributes of the image in overall quality by stating the correlation between the overall quality and the other attributes scores. This finding allows image quality to be defined by means of brightness, color, detail, and artifacts. Moreover, the influences of other attributes among each other were also examined. Statistical analysis showed that two experiment results are not statistically significant. It was concluded that the real scene reference is not necessary for perceptual evaluation of the tone mapping operators for this study unlike Ashikhmin and Goyal's work [7]. Subjects were statistically consistent among each other. The tone mapping performances changed for different input scenes. Results show that overall image quality was achieved best by linear clip followed by a group of tone mapping techniques which were global or have global tone mapping characteristics. The other methods had average scores and 2 operators [12]and [20] had the worst performances. Overall quality was influenced by the contrast, color, and artifact attributes, and brightness contributed indirectly by distributing its effect to the other attribute scores. The study involved several methods (14 tone mapping techniques); however, they are limited with tone mapping operators. In addition, they used 2 outdoor scenes which have high risk of the ground truth change for the referenced experiment.

A recent study which influenced this proposed work was conducted on both standard screens and small screen devices (SSD) by Urbano et al. [50]. The purpose of the study was to compare tone mapping operator performances in different screen sizes and find out whether the screen size is effective in subjects' choices. In the experiment, LCD, CRT, and PDA (personal digital assistant) screens were used to evaluate 7 tone mapping operator results of 2 indoor scenes. For each scene and device, 19 different subjects (114 subjects in total) participated in the evaluation. Tone mapping results were displayed in pairs and participants were asked to choose the image which is similar to the real scene based on the color, detail, contrast, and naturalness attributes, separately. Subject consistency and general agreement are ensured by the statistical analysis. LCD and CRT screen results were statistically similar but SSD results were different from both LCD and CRT screen results. The tone mapping operator which produced saturated colors and more details received better scores in SSD since color and detail attributes are essential in order to compensate limited screen size and color depth in SSDs. This study had an important contribution to the field of tone mapping evaluation by demonstrating the effect of the display properties on tone mapping performance. However, the experiment does not contain exposure fusion algorithms and excluded device dependent tone mapping operators which may be suitable for this type of experiment such as Mantiuk et al.'s display adaptive tone mapping [34].

A literature survey was recently conducted by Ganga et al. [22] for exposure fusion methods. 4 different exposure fusion approaches were introduced and discussed. In Goshtasby's work [24], every image is divided into image blocks and the image which contains the most information is selected among the different exposures of the block. Finally, image blocks are combined and blended. This method was described as not having a side effect such as increasing the contrast and color saturation. Mertens et al.'s work [36] weights each pixel resides in the same positions of different exposures according to the contrast, saturation and well exposedness quality. It was explained that the algorithm performs high contrast and good color reproduction. Another method proposed by Li et al. [31] uses features extracted from discrete wavelet frame transform coefficients in order to train support vector machines and selects the image that has the best focus for each pixel. However, it was mentioned that results are affected by slight object movements and defections in exposure sequence. Shen et al. [43] introduced a method in which generic random walk (GRW) model [30] applied to multi exposure image fusion. GRW is applied in an undirected graph in which a fused image pixels and input images are represented as nodes and local contrast and color consistency are used as quality measures. The survey concluded that GRW approach produces more natural images and has lower time complexity compared to other image fusion methods.

CHAPTER 3

EXPERIMENTAL METHODOLOGY

In the previous chapter, the algorithms used in the experiments are introduced with the other fundamental studies in literature, and some earlier similar subjective evaluations are discussed. In this chapter, the pairwise comparison is briefly mentioned and the statistical analysis methods of the pairwise experimental data used in this study are given in detail.

In this study, subjective evaluation which can be defined as human based evaluation was used in image quality assessment since it is accepted to be the best evaluation method for this task [41]. Subjective algorithm evaluation can be conducted with several types of comparison methods. Some of the previous subjective studies used rating [11, 16, 55] and ranking [11] evaluations. In this study, the pairwise comparison evaluation was preferred to compare image performances. Although comparing the image set in pairs is a time consuming task, it enforces the subjects to assess the images more than once unlike rating and ranking methods. Besides, it offers a reliability measurement of the subjects with a consistency analysis method. The following part of the chapter contains further information about the statistical data analysis of the pairwise comparison results.

3.1 The Method of Paired Comparisons

While processing experimental data obtained from pairwise comparison, it is important to validate the significance of the results in order to declare any conclusion. Therefore, reliability of the results should be tested by statistical analyses. Two analysis methods are conducted to resulting experimental data in order to state consistency of subject preference and significance of method scores. Detailed information about analysis methods are provided in the following part of this chapter.

3.1.1 Method of Consistency Analysis

Pairwise comparison aims to scale and rank quality of compared object. It is carried out by asking to participants to prefer one object to the other in pair and all possible pairwise preference results lead to an overall ranking among all objects. However, misunderstanding of the experiment objectives or reluctance of the subject himself may cause inconsistencies in the final ranking even if the preference process is done for every pair of objects. Furthermore, the difficulty of the experimental task may cause the participants to make inconsistent decisions.

Suppose a subject participate in an experiment in which the evaluated objects involve A, B, and C. If a subject preferred A to B and B to C, a ranking can intuitively be done as $A > B > C$ for the subject's overall evaluation. However, if the subject chooses C in object pair (A, C) then this preference causes inconsistency in the final ranking.

Kendall and Smith [26] introduce a metric for subject consistency that stems from the case explained above. Suppose, in a pairwise comparison experiment, the demonstration of preference in object pair (A, B) is $A \rightarrow B$ or $B \leftarrow A$, if A is preferred to B. With n objects to compare, this gives rise to a total of $t(t-1)/2$ preferences. These preferences can be displayed in a complete directed graph in which each node corresponds to an object. An example preference matrix and its preference graph are given in Table 3.1 and Figure 3.1.

Table 3.1: An example preference matrix.

	A	**B**	**C**	**D**	**E**
A	0	1	1	1	0
B	0	0	1	0	0
C	0	0	0	1	0
D	0	1	0	0	1
E	1	1	1	0	0

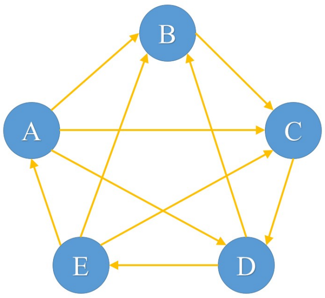


Figure 3.1: Preference graph for matrix in Table 3.1.

A preference set consisting of 3 objects with their connecting edges in a preference graph is called a triads (e.g. $A \rightarrow B$, $B \rightarrow C$, and $A \rightarrow C$) [26]. Some triads form a loop in the overall preference graph, causing inconsistencies in the final evaluation results of a subject. These are called circular triads (e.g. $A \rightarrow B$, $B \rightarrow C$, and $C \rightarrow A$) which are considered to be the simplest unit of inconsistency.

Although circular n-ads are also possible for a set of preferences, the number of triads are used for consistency analysis since circular n-ads contain at least $(t-2)$ circular triads. The consistency of a subject can be evaluated by comparing the number of circular triads with the total possible number of triads as explained below.

Let $\alpha_1, \alpha_2, ..., \alpha_t$ be the number arrows pointing out from each object r for $r = 1, 2, \ldots, t$ (that is the number of preference of each object):

$$\sum_{r=1}^{t} (\alpha_r) = \left(\begin{array}{c} t \\ 2 \end{array} \right). \tag{3.1}$$

The mean of α_r is $(t-1)/2$. A T value can be defined as:

$$T = \sum_{r=1}^{t} \left(\alpha_r - \frac{t-1}{2} \right)^2, \tag{3.2}$$

which substitutes into

$$T = \sum_{r=1}^{n} \left(\alpha_r^2 \right) - \frac{t\,(t-1)^2}{4}. \tag{3.3}$$

The number of circular triads change resulting from inversion of an arrow in preference graph will be examined. Suppose that in a subject evaluation, $A \to B$ preference exists. There are α preferences with $A \to X$ including $A \to B$ and β preferences with $B \to X$ where X is any object in experiment. Then four types of triads are possible in the preference graph:

$A \to X \leftarrow B$, p number of preferences in total
$A \leftarrow X \to B$,
$A \to X \to B$, $\alpha - p - 1$ number of preferences in total
$A \leftarrow X \leftarrow B$, $\beta - p$ number of preferences in total

When $A \to B$ is changed to $A \leftarrow B$, the first two types of triads stay as non-circular triads. The third type is turned into circular triads and the fourth one becomes non-circular when they were circular before. Therefore, the increase in circular triad count is equal to:

$$(\alpha - p - 1) - (\beta - p) = \alpha - \beta - 1 = d, \tag{3.4}$$

and the decrease in T value is:

$$(\alpha^2 + \beta^2) - \left((\alpha - 1)^2 + (\beta + 1)^2 \right) = 2\,(\alpha - \beta - 1) = 2d, \tag{3.5}$$

where d value is the change in the number of circular triads. This shows that if the circular triad count increases by d, then the T value decreases by $2d$.

T value reaches its maximum value when α_r's are successive $(1, 2, \ldots, t-1)$ since the overall result forms a normal ranking. Therefore, the maximum T value becomes $T = \left(t^3 - t\right)/12$. The minimum T value is reached when α_r values are close to each other or equal if possible. This is achieved when $\alpha_r = (t-1)/2$ when n is odd and $t/2$ nodes have a preference of $t/2$ and the remaining $t/2$ nodes have a preference of $(t-2)/2$. In this case T becomes equal to $t/4$.

Therefore, the range of T is zero to $\left(t^3 - t\right)/12$ if t is odd and $t/4$ to $\left(t^3 - t\right)/12$ if t is even. We can derive from the previous equations that the circular triad count change is equal to the half of the variation in T value. Therefore, the maximum number of circular triads in a preference graph is $(t^3 - t)/24$ if t is odd and $(t^3 - 4t)/24$ if t is even.

Finally, the coefficient of consistence ζ is calculated by finding the ratio of the current number of circular triads to the maximum number of circular triads and subtracting this quantity from unity:

$$\zeta = 1 - \frac{24\left(\frac{t^3-t}{24} - \frac{1}{2}\sum_{i}^{t}\left(\alpha_i - \frac{t-1}{2}\right)^2\right)}{t^3 - t} \qquad \text{if t is odd,} \tag{3.6}$$

$$\zeta = 1 - \frac{24\left(\frac{t^3-t}{24} - \frac{1}{2}\sum_{i}^{t}\left(\alpha_i - \frac{t-1}{2}\right)^2\right)}{t^3 - 4t} \qquad \text{if t is even.} \tag{3.7}$$

3.1.2 Method of Significance Analysis

Preference testing scores contribute to a ranking when combining pairwise comparison choices. However, these data need to be analyzed to determine whether the score differences are significant enough to constitute a proper ranking. Starks and David [44] propose a statistical data analysis that provides a significance test for paired comparison experiments. In this thesis, the evaluation scores are analyzed by using this method as explained below.

Assume that a pairwise experiment is conducted with t number of treatments (items or objects) and with n number of subjects (participants). The number of times a treatment i where $i = 1, \ldots, t$ is preferred is a_i and the probability of treatment i being preferred to treatment j is π_{ij}. An average preference probability of treatment u is then given by:

$$\pi_{u.} = \sum_{j=1}^{t} \pi_{uj} / \left(t - 1\right), \tag{3.8}$$

where $j \neq u$.

Two tests are used in order to analyze the significance of score difference for any a_i and a_j.

3.1.2.1 Test of Equality of Two Preassigned Treatments

This test is conducted to determine if any distinguishable difference between treatments u and v exists. The following null hypothesis is tested:

$$H_0 : \pi_{u.} = \pi_{v.}, \tag{3.9}$$

against

$$H_a : \pi_{u.} \neq \pi_{v.}. \tag{3.10}$$

The test is developed under

$$H_0' : \pi_{ij} = \frac{1}{2} \quad \text{for all } (i, j). \tag{3.11}$$

The test is done by the probability calculation of $|a_u - a_v| \geq m$ where m is a positive integer. Ultimately, for a predetermined significance level, an m_c number is determined. If $|a_u - a_v| \geq m_c$ then H_a is accepted. That means, if the previous condition is achieved, a_u is significantly better than a_v. A table that gives the appropriate m_c values for various values of n and t and two significance levels $\alpha = 0.01$ and $\alpha = 0.05$ is given in [44].

Although this test can state that a difference between two preassigned treatments is significant or not, it does not provide a general difference metric in order to separate any two treatment scores. To achieve this, one needs to perform a multiple comparison test.

3.1.2.2 Multiple Comparison Test

In [44], two types of multiple comparison tests, multiple comparison range test and least significant difference method, are introduced. The latter is used in this work in order to compare the results of multiple algorithms. In this method, sample means are compared using variance analysis. The method initially tests

$$H_0 : \pi_i = \frac{1}{2} \quad \text{for all } i, \tag{3.12}$$

against

$$H_\alpha : \pi_i \neq \frac{1}{2} \quad \text{for some } i. \tag{3.13}$$

If H_0 is not rejected, that means there are no significant differences between scores. If not, the steps below are applied to determine the threshold beyond which the differences can be considered significant:

1. Determine the significance level α
2. (a) For small n and t values, use tables in [9], [8] and [15] to test H_0.
   (b) For larger values which do not exist in these tables, use Durbin's method [18] with the following formula:
   $$D = \sum_{i=1}^{t} d_i^2 = 4\left[\sum_{i=1}^{t} (a_i - \overline{a})^2\right] \Big/ (nt), \tag{3.14}$$
   which substitutes into
   $$D = \sum_{i=1}^{t} d_i^2 = 4\left[\sum_{i=1}^{t} a_i^2 - \frac{1}{4} tn^2 (t-1)^2\right] \Big/ (nt), \tag{3.15}$$
   and compare the result with the upper $100\alpha\%$ point of the χ^2-distribution with $(t-1)$ degrees of freedom. Reject the null hypothesis H_0 if D is greater than the critical χ^2 value.
3. If the null hypothesis H_0 cannot be rejected then the test is completed and no significant difference is found. Otherwise, apply the "equality of two pre-assigned treatments test" to find the critical value m_c for the two sided test in Table 1 in [44]. Any two treatments which differ at least by m_c are declared to be significantly different.

CHAPTER 4

EXPERIMENTS AND RESULTS

In this thesis, four tone mapping operators and three exposure fusion algorithms are compared with each other. Below, first the motivation for selecting these algorithms are given. This is followed by the description of the stimuli used in the experiments. Finally, the details and results of each experiment are elaborated.

4.1 Selected Algorithms

Although there is a myriad of tone mapping and exposure fusion algorithms comparing all of them in a single study is impossible for practical reasons. Therefore, in this study the comparison is made between four tone mapping and three exposure fusion algorithms. These are (the symbols in parenthesis serve as their identifiers):

- Block based EFA (A) [24]
- Subband based TMO (B) [32]
- Display adaptive TMO (C) [34]
- Multiscale EFA (D) [36]
- Global photographic TMO (E) [40]
- Linear windowed TMO (F) [42]
- Gradient based EFA (G) [56]

The selected algorithms include the state-of-the-art methods in both fields and some of them have been found to be the best methods in earlier validation experiments.

The photographic TMO had good performance in the earlier case studies [6, 16, 27, 29, 50, 55]. The presence of the photographic TMO is also motivated by its existence in several evaluation studies. This may allow to compare any algorithm performance of this study and the other studies by taking the common operator result as a reference. Display adaptive TMO was selected to test an operator designed for handling various device and ambient lighting conditions and the second experiment contained a proper test environment for display adaptive TMO since it provided standard and small screen evaluations and various illumination conditions. Li et al. and Shan et al.'s operators were included in the algorithm set since they produce high quality results and author implementation of the operators are available.

Mertens's et al. and Goshtasby's EFAs are the two fundamental studies among the EFAs. Therefore their presences are very important for this experiment, which is the first subjective evaluation of exposure fusion in literature. Zhang and Cham's EFA were chosen due to its high quality of image results.

4.2 Stimuli

The selected algorithms were tested using 4 photographic scenes. In order to create different test cases, the scenes were selected in a variety of content, lighting condition, dynamic range, and detail. 4 scenes used in the experimental evaluation are listed below:

- Ametyst
- Lamp
- Toys
- Trail

Ametyst is an indoor scene with different-colored object illuminated by a desk lamp. Strong highlights are visible on some objects. Lamp scene, on the other hand, represents an outdoor night environment that contains a bright light source with heavy foliage. Toys is an another indoor scene but illuminated with standard florescent lights and has relatively low dynamic range. Trail scene represents a bright daylight environment that is partly shadowed due to arching trees.

Table 4.1: The dynamic ranges of the images used in the experiments. DR stands for the order of magnitude difference between the maximum and minimum luminances. DR^1 represents the dynamic range after the removal of %1 of the least and greatest pixel values.

	Ametyst	**Lamp**	**Toys**	**Trail**
DR	4.27	4.51	2.99	4.33
DR^1	2.40	3.75	2.08	2.46

The dynamic ranges of the images are stated in Table 4.1. As illustrated in this table, all images have similar dynamic ranges except toys scene. In the second row of the table, the darkest and lightest pixels are excluded from dynamic range calculations in order to obtain the dynamic range values without outliers. In both cases, lamp has the highest dynamic range, especially after the removal of the outliers, there is a remarkable difference with the other dynamic range values.

The scenes were photographed in a way that the dynamic range of the medium lighting was covered. To achieve this, multiple shots of the same static scenes with different exposure times were taken as shown in Figure 4.1. Each bracketed sequence was taken in RAW format using a Canon EOS550D camera in order to avoid the necessity of determining the camera response curve (CRF) (The pixel values in RAW images are linear with respect to the scene luminance). The exposure spacing between each exposure was 1-fstop.

The images were then converted to the JPEG format using the sRGB non-linearity which serves as a common starting point for both tone mapping and exposure fusion algorithms. In the case of TMOs, first an HDR image was created using a standard HDR assembly equation:

$$I_j = \sum_{i=1}^{N} \frac{f^{-1}\left(p_{ij}\right)\omega\left(p_{ij}\right)}{t_i} \Big/ \sum_{i=1}^{N} \omega(p_{ij}), \tag{4.1}$$

where N is the number of exposures, p_{ij} is pixel value in position j in the i^{th} image, f is the camera response function, which is the inverse sRGB gamma in this case, ω is the weighting function to reduce the effect of under and over exposed pixels [38], and t_is are exposure times. The generated HDR images were tone mapped by the 4 tone mapping algorithms listed in Section 4.1. As for the exposure fusion results, the JPEG images are directly combined using the selected EFAs.

Table 4.2: Parameters of the algorithms used in the experiments. For all TMOs, gamma is set to 0.45. The value of the ambient light parameter for toys scene and the maximum luminance parameter for the camera screen are given in parenthesis.

Alg.	Parameters
A	$d = 160, \sigma = 160, \Delta = 32$
B	$\alpha = 0.2, \gamma = 0.6$
C	$E_{amb} = 10(250), L_{min} = 0.5,$ $k = 0.01, L_{max} = 80(125)$
D	$\omega_C = 1, \omega_E = 1, \omega_S = 1$
E	$\alpha = 0.18, L_{white} = 1e20$
F	$\epsilon = 0.1, \kappa = 0.05, windowsize = 3,$ $\beta_1 = 0.6, \beta_2 = 0.2, \beta_3 = 0.1, s = 1$
G	$l = 9, \tau = 0.9$

All of the selected algorithms were implemented with the original author implementations except Zhang and Cham's TMO since its implementation has not been made publicly available by their authors. Default parameters are used in the implementations of algorithms. However, parameters of display adaptive TMO are set according to the black and white luminance values of the display devices and ambient illumination of the experiment room. The parameter set for all algorithms are reported in Table 4.2. All of the TMO and EFA results are shown in Figure 4.2.

The final tone mapping results together with fused images were transferred to the screen devices in experiment location. Overall illustration of image acquisition chart is provided in Figure 4.3.

4.3 Experimental Setup

After generation of tone mapping and exposure fusion images, two subjective experiments were conducted. The aim of the first experiment was to evaluate which algorithm better preserves the attributes of color, contrast, and detail on a standard desktop screen. Computer screens are defined to be standard screens since they are commonly used in everyday life, especially when displaying a digital photograph. The second experiment was aimed to evaluate the similarity of the images to their real-world counterparts.

Both experiments were conducted in an experiment room which had no windows and is totally isolated from the outside environment (see Figure 4.4). During the first experiment no additional light source was used to illuminate the experiment room. When conducting the second experiment, the room was illuminated by a desk lamp or florescent lights depending on the real world reference scene being tested at that moment. The LCD monitor was calibrated to the sRGB profile using an X-Rite i1Display Pro colorimeter. The camera's LCD display was used only in the second experiment. The details of the display devices are provided in Table 4.3

Table 4.3: Screen device properties.

	Camera display	**Standard LCD display**
Name & brand	Canon EOS 550D	NEC SpectraView Reference 241W
Size	3 inches in diagonal	$51.9cm \times 32.2cm$
Resolution	720×480	1920×1200
Min. luminance	$0.5cd/m^2$	$0.5cd/m^2$
Max. luminance	$125cd/m^2$	$80cd/m^2$

4.4 Experiment One: Color, Contrast, and Detail

In order to examine the performance of the selected algorithms, the resulting images were tested by a set of quality attributes namely color, contrast, and detail. These attributes were also used as the quality measures by several earlier tone mapping evaluation studies such as [11, 16, 50, 55]. To ensure that the meaning of these attributes are understood, they were defined in the instruction given to the participants (see Appendix A).

The main objective was to rank the methods in different scenes according to these quality attributes. In the first experiment, ametyst, lamp, and trail scenes were tested. The subjects were asked to compare the image results in pairs and choose the better one according to their preference. Each attribute-scene combination was evaluated once at a time and after all possible pairs of an attribute-scene combination were finished, the set of random image pairs for the next scene-attribute combination was shown to the subject. Attributes were evaluated in the order of color, contrast, and detail. The scenes were ordered as ametyst, lamp, and trail within each attribute evaluation.

The images were displayed on a standard LCD display with a custom experiment software implemented using Physchtoolbox [1] in Matlab environment. The program first created all possible image pairs of a scene and presented them to the participants in random order. For each pair, the participants could view one image at a time in full screen and switch between the images by using the arrow keys. The participants indicated their preference by pressing the enter key while the preferred images were shown. This choice was recorded by the program and a neutral gray screen was displayed for 5 seconds before switching to the next pair. Each image pair was tested only once. At the end of the experiment, the preference matrices of the subject were written in a file in order to be used in the analysis phase.

There were 15 people which participated in the first experiment. The participants involved graduate students and instructors mostly from the computer engineering department. Before the experiment, an instruction manual was provided to the subjects in order to clarify regulations and objectives of the experiment (see Appendix A). Each subject was allowed to take the experiment individually. For each subject, the experiment took about 30 minutes on average.

Every subject made 189 comparisons ($\begin{pmatrix} 7 \\ 2 \end{pmatrix}$ image pairs $\times$ 3 scenes $\times$ 3 attributes) in total. $189 \times 15 = 2835$ pairwise comparisons were made for all subjects.

4.4.1 Results

After all the image evaluations were finished, the collected data was processed by using the paired comparisons analysis as explained in Section 3.1. The data are gathered as preference matrices formed by choices of the subjects. The accumulated preference matrices are given in Table 4.4. The algorithm names are abbreviated with the identifiers introduced in Section 4.1.

4.4.1.1 Consistency Analysis

In order to assess the reliability of the results, a measurement has to be made to determine whether preferences of a subject can represent a ranking. Therefore, Kendall's consistency test discussed in Section 3.1.1 was conducted to every preference matrix of every subjects. In these experiments, the number of objects t is equal to 7. The resulting consistency values are listed in Table 4.5.

In Table 4.5 it is observed that all subjects has a high degree of consistency. Therefore, no subject result is excluded from the data analysis. It can be stated that the consistency values of

Table 4.4: Accumulated preference matrices for the first experiment. T stands for the total test score of the algorithm for the corresponding accumulated preference matrix.

		Color								**Contrast**								**Detail**							
		A	**B**	**C**	**D**	**E**	**F**	**G**	**T**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**T**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**T**
Ametyst	**A**	0	12	3	2	5	7	2	31	0	3	11	8	14	14	9	59	0	0	9	11	9	6	13	48
	B	3	0	2	5	3	6	5	24	12	0	12	8	15	15	14	76	15	0	15	13	14	15	15	87
	C	12	13	0	10	10	9	11	65	4	3	0	5	13	15	9	49	6	0	0	10	5	5	13	39
	D	13	10	5	0	6	8	10	52	7	7	10	0	14	14	13	65	4	2	5	0	7	4	15	37
	E	10	12	5	9	0	9	11	56	1	0	2	1	0	5	2	11	6	1	10	8	0	8	12	45
	F	8	9	6	7	6	0	5	41	1	0	0	1	10	0	3	15	9	0	10	11	7	0	13	50
	G	13	10	4	5	4	10	0	46	6	1	6	2	13	12	0	40	2	0	2	0	3	2	0	9
Lamp	**A**	0	7	9	5	11	11	3	46	0	10	5	4	5	13	8	45	0	0	8	3	8	15	10	44
	B	8	0	9	6	13	8	4	48	5	0	6	4	7	14	8	44	15	0	14	14	14	15	14	86
	C	6	6	0	4	12	6	3	37	10	9	0	12	9	13	10	63	7	1	0	4	5	15	13	45
	D	10	9	11	0	14	10	5	59	11	11	3	0	6	13	11	55	12	1	11	0	10	15	15	64
	E	4	2	3	1	0	4	3	17	10	8	6	9	0	13	11	57	7	1	10	5	0	15	11	49
	F	4	7	9	5	11	0	3	39	2	1	2	2	2	0	2	11	0	0	0	0	0	0	1	1
	G	12	11	12	10	12	12	0	69	7	7	5	4	4	13	0	40	5	1	2	0	4	14	0	26
Trail	**A**	0	12	8	9	10	10	5	54	0	3	2	0	10	10	2	27	0	1	10	0	7	10	4	32
	B	3	0	3	0	3	5	2	16	12	0	10	12	14	14	13	75	14	0	15	14	15	14	15	87
	C	7	12	0	7	8	9	4	47	13	5	0	5	10	13	11	57	5	0	0	0	3	10	0	18
	D	6	15	8	0	10	10	9	58	15	3	10	0	14	13	11	66	15	1	15	0	12	14	13	70
	E	5	12	7	5	0	11	4	44	5	1	5	1	0	9	3	24	8	0	12	3	0	12	3	38
	F	5	10	6	5	4	0	4	34	5	1	2	2	6	0	2	18	5	1	5	1	3	0	1	16
	G	10	13	11	6	11	11	0	62	13	2	4	4	12	13	0	48	11	0	15	2	12	14	0	54
Total	**A**	0	31	20	16	26	28	10	131	0	16	18	12	29	37	19	131	0	1	27	14	24	31	27	124
	B	14	0	14	11	19	19	11	88	29	0	28	24	36	43	35	195	44	0	44	41	43	44	44	260
	C	25	31	0	21	30	24	18	149	27	17	0	22	32	41	30	169	18	1	0	14	13	30	26	102
	D	29	34	24	0	30	28	24	169	33	21	23	0	34	40	35	186	31	4	31	0	29	33	43	171
	E	19	26	15	15	0	24	18	117	16	9	13	11	0	27	16	92	21	2	32	16	0	35	26	132
	F	17	26	21	17	21	0	12	114	8	2	4	5	18	0	7	44	14	1	15	12	10	0	15	67
	G	35	34	27	21	27	33	0	177	26	10	15	10	29	38	0	128	18	1	19	2	19	30	0	89

color evaluation is relatively small compared to others. This suggests that it was more difficult to decide color quality than contrast and detail. It should also be noted that the average consistency for the ametyst-color combination was the lowest among all other combinations. This can be due to showing this combination as the first one to all subjects. That is, as the experiment progressed, the subjected learned to make more consistent decisions.

4.4.1.2 Significance Analysis

After the consistency analysis, the second test was conducted to understand how significant the differences among total method scores were. To this end, the multiple comparison test described in Section 3.1.2.2 was applied to find the m_c value which was used to test the significance of difference between two scores.

For each scene-attribute combination, the n and t values were set to 15 and 7 to determine the D value using Equation 3.14. D values of all tests are given in Table 4.6. χ^2 distribution with $(t - 1) = 6$ degrees of freedom at significance level $p = 0.05$ is 12.59. Therefore, all test results were found to be statistically significant. For the aggregated results, n was set to 45 as the results of three scenes were combined. This gave rise to the D values shown in Table 4.7. Note that significance level p is chosen to be 0.05 as a common practice.

Table 4.5: Subject consistency values for the first experiment.

	Color			**Contrast**			**Detail**		
Sub. #	**Ametyst**	**Lamp**	**Trail**	**Ametyst**	**Lamp**	**Trail**	**Ametyst**	**Lamp**	**Trail**
1	0.786	0.929	0.929	1.000	1.000	1.000	1.000	0.929	1.000
2	0.429	1.000	0.929	0.857	1.000	0.929	0.643	1.000	0.929
3	0.571	0.929	0.714	0.786	0.929	1.000	0.714	0.929	1.000
4	0.714	1.000	0.714	0.929	1.000	0.929	0.714	1.000	1.000
5	0.714	0.643	0.500	0.857	0.857	0.571	0.929	0.714	0.571
6	1.000	0.786	0.714	1.000	0.786	0.714	1.000	0.929	0.571
7	0.571	0.643	0.714	0.786	1.000	0.929	0.929	0.929	1.000
8	0.286	0.929	1.000	1.000	1.000	1.000	1.000	0.857	1.000
9	0.571	0.857	0.929	0.786	0.857	0.643	0.929	1.000	0.857
10	0.714	0.643	0.714	0.857	0.857	0.929	0.643	1.000	0.929
11	0.286	0.429	0.571	0.929	0.857	1.000	0.929	0.786	0.857
12	1.000	0.571	0.857	0.500	1.000	0.929	1.000	1.000	1.000
13	0.571	0.571	0.786	1.000	0.786	1.000	0.929	0.857	1.000
14	0.786	0.571	0.857	0.929	0.929	1.000	0.714	0.857	0.929
15	0.571	0.857	1.000	0.857	0.857	0.857	0.714	1.000	1.000
Avg.	0.638	0.757	0.795	0.871	0.914	0.895	0.852	0.919	0.910

Table 4.6: D values for the first experiment.

	Ametyst	**Lamp**	**Trail**
Color	46.629	63.467	57.371
Contrast	139.200	66.667	113.828
Details	121.676	165.942	162.209

According to Table 1 in [44], m_c value was calculated with the following formula:

$$m_c = \lceil 1.96\,(0.5nt)^{0.5} + 0.5 \rceil. \tag{4.2}$$

For $n = 15$ and $t = 7$, m_c is equal to 15. This means that two scores with a difference not less than 15 are statistically significant. For the aggregated results m_c is equal to 26. Based on these, the similarity groups for the individual scene-attribute combinations as well as the aggregated results for each attribute are shown in Figures 4.5, 4.6, 4.7, and 4.8. In these figures, the algorithms underlined by the same line belong to the same similarity groups. That is, the differences between them are not statistically significant.

To illustrate, in Figure 4.8, the difference between Mertens (D) and Zhang (G)'s color attribute scores is 8 which is less than m_c value 26. Therefore, the scores are accepted to be statistically similar and these algorithms share the same horizontal line below their names in the figure. However, the scores of Li (B) and Shan (F) in the same attribute test differ 26, equal to the m_c value. Thus, they are stated to be significantly different and they belong to the different significance groups in the figure.

Table 4.7: Aggregated D values for the first experiment.

Color	Contrast	Details
77.537	222.883	315.937

4.4.1.3 Evaluation of Results

In the previous section, the experimental data is validated by the statistical models which test the significance of the collected results. In this section, these findings are interpreted.

The algorithms were tested in 3 scenes which have different characteristics by means of luminance, detail, and content. The accumulated scores are generally significantly different from each other and significant groups do not contain more than three operators (see Figure 4.8). The results show that most EFAs are better in color reproduction, while contrast and detail performances of the best TMOs outperform EFAs. It can also be stated that parameter setting is important in the ranking of the algorithms since bad image results due to improper parameter setting for a specific scene causes a drastic change in the ranking of the algorithm. In the following, the performance of each algorithm in the first experiment is discussed separately.

Goshtasby (A) results have moderate performance on average, although images are relatively pale compared to other TMO and EFA results. Contrast performance in trail image is the worst. It can be observed that Goshtasby's EFA could not succeed in extending the luminance range possibly resulting from the fact that small size of well exposed regions increased the number of blocks, leading to too much blending and loss of contrast in the fused image. Detail performance of the algorithm is not worse than others although the algorithm is stated to fail to spill information across object boundaries [36].

In general, Li et al. (B) is observed to produce images that have saturated colors. Although this algorithm provides the most vivid colors among the other algorithms, the colors of the images are unnatural. Therefore, color performances of the algorithm are relatively worse since the subjects were asked to evaluate the naturalness of colors in this task (See Appendix A). The detail performance of this algorithm is significantly better than the other TMOs and EFAs. This can be attributed to the fact that this is a local tone mapping operator.

Mantiuk et al. (C) performances are average in the image attributes, but relatively better in contrast. Lamp result has glare around the light source. Also trail image output has over exposed regions. Therefore, the operator has the worst performance in detail reproduction for the trail scene. Both lamp and trail outputs show that direct lighting and sharp bright regions cause burn-out in this algorithm. In contrast, the algorithm performed well in contrast reproduction, considering the image result of the dark lamp scene.

Mertens et al.'s method (D) has a moderate performance on average, similar to Goshtasby's results but it produces relatively colorful images. Contrast reproduction is better than the average possibly resulting from the contrast quality measure imposed by the weight calculation in the algorithm's implementation.

Reinhard et al. (E) scores are mostly in the middle of performance rankings. Color performances of the lamp scene are significantly the worst amongst the others. In color evaluation, the subjects may be distracted by the glare around the lamp which deteriorates the naturalness of the lamp image. Global TMOs do not succeed in preserving details in high contrast images [41]. Therefore, both Mantiuk et al. and Reinhard et al. stays in medium range in detail reproduction ranking.

Shan et al. (F) can be considered as the most unsuccessful algorithm among the other TMOs and EFAs. It is possible that the default parameters for this algorithm did not work very well for the images used in this experiment. Color results are moderate, but contrast and detail quality of every scene have the worst performance except detail reproduction of the ametyst scene. The algorithm tends to smooth small details. Thus the algorithm fails in lamp and trail scene and produces reasonable result in ametyst scene. Lamp image result of the operator is dark, leading to be significantly worst in contrast and detail reproduction.

Zhang and Cham (G) produced the most preferred color quality together with the Mertens et al. image results. Color quality is better in lamp and trail scene compared to the ametyst scene since the lamp and trail scenes have small details that increase the gradient and, as a result, well exposed region representation in the implementation. The algorithm has the worst detail representation in ametyst and lamp scene since the dark regions are mapped to zero or low luminance values, causing to the loss of details in the final image. The trail scene is not affected from the bad reproduction of details, for it is relatively bright and does not have large dark regions.

4.5 Experiment Two: Similarity

In this experiment, the similarity attribute of the methods was tested on different screen sizes. Ametyst and toys scenes were evaluated with their real scene references. Since outdoor scenes are vulnerable to content and lighting condition modifications, 2 static indoor scenes with different dynamic ranges were used. The experiment was aimed to measure the similarity performances of TMOs and EFAs in different display devices and determine whether the display device properties are effective in the similarity performance.

2 scenes were first evaluated on a standard LCD screen. After the standard screen evaluation is finished, the camera screen evaluation is conducted. The scenes were ordered as ametyst and toys within each screen evaluation. The image pairs were organized randomly similar to the first experiment. Ametyst scene was common in both experiments in order to observe the correlation among the similarity and the other 3 attributes tested in the first experiment.

In the second experiment, the algorithm results were organized as image pairs similar to the first experiment and the subjects were asked to determine which image in the image pair is more similar to the corresponding real scene residing to the left and right side of the LCD display (see Figure 4.4). Standard LCD monitor test was conducted with the program used in the first experiment. In the camera screen test, the experimental setup was organized manually. A randomized set of image pairs were created by using a bash script. A neutral gray image was placed in between every image pair. The resulting image sequence was transferred to the camera with an SD memory card. Subjects could use buttons near the camera screen to change the displayed image in the current pair and pass to the next image pair. Similar to the standard LCD display test, the subjects could view only one image at a time in full screen mode. For every image pair, the subjects stated the preferred image name verbally to the experimenter. These image names were noted down in order to compose the preference matrices of the participant for further analysis.

There were 15 people which participated in the second experiment. The subject profile of this experiment was similar to that of first experiment. Before the experiment, an instruction manual was provided to the subjects in order to clarify the regulations and objectives of the experiment (see Appendix B). Each subject was allowed to take the experiment individually. For each subject, the experiment took about 20 minutes on average.

Every subject made 84 comparisons ($\binom{7}{2}$ image pairs $\times$ 2 scenes $\times$ 2 screens) in total. $84 \times 15 = 1260$ pairwise comparisons were made for all subjects.

4.5.1 Results

After the second experiment was finished, the collected data was processed by using the paired comparison analysis similar to the first experimental data evaluation. The accumulated preference matrices are given in Table 4.8. The algorithm names are abbreviated with the identifiers introduced in 4.1.

4.5.1.1 Consistency Analysis

Similar to the first experiment, the reliability of the results are measured by Kendall's consistency test mentioned in Section 3.1.1. The number of objects t is the same as the first experiment's number of objects, 7. The resulting consistency values are listed in Table 4.9.

Table 4.9 shows that subject are consistent on average. It can be observed that the average consistency value of ametyst scene evaluation in standard LCD display device is relatively low. Therefore, it can be concluded that the similarity evaluation conducted in standard LCD screen is harder than the same evaluation conducted in small camera screen. Similar to the first experiment, the first evaluated image pair set, which is the image pair set of computer-ametyst combination in this experiment, has the lowest average consistency value. This shows that the reliability of subjects improved by the time during the experiment.

Table 4.8: Accumulated preference matrices for the second experiment. Camera columns represents small camera screen and computer column represents standard LCD screen. T stands for the total test score of the algorithm for the corresponding accumulated preference matrix.

		Computer								**Camera**							
		A	**B**	**C**	**D**	**E**	**F**	**G**	**T**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**T**
Ametyst	**A**	0	4	12	7	7	10	9	49	0	9	6	6	7	8	7	43
	B	11	0	13	10	12	12	12	70	6	0	4	7	8	9	11	45
	C	3	2	0	3	7	11	5	31	9	11	0	4	9	10	8	51
	D	8	5	12	0	10	12	9	56	9	8	11	0	12	10	10	60
	E	8	3	8	5	0	10	7	41	8	7	6	3	0	8	6	38
	F	5	3	4	3	5	0	6	26	7	6	5	5	7	0	8	38
	G	6	3	10	6	8	9	0	42	8	4	7	5	9	7	0	40
Toys	**A**	0	14	3	10	7	4	13	51	0	15	2	11	10	5	14	57
	B	1	0	1	2	3	0	6	13	0	0	0	2	2	2	4	10
	C	12	14	0	14	12	7	14	73	13	15	0	12	13	9	15	77
	D	5	13	1	0	4	2	15	40	4	13	3	0	6	2	14	42
	E	8	12	3	11	0	3	14	51	5	13	2	9	0	1	14	44
	F	11	15	8	13	12	0	13	72	10	13	6	13	14	0	13	69
	G	2	9	1	0	1	2	0	15	1	11	0	1	1	2	0	16
Total	**A**	0	18	15	17	14	14	22	100	0	24	8	17	17	13	21	100
	B	12	0	14	12	15	12	18	83	6	0	4	9	10	11	15	55
	C	15	16	0	17	19	18	19	104	22	26	0	16	22	19	23	128
	D	13	18	13	0	14	14	24	96	13	21	14	0	18	12	24	102
	E	16	15	11	16	0	13	21	92	13	20	8	12	0	9	20	82
	F	16	18	12	16	17	0	19	98	17	19	11	18	21	0	21	107
	G	8	12	11	6	9	11	0	57	9	15	7	6	10	9	0	56

4.5.1.2 Significance Analysis

After the subject consistency analysis, the second test was conducted on the experimental data to measure the significance of the algorithm scores. To achieve this, the test explained in Section 3.1.2.2 was applied to find m_c value similar to the significance evaluation of the first experiment.

For any screen type-scene combination analysis, the n and t values were equal to 15 and 7 respectively. These values were used in calculation of D values by using the Equation 3.14. D values of all tests are given in Table 4.10. χ^2 distribution with $(t - 1) = 6$ degrees of freedom at significance level $p = 0.05$ is 12.59. Therefore, all test results were found to be statistically significant similar to the significance analysis of the first experiment. For the aggregated results, n was set to 30 since the results of two scenes were combined. D values of aggregated results are shown in Table 4.11. The aggregated results were also found to be statistically significant in the same significance level.

For $n = 15$ and $t = 7$, m_c is equal to 15. This means that two scores with a difference not

Table 4.9: Subject consistency values for the second experiment. Camera columns represent the small camera screen and computer columns represent the standard LCD screen.

	Camera		**Computer**	
Sub. #	**Ametyst**	**Toys**	**Ametyst**	**Toys**
1	0.857	0.786	0.643	0.643
2	1.000	0.929	0.571	0.286
3	1.000	1.000	0.643	1.000
4	0.714	1.000	0.571	1.000
5	0.929	1.000	0.643	0.786
6	0.929	0.786	0.357	0.714
7	0.857	1.000	0.357	0.929
8	1.000	0.929	0.714	1.000
9	0.714	0.214	0.286	0.857
10	0.929	0.929	0.857	0.929
11	0.786	1.000	0.500	0.929
12	0.929	1.000	0.429	0.929
13	0.929	0.714	1.000	0.929
14	1.000	0.857	0.929	1.000
15	1.000	0.857	1.000	0.714
Avg.	0.905	0.867	0.633	0.843

less than 15 are statistically significant. For the aggregated results m_c is equal to 21. Based on these, the similarity groups for the individual scene-attribute combinations as well as the aggregated results for each attribute are shown in Figures 4.5, 4.6, 4.7, and 4.8. In these figures, the algorithms underlined by the same line belong to the same similarity groups. That is, the differences between them are not statistically significant.

Table 4.10: D values for the second experiment.

	Ametyst	**Toys**
Camera	51.2	134.63
Computer	14.781	145.52

Table 4.11: Aggregated D values for the second experiment.

Camera	**Computer**
29.295	84.227

4.5.1.3 Evaluation of Results

The second experiment was conducted on ametyst and toys scenes which have different characteristics such as lighting conditions and content. The resulting image scores are scene dependent. Since an operator which shows good performance in a scene may fail in other scene, it is difficult to find out a correlation between operator scores except the consistent performance of Mantiuk et al.'s operator.

In ametyst scene evaluation results have a small number of significance groups. This means, most of the algorithm performances can be considered to be the same for ametyst scene. Besides, the subject consistency values in ametyst scene are also small in the standard LCD test (see Table 4.9). In contrast, toys scene produced the most salient significance groups in both camera and standard display tests. Moreover, operator scores are consistent in between the standard and camera display evaluation results of toys scene.

Mantiuk et al.'s display adaptive TMO has the best performance in overall results since it proposes an HVS contrast response model leading to improved similarity performance of the resulting images. Besides, the algorithm takes the ambient light and the maximum and minimum luminance values of the display devices as parameters. Therefore, the algorithm coped with the display device and viewing condition change by producing the final images according to the device and ambient light modifications performed in the second experiment.

For most of the cases, the similarity performance of the camera is related to the contrast performance. If an operator has good contrast scores, then it generally performs well in small screen devices since high contrast tolerates the detail loss in small screen devices. As a result, it makes the displayed image more similar to the real scene. This can be concluded by examining ametyst performance of Mertens et al.'s EFA results.

Drastic change in an algorithm score may result from the low amount of overall brightness or artifacts. Zhang and Cham's toys scene produced dark image and Li et al.'s result of the same scene has halo artifact in sharp contrast transition (see Figure 4.12). Therefore, they are significantly worse than the other algorithms in toys scene. On the other hand, Shan et al.'s bright image result has one of the best score in the same scene although it has the worst overall performance in the other quality evaluations. An interesting observation is that, Goshtasby's result of the same scene has an average score although it has seam artifacts in the background (see Figure 4.13). Considering the Goshtasby's implementation, toys scene has large flat regions and this led to large use of block size in the selection of the well exposed regions from the exposure sequence. Large block usage in blending operation caused the seam artifact in the background of the image.

The results show that, for the small screens, the differences between the algorithm's scores are minimized. Therefore, it can be argued that, simpler and computationally efficient versions of these algorithms can be used when preparing HDR images for display on a small display device. An explanation for this based on the human contrast sensitivity is provided in the following chapter.

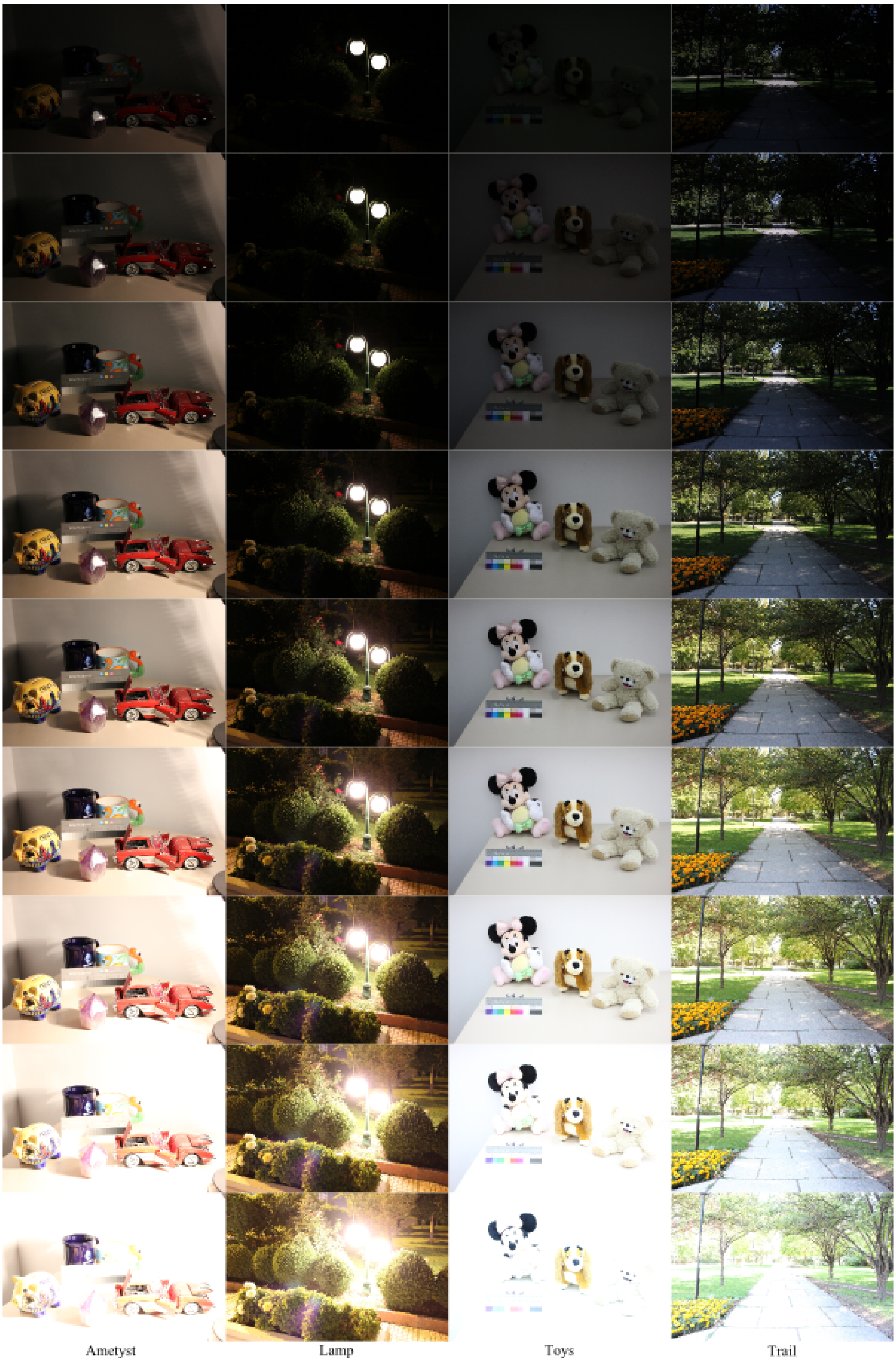


Figure 4.1: Exposure sequence of input scenes in JPEG format.

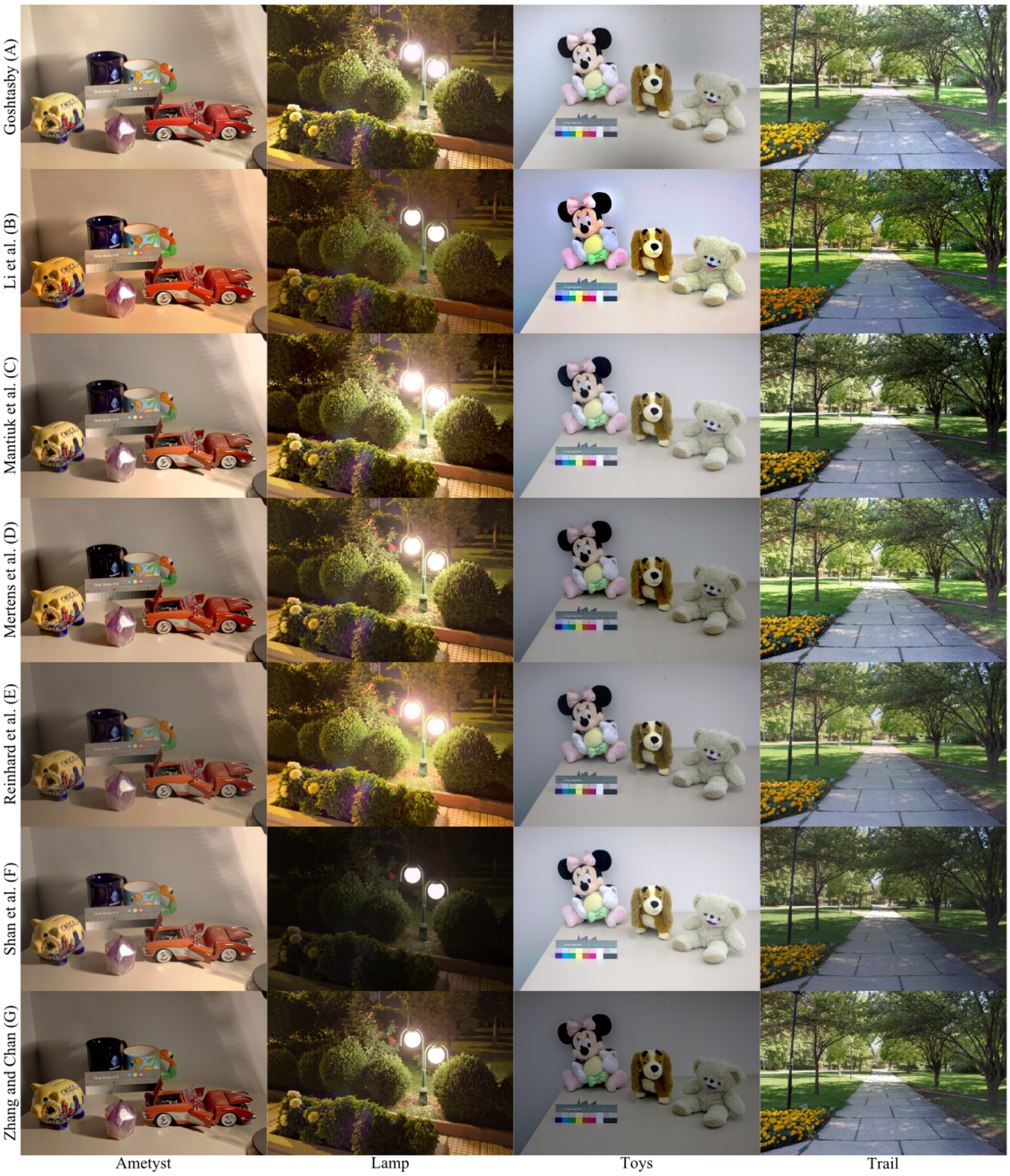


Figure 4.2: Outputs of all algorithms for all images.

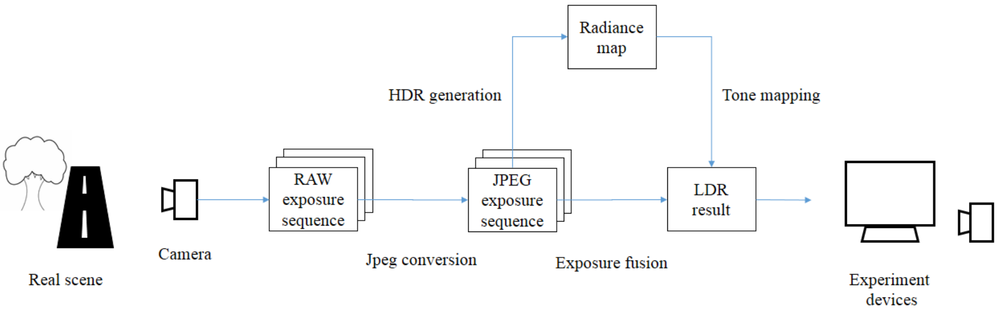


Figure 4.3: Image acquisition chart.

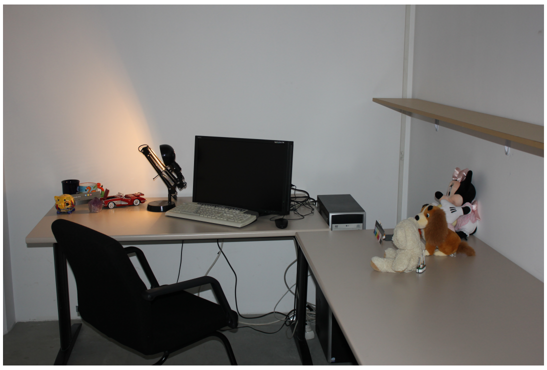


Figure 4.4: A photograph from the experiment room.

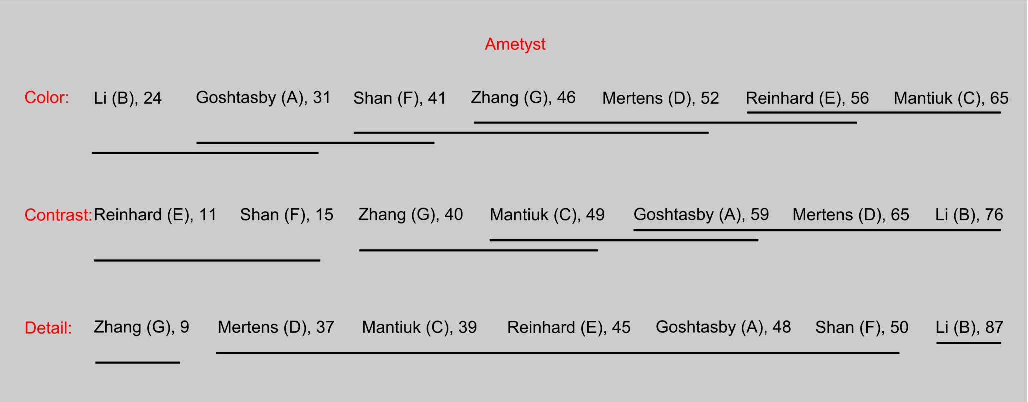


Figure 4.5: Significance groups of ametyst scene for the first experiment.

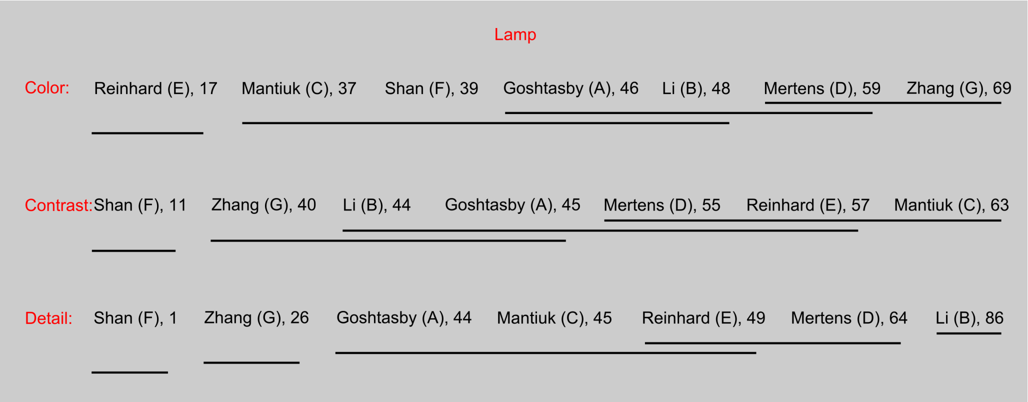


Figure 4.6: Significance groups of lamp scene for the first experiment.

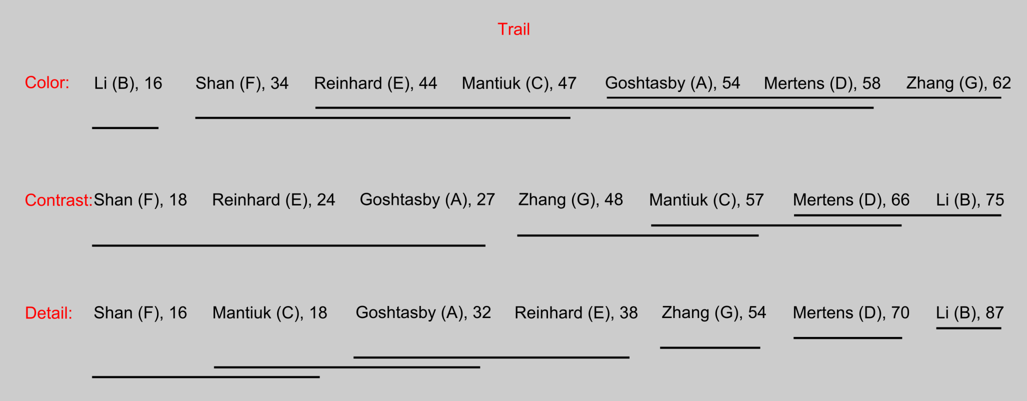


Figure 4.7: Significance groups of trail scene for the first experiment.

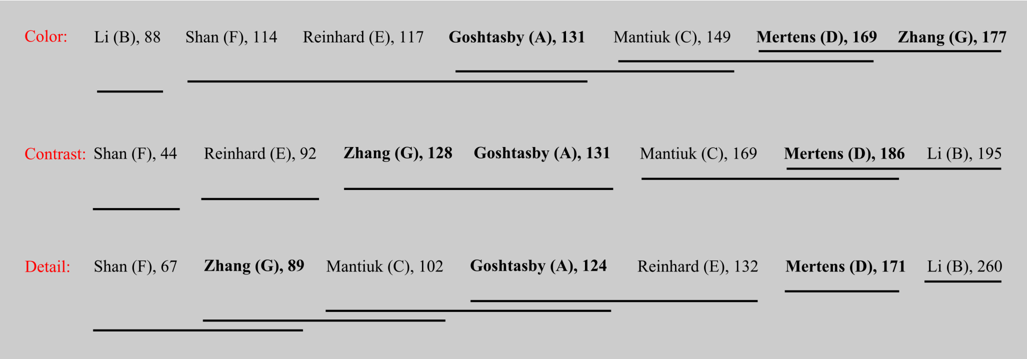


Figure 4.8: Significance groups of aggregated results for the first experiment. EFA names are written in bold.

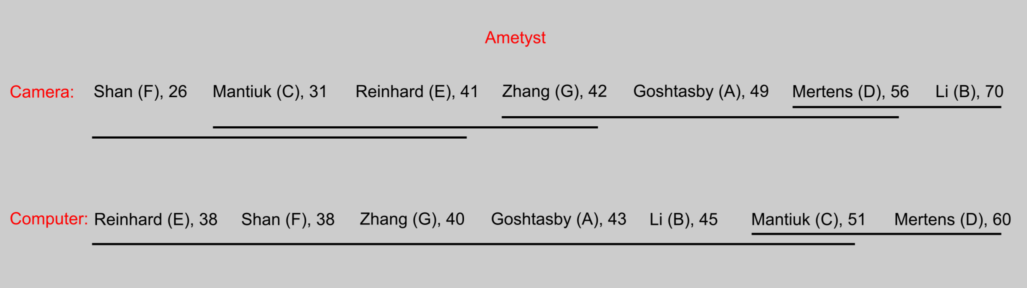


Figure 4.9: Significance groups of lamp scene for the second experiment.

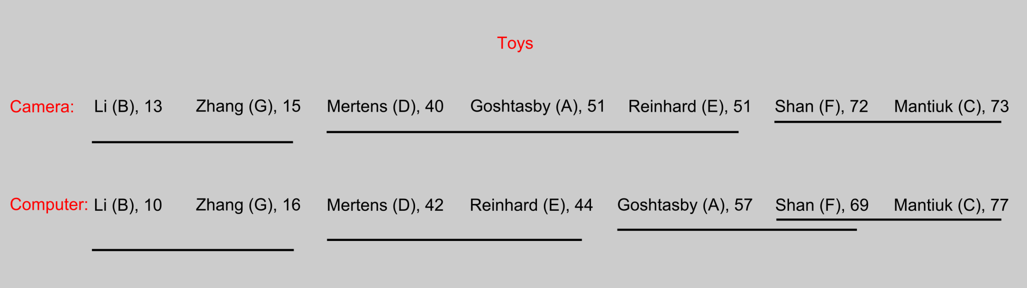


Figure 4.10: Significance groups of trail scene for the second experiment.

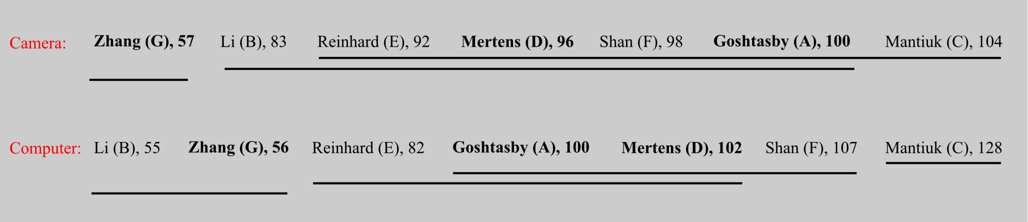


Figure 4.11: Significance groups of aggregated results for the second experiment. EFA names are written in bold.

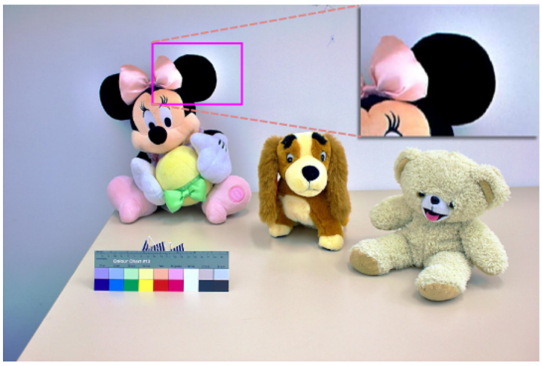


Figure 4.12: Halo artifact in toys scene result of Li et al.'s TMO.

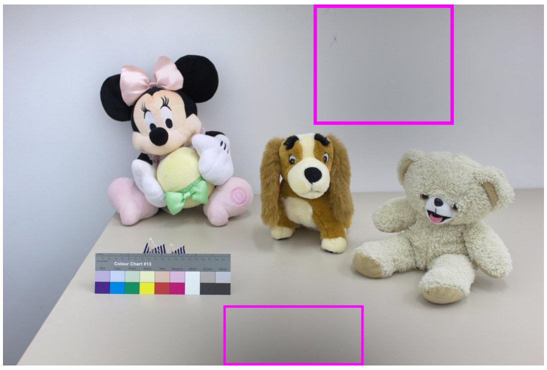


Figure 4.13: Seam artifact in toys scene result of Goshtasby's TMO.

CHAPTER 5

DISCUSSION

In this chapter, further analysis on the results of subjective experiments are made based on the main findings of the statistical evaluation. The details of these findings are highlighted by introducing some reasonings and models. Also relationships to earlier validation studies are discussed.

Three main observations can be made while examining the evaluation results: First, Li et al.'s TMO performs better in contrast and detail attribute evaluations and Mantiuk et al.'s TMO produces images that have a high quality of similarity performances on the standard display. Second, colors of EFA results are more natural than that of TMOs. Third, most algorithms have similar scores in the small screen evaluation.

The reason of the first two findings can be clarified by explaining the overall mechanism of the EFA and TMO algorithms. TMOs process the radiance map which is obtained from the exposure sequence. The radiance map is a linearized data generated by recovering the camera response function (CRF) of the capturing device and using the exposure times of the bracketed sequence. Thus, the resulting HDR image is the exact representation of the real world scene in a linear color space. EFAs, on the other hand, skip the HDR image reproduction steps and directly fuse the multiple exposures into one final image.

EFAs discard the radiance data and attempt to form visually appealing images by weighting the exposure sequence. However, TMOs operate on the linear HDR data that refers to the pixel-wise lighting distribution of the real scene in a color space. Therefore, the similarity performances of the tone mapped images are expected to be better than that of fused images.

Both TMO and EFA pipelines start with a set of exposures as illustrated in 5.1. The operation which affects the detail performance of EFAs is the exposure weight map generation. EFAs blend the exposures' weights in order to smooth the boundaries of selected well-exposed regions or remove the seam artifacts of the fused image. This weight map modification causes small details being smoothed in the final image. In the tone mapping pipeline, the exposures are weighted in the HDR generation phase before the tone mapping operation. The linear hat function ω in Equation 4.1 is used for weighting the exposures in order to reduce the contribution of over and under exposed pixels to the HDR image. ω is not a spatially- varying function so it does not cause a deterioration in the final image. Therefore, the detail loss is expected to be lower in TMOs, so TMOs perform better than EFAs by means of detail attribute.

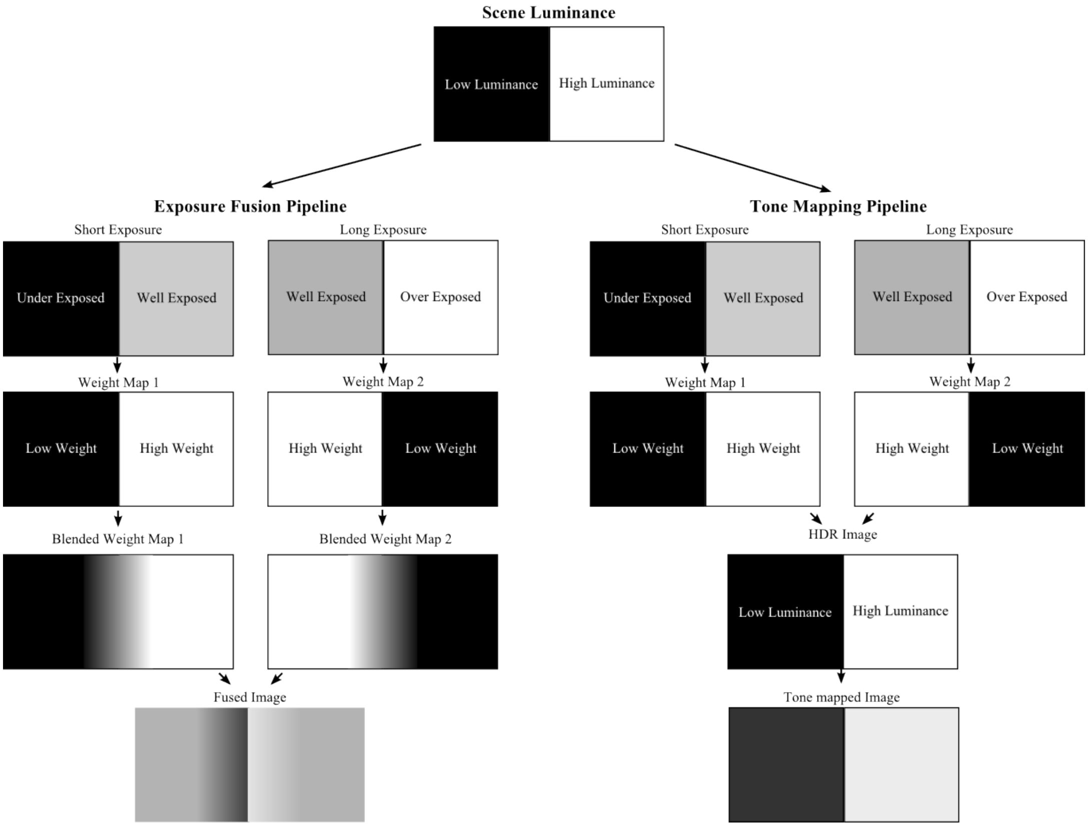


Figure 5.1: The exposure fusion and tone mapping imaging pipeline. An HDR scene which has a dark region on the left side and a bright region on the right side is captured with 2 exposure times. Weight maps used in the exposure fusion pipeline is blended, causing detail loss in the final image. Weight maps used in the tone mapping pipeline are not blurred.

EFAs directly combine well-exposed pixels by discarding the exposure times of the bracketed sequence. Thus, it is possible that both dark and bright regions is mapped to the similar luminance values. Therefore, the EFA results have a low degree of contrast compared to the TMO results.

EFAs work on the color space of the individual exposures which is designed to be visually appealing to humans. On the other hand, the HDR generation accumulates the color values of each exposure into one image by using the camera response curve, leading the correlation among the color values in this visually appealing space being lost.

A second issue for tone mapping operators is the Hunt effect. According to the Hunt effect colorfulness is incremented as the luminance is increased [19]. Compressing only the luminance does not preserve colors. The colors are mostly reproduced by the following equation:

$$C_o = \left(\frac{C_i}{L_i}\right)^s L_o, \tag{5.1}$$

where C_i represents the input color channels (R_i, G_i, B_i), C_o represents the output color channels, L_i is the input luminance, L_o is the output color luminance, and s is the saturation parameter. The problem of this calculation is the difficulty in setting this parameter in an automatic way. Although some methods exist that suggest a solution [35], the proposed solutions are limited only for gamma-like luminance compressions. Therefore, it could be expected that in this study, TMOs produced less natural colors compared to EFAs.

In the second experiment, the angular sizes for the camera and standard LCD display evaluation were different. The approximate angular sizes of the central pixel for the standard and camera screen were 0.022^o and 0.017^o respectively. These angular size values led to spatial frequency of 45 cycles per degree for the standard display and 59 cycles per degree for the camera display. The contrast sensitivity function (CSF) for modeling the HVS contrast sensitivity introduced by Mannos and Sakrison [33] is as follows:

$$A(f) = 2.6(0.0192 + 0.114f)\exp(-0.114f^{1.1}), \tag{5.2}$$

where f is the spatial frequency. The CSF plot given in Figure 5.2 shows that the CSF value has the maximum value for $f = 8\ cycles/degree$ and it decreases for the higher values of f. Therefore, it can be concluded that the contrast sensitivity is higher in the standard display than the camera display. In the second experiment, contrast sensitivity decreased in the camera screen evaluation and this prevented subjects to distinguish the similarity performances of the images. Therefore, evaluation results of the camera screen evaluation are similar to each other for all TMOs and EFAs.

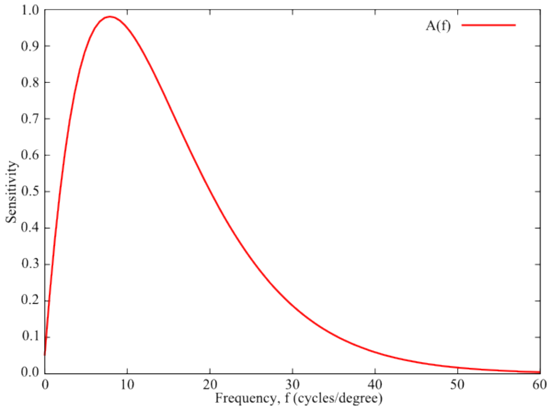


Figure 5.2: The contrast sensitivity function of Mannos and Sakrinson [33].

It is impractical to make a detailed comparison between the results of any other different subjective evaluation studies although a common operator, Reinhard et al.'s photographic tone reproduction [40] is included in some of the earlier evaluation studies since the experiment conditions and the selected scenes are not the same. However, it is worth to indicate some of the relationship between this study and the other subjective evaluation results.

The photographic TMO performed well in [6, 16, 27, 29, 50, 55], but it was generally outperformed by Li et al.'s TMO [32] and Mantiuk et al.'s TMO [34] operators. Also mostly the EFAs were found to be generally better than the photographic TMO. Therefore, it can be concluded that these algorithms which have not been evaluated by any validation study may perform better than the other algorithms which are tested with the photographic TMO in the earlier subjective evaluations.

Finally, the importance of parameter settings should be emphasized. All of the tested algorithms include a set of parameters and using different settings can drastically change their output. Therefore, the presented results should be considered valid only for the parameter settings used in this study (see Table 4.2).

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this thesis, a subjective validation study was conducted with two main techniques in HDR imaging, TMOs and EFAs in order to assess the reproduced image quality. 4 TMOs and 3 EFAs were selected and 4 different scenes were captured. The exposure sequence and the corresponding HDR images were reproduced by these selected algorithms. The reproduced images were evaluated by two different experiments of pairwise comparison in an isolated experiment room. In the first experiment the subjects were asked to evaluate that 7 algorithm results of each 3 scenes by means of color, contrast, and detail attributes in the standard LCD display. In the second experiment, the subjects were asked to compare that 7 algorithm results of each 2 scenes by means of similarity attribute in the standard LCD display and camera display. There were 15 participants in each experiments. The results were analyzed by Kendall's consistency analysis [26] and multiple comparison test in [44].

Significant findings are observed in the statistical evaluation of the experimental results. It is concluded that color, contrast, and detail performance of EFAs are surpassed by the best TMO operator of that attribute evaluation. It is also stated that color reproduction of EFAs is better compared to that of TMO. This observation shows that TMOs needs to be improved for better color reproduction. The screen size and the viewing distance is effective in the similarity performance. In a small screen, the ranking of the algorithm performances changes as well as the differences between the algorithm performance scores are minimized due to the contrast sensitivity change of the human observer. This suggests that a simpler and more efficient algorithm can be used in the embedded program of the camera in order to receive an instant feedback of HDR capture which is expected to be a common feature for the future cameras.

Further improvements can be included in the study in order to conduct the experiments in a more controlled manner. For example, the experiments started with the same scene-attribute or scene-image device combinations. This led to less amount of overall consistency value for the first combination since subjects were learning the experiment objectives in the first combination. Instead of starting with the same combination, each subject can start to the experiments with a different combination in order to distribute the influence of the most inconsistent part of the experiment to the other image combination results. Also subjects can be trained with a brief pilot study before the experiment for covering the learning phase before the first evaluated combination and for better understanding of experiment objectives. The angular size can be controlled by fixing subject to a certain distance from the display device, for instance by using head-mounting. Moreover, the experiment can be extended by increasing the number of scenes and algorithms.

REFERENCES

[1] Psychtoolbox Wiki : HomePage. http://psychtoolbox.org/HomePage, 2011. [Online; accessed 22-August-2013].

[2] iPhone 4 Tech Specs. http://www.apple.com/iphone/iphone-4/specs.html, 2013. [Online; accessed 23-August-2013].

[3] A. Adams. *The Negative*. Boston, Little, Brown & Company, 1997.

[4] A. Adams and R. Baker. *The Camera*. New Ansel Adams Photography Series, Book 1. New York Graphic Society, 1980.

[5] A. Adams and R. Baker. *The Print*. The New Ansel Adams photography series. Little Browm, 1994.

[6] M. Čadík, M. Wimmer, L. Neumann, and A. Artusi. Image attributes and quality for evaluation of tone mapping operators. In *National Taiwan University*, pages 35–44. Press, 2006.

[7] M. Ashikhmin and J. Goyal. A reality check for tone-mapping operators. *ACM Trans. Appl. Percept.*, 3(4):399–411, Oct. 2006.

[8] R. A. Bradley. Rank analysis of incomplete block designs: Ii. additional tables for the method of paired comparisons. *Biometrika*, 41(3/4):pp. 502–537, 1954.

[9] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):pp. 324–345, 1952.

[10] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. *Communications, IEEE Transactions on*, 31(4):532–540, 1983.

[11] M. Čadík, M. Wimmer, L. Neumann, and A. Artusi. Evaluation of hdr tone mapping methods using essential perceptual attributes. *Computers & Graphics*, 32(3):330–349, June 2008.

[12] K. Chiu, M. Herf, P. Shirley, S. Swamy, C. Wang, and K. Zimmerman. Spatially nonuniform scaling functions for high contrast images. In *In Proceedings of Graphics Interface '93*, pages 245–253, 1993.

[13] C. Connolly and T. Fleiss. A study of efficiency and accuracy in the transformation from rgb to cielab color space. *Image Processing, IEEE Transactions on*, 6(7):1046–1048, 1997.

[14] S. Daly. Digital images and human vision. chapter The visible differences predictor: an algorithm for the assessment of image fidelity, pages 179–206. MIT Press, Cambridge, MA, USA, 1993.

[15] H. A. David. Tournaments and paired comparisons. *Biometrika*, 46(1/2):pp. 139–149, 1959.

[16] F. Drago, W. Martens, K. Myszkowski, and H.-P. Seidel. Perceptual evaluation of tone mapping operators with regard to similarity and preference. Research Report MPI-I-2002-4-002, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, August 2002.

[17] F. Durand and J. Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. *ACM Trans. Graph.*, 21(3):257–266, July 2002.

[18] J. DURBIN. Incomplete blocks in ranking experiments. *British Journal of Statistical Psychology*, 4(2):85–90, 1951.

[19] M. Fairchild. *Color Appearance Models*. The Wiley-IS&T Series in Imaging Science and Technology. Wiley, 2005.

[20] R. Fattal, D. Lischinski, and M. Werman. Gradient domain high dynamic range compression. *ACM Trans. Graph.*, 21(3):249–256, July 2002.

[21] J. Ferwerda. Elements of early vision for computer graphics. *Computer Graphics and Applications, IEEE*, 21(5):22–33, 2001.

[22] R. Ganga and T. A. Ramena. Literature survey for fusion of multiple-exposure images. *"International Journal of Engineering Research & Technology"*, 2, January 2013.

[23] A. Goshtasby. Design and recovery of 2-d and 3-d shapes using rational gaussian curves and surfaces. *International Journal of Computer Vision*, 10(3):233–256, 1993.

[24] A. A. Goshtasby. Fusion of multi-exposure images. *Image Vision Comput.*, 23(6):611–618, June 2005.

[25] K.-H. Jo and A. Vavilin. {HDR} image generation based on intensity clustering and local feature analysis. *Computers in Human Behavior*, 27(5):1507 – 1511, 2011. <ce:title>2009 Fifth International Conference on Intelligent Computing</ce:title> <ce:subtitle>ICIC 2009</ce:subtitle> <xocs:full-name>2009 Fifth International Conference on Intelligent Computing</xocs:full-name>.

[26] M. G. Kendall and B. B. Smith. On the method of paired comparisons. *Biometrika*, 31(3/4):pp. 324–345, 1940.

[27] J. Kuang, H. Yamaguchi, G. M. Johnson, and M. D. Fairchild. Testing hdr image rendering algorithms. In *Color Imaging Conference*, pages 315–320. IS&T - The Society for Imaging Science and Technology, 2004.

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

[29] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen. Evaluation of tone mapping operators using a high dynamic range display. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, pages 640–648, New York, NY, USA, 2005. ACM.

[30] S. Li. Markov random field models in computer vision. In J.-O. Eklundh, editor, *Computer Vision — ECCV '94*, volume 801 of *Lecture Notes in Computer Science*, pages 361–370. Springer Berlin Heidelberg, 1994.

[31] S. Li, J.-Y. Kwok, I. Tsang, and Y. Wang. Fusing images with different focuses using support vector machines. *Neural Networks, IEEE Transactions on*, 15(6):1555–1561, 2004.

[32] Y. Li, L. Sharan, and E. H. Adelson. Compressing and companding high dynamic range images with subband architectures. *ACM Trans. Graph.*, 24(3):836–844, July 2005.

[33] J. Mannos and D. Sakrison. The effects of a visual fidelity criterion of the encoding of images. *Information Theory, IEEE Transactions on*, 20(4):525–536, 1974.

[34] R. Mantiuk, S. Daly, and L. Kerofsky. Display adaptive tone mapping. *ACM Trans. Graph.*, 27(3):68:1–68:10, Aug. 2008.

[35] R. Mantiuk, R. Mantiuk, A. M. Tomaszewska, and W. Heidrich. Color correction for tone mapping. *Comput. Graph. Forum*, 28(2):193–202, 2009.

[36] T. Mertens, J. Kautz, and F. V. Reeth. Exposure fusion. *Computer Graphics and Applications, Pacific Conference on*, 0:382–390, 2007.

[37] N. Mitchell. *Photographic Science*. Wiley, 1984.

[38] T. Mitsunaga and S. K. Nayar. Radiometric self calibration. In *1999 Conference on Computer Vision and Pattern Recognition (CVPR 99), 23-25 June 1999, Ft. Collins, CO, USA*, pages 1374–1380. IEEE Computer Society, 1999.

[39] S. Paris and F. Durand. A fast approximation of the bilateral filter using a signal processing approach. *Int. J. Comput. Vision*, 81(1):24–52, Jan. 2009.

[40] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda. Photographic tone reproduction for digital images. *ACM Trans. Graph.*, 21(3):267–276, July 2002.

[41] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. The Morgan Kaufmann Series in Computer Graphics. Elsevier Science, 2005.

[42] Q. Shan, J. Jia, and M. S. Brown. Globally optimized linear windowed tone mapping. *IEEE Transactions on Visualization and Computer Graphics*, 16(4):663–675, 2010.

[43] R. Shen, I. Cheng, J. Shi, and Basu. Generalized Random Walks for Fusion of Multi-Exposure Images. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, X(X):1–28, May 2011.

[44] T. H. Starks and H. A. David. Significance tests for paired-comparison experiments. *Biometrika*, 48(1/2):pp. 95–108, 1961.

[45] T. G. Stockham. Image processing in the context of a visual model. *Proceedings of the IEEE*, 60(7):828–842, 1972.

[46] L. Stroebel. *Basic Photographic Materials and Processes*. Focal Press, 2000.

[47] B. G. Tabachnick and L. S. Fidell. *Using Multivariate Statistics (5th Edition)*. Allyn & Bacon, 5 edition, Mar. 2006.

[48] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, pages 839–, Washington, DC, USA, 1998. IEEE Computer Society.

[49] J. Tumblin and H. Rushmeier. Tone reproduction for realistic images. *Computer Graphics and Applications, IEEE*, 13(6):42–48, 1993.

[50] C. Urbano, L. Magalhães, J. Moura, M. Bessa, A. Marcos, and A. Chalmers. Tone Mapping Operators on Small Screen Devices: An Evaluation Study. *Computer Graphics Forum*, 29(8):2469–2478, 2010.

[51] B. Wandell. *Foundations of vision*. Sinauer Associates, Incorporated, Sunderland, Mass., 1995.

[52] G. Ward and M. Simmons. Subband encoding of high dynamic range imagery. In *Proceedings of the 1st Symposium on Applied perception in graphics and visualization*, APGV '04, pages 83–90, New York, NY, USA, 2004. ACM.

[53] H. Wilson. A transducer function for threshold and suprathreshold human vision. *Biological Cybernetics*, 38(3):171–178, 1980.

[54] Z. Xiang. Color image quantization by minimizing the maximum intercluster distance. *ACM Trans. Graph.*, 16(3):260–276, July 1997.

[55] A. Yoshida, V. Blanz, K. Myszkowski, and H.-P. Seidel. Perceptual evaluation of tone mapping operators with real-world sceness. In B. E. Rogowitz, T. N. Pappas, and S. J. Daly, editors, *Human Vision and Electronic Imaging X, IS&T/SPIE's 17th Annual Symposium on Electronic Imaging (2005)*, volume 5666 of *SPIE Proceedings Series*, pages 192–203, San Jose, USA, January 2005. SPIE.

[56] W. Zhang and W.-K. Cham. Gradient-directed composition of multi-exposure images. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 0:530–536, 2010.

Appendix A

INSTRUCTIONS FOR EXPERIMENT ONE

In this experiment, you'll be presented with multiple image pairs. For each pair, you'll be asked to choose the image that you prefer according to three different criteria, namely:

1. Naturalness of colors
2. Sensation of contrast
3. Visibility of details

For (1), please choose the image whose colors appear more natural to you. For (2), please choose the image that appears to have more contrast. For (3), please choose the image where details are more visible.

The experiment should take about 30-40 minutes assuming that you spend 10-15 seconds on each image pair. To avoid fatigue, please do not spend too much on each stimulus. Please note that you can terminate the experiment if you feel any discomfort.

Thank you for your participation. Please ask the experimenter if you have any questions.

Appendix B

INSTRUCTIONS FOR EXPERIMENT TWO

In this experiment, you'll be presented with multiple image pairs in two different display devices. For each pair, you'll be asked to choose the image that you think is more similar to its real world version.

For the desktop monitor, please switch between the images using the keys "1" and "2" on the numpad. When you make your decision, please press the "enter" key to continue with the next pair.

For the camera LCD monitor, please switch between the images using the camera's "left" and "right" buttons. When you make your decision, please tell the name of that image to the experimenter. You can then move on to the next pair (each pair is separated by a gray image).

The experiment should take about 15-20 minutes assuming that you spend 10-15 seconds on each image pair. To avoid fatigue, please do not spend too much on each stimulus. Please note that you can terminate the experiment if you feel any discomfort.

Thank you for your participation. Please ask the experimenter if you have any questions.