https://doi.org/10.2352/issn.2169-2629.2019.27.48

Tone Mapping Operators Evaluation Based on High Quality Reference Images

Imran Mehmood, Muhammad Usman Khan, Ming Ronnier Luo, Muhammad Farhan Mughal; State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, Hangzhou, China.

Abstract

High Dynamic Range (HDR) imaging applications have been commonly placed recently. Several tone mapping operators (TMOs) have been developed which project the HDR radiance range to that of a display. Currently, there is no agreement on a technique for evaluation of tone mapping operators. The goal of this study is to establish a method based on reference images to evaluate the TMOs. Two psychophysical experiments were carried out for the evaluation of tone mapping operators. In the first experiment, a set of high quality images were generated to possess right extents of image features including contrast, colourfulness and sharpness. These images were further used in the second experiment as reference images to evaluate different TMOs. It was found Reinhard's photographic reproduction based on local TMO gave an overall better performance. CIELAB(2:1) and S- CIELAB metrics were also used to judge colour image quality of the same TMOs. It was found that both metrics agreed well with the visual results.

Introduction

Nowadays, HDR technology is becoming popular. The original idea of the HDR is to map the luminance range of capturing scene to that of display device while preserving the contrast of the scene [1]. Several TMOs have been proposed over the years [2-8]. These TMOs are based on human vision theory, and inspired by various fields such as, photographic reproduction and image processing. In the process of mapping, some information may loss and the results may not have good quality. However, there is a lack of means for the evaluation of TMOs. There are two types of evaluation, i.e. the subjective assessment based upon psychophysical experiment and the objective measurement based on image attributes. The evaluation of the operators could be further divided into the external reference image based where the tone mapped image is compared with the reference image and the non-reference image based where the image is compared with the observer memory. There are many studies in literature on subjective evaluation of the quality of TMOs. Subjective evaluation of various TMOs have been typically studied [9] including preference and similarity. Ledda et al. [10] compared 8 TMOs. Čadík et al. [11] conducted an experiment using an HDR display and the tone mapped images were compared to the HDR displayed images. Some studies [12] investigated different TMOs by asking observers to judge in terms of image attributes i.e. overall brightness, quality, contrast and colours.

The image quality metrics (IQMs) are classified into three types: full Reference, reduced reference and no reference. In full reference IQMs, the reference image is available for comparison. In reduced reference IQMs, a specific set of features of reference image is available for quality measurement. In no reference IQMs, the information of the reference image is not available. The scope of this research is focused on full reference IQM.

Currently, there is a lack of studies based on reference images. There are many image database [13, 14] but HDR reference images have not been developed. Therefore in this study one objective was to produce high quality images for evaluation of TMOs. Two psychophysical experiments were conducted. Experiment 1 was aimed to produce a set of reference images. Experiment 2 was an evaluation of the performance of some commonly used TMOs using the reference images obtained in Experiment 1. CIELAB(2:1) [15] and spatial extension of CIELAB called S-CIELAB [16] were also used to evaluate these TMOs against the reference image. Their results are also be compared with visual results.

Methodology

The method used in this study includes generation of image renderings for high quality image production by varying three image attributes, contrast, sharpness and colourfulness. Reinhard et al. [3] developed two versions of the photographic tone reproduction operator, denoted as the global operator and the local operator. It was found in our initial study that local operator performed well so that it was chosen to be used as the base line TMO in Experiment 1. The RGB radiance map was first transformed to XYZ coordinates and its luminance was non-linearly compressed using the Reinhard local operator. Subsequently, the XYZ values were converted to CIELAB colour space for rendering of each scale.

Experiment 1: Reference Image Development

Experiment Design

Firstly, some high quality HDR images were selected from the Fairchild's database at Rochester Institute of Technology. Each image was then rendered according to image characteristics of contrast and colourfulness, as reported by [17]. Sharpness is also desired and is defined as the contrast at the edges. Therefore, the three image scales, contrast, sharpness and colourfulness, were used for rendering high quality images. Different contrast enhancement techniques [18] have been proposed. The most common and effective method of contrast enhancement is the histogram equalization method. For various contrast rendering, adaptive histogram equalization (AHE) method [19] with uniform distribution is very popular. However, the convectional histogram equalization method usually results in over contrast. The variation of AHE proposed by Reza [20] with clip limits was used to control the over enhancement of contrast. Therefore in the contrast renderings clip limit histogram equalization (CLAHE) method was used.

For image sharpness renderings, unsharp masking technique was used. Let the original image is denoted by $f(x,y)$ and its Guassian blurred image is denoted by $\bar{f}(x,y)$ then the masks of the images is first obtained as

$$g_{mask}(x,y) = f(x,y) - \bar{f}(x,y) \quad (1)$$

The mask $g_{mask}(x,y)$ is then added to the original image i.e.

$$g(x,y) = f(x,y) + g_{mask}(x,y) \quad (2)$$

The processed image will produce a sharpened image as compared to the original image.

Colourfulness is one of the measures of colour quality. Boosting and enhancing colourfulness are operations often performed for improving image aesthetics. The Chroma channel of CIELAB $L^*C_{ab}^*h_{ab}$ space was used to enhance the colourfulness linearly here.

For each of the three scales, five rendered images were produced including one original image, one most pleasing, one acceptable and two images in between the latter two types. The acceptable image was chosen visually such that after this rendering the images would appear unnatural or have defects. For example, in case of contrast rendering, one original contrast was used, one best version of the CLAHE image, and one version with highest possible contrast. The other two images were chosen by comparing with the best version. For contrast enhancement, the clip limit of the CLAHE was adjusted to obtain different renderings. For sharpness renderings, the standard deviation of the Gaussian distribution was applied to achieve various blurring or sharpening effects. The colourfulness variation was used straightforward by linearly increasing the CIELAB Chroma.

For generating these rendered images, the HDR image database from Rochester Institute of Technology (RIT) was used [21]. Ten natural images with quite high dynamic range were selected. As mentioned earlier, each scale had five renderings, therefore 125 rendered images were produced for one image and 1,250 images were processed in total. For analysis of intra-observer variability, 20% of the images were used. In total, 1,500 images were assessed in Experiment 1.

The experiment was conducted on a display which is shown in Figure 1. Each observer was asked to judge the displayed image either 'High Quality' or 'Low Quality'. After made the decision, s/he clicked the 'forward' button to progress into the next image or the 'backward' button to redo the judgment of the present image.

Procedure

The images were displayed on a NEC PA302W AH-IPS LCD display. It was located in a dark room. The wall reflectance of the dark room was approximately 4%. The peak luminance of the peak white of the display was set at CIE D65 chromaticity at a luminance of 287 cd/m^2. For evaluation of spatial uniformity,

Figure 1: *Experiment 1 setup*

the display was divided into 3 by 3 segments and the mean colour difference calculated between the center and each segment was 1.21 ΔE^*_{ab}. The GOG model [22] was implemented for display characterization and gave a performance of 0.64 with a range from 0.37 to 1.66 calculated from the 24 colours of the Macbeth ColourChecker chart. The display was located at 45 cm away from the observer's eyes. Each observer passed the Ishihara Test to ensure that they had normal colour vision.

Twenty observers, 14 males and 6 females, participated in the experiment. They had a mean age of 25.7 and a standard deviation of 3.54**.** Each observer did 1,500 judgments.

Each observer sit in front of the display and did the Ishihara Test. In the experiment observers of different disciplines participated therefore the observers were presented instructions and a training session was provided before carried out the experiment. In the training session, the observers were shown three renderings of each scale one at a time to understand the effects of each scale on the image. Before displaying the experimental images, the observers were displayed a gray image for 30 seconds to adapt in the environment. The images presented to each observer were presented randomly. The experiment took 100 minutes to complete and it was conducted in two sessions of 50 minutes. In total, 30,000 judgments i.e. 5 (contrast) × 5 (sharpness) × 5 (colourfulness) × 10 (images) + 250 (repeatable images)) ×20 (observers) were made.

Experiment 2: TMO Evaluation

Experiment Design

As mentioned earlier, the purpose of Experiment 1 was to produce reference images for judging the quality of HDR images. Experiment 2 was to evaluate TMOs quality. Five commonly used TMOs, Reinhard photographic tone reproduction local and global operators [3], Drago's adaptive logarithm mapping [2], Schilik's quantization method [5] and Reinhard and Devlin's photoreceptor physiology based dynamic range reduction [8] were implemented. The parameters of each TMO were tuned to produce visually satisfactory versions based on the authors. The best versions were selected on the same display under the real experimental viewing condition. For TMO evaluation same ten images from RIT database were used in tone mapping. Ten images tone mapped by five TMOs produced 50 images. These tone mapped images are named as test images. The ten high quality image obtained in Experiment 1 (discussed later) were used as reference images. For observer variability 20% images were repeated and total number of images for Experiment 2 were 72.

Figure 2 shows the experimental layout. A six-point categorical scale was defined for ranking each image with respect to the reference image. The observer were asked to assess the image difference with respect to the reference image. The six categories were 1) "No Difference", 2) "Just Noticeable Difference", 3) "Small Difference", 4) "Acceptable difference", 5) "Large Difference" and 6) "Extremely Large Difference".

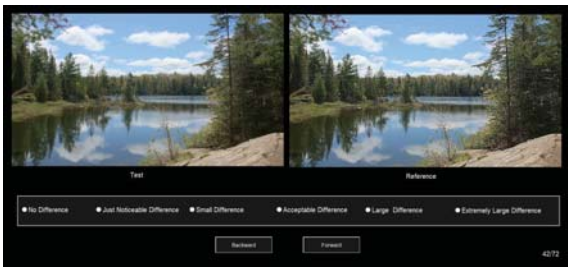

Figure 2: *Experiment 2 Setup*

Procedure

The experiment was conducted in the same display conditions as mentioned in Experiment 1 and similar procedure was adapted. The observer was presented a pair of reference and test images. The positions of both reference and test images were interchanged randomly. The observers were asked to make judgment in the 1 to 6 scale. Each observer took about 15 minutes to complete the experiment.

In total, 20 observers, 15 males and 5 females, took part with a mean age of 27.35 and standard deviation of 8.6. In total, 1,440 assessments were made.

Results and Discussion

Observer Variability

To calculate the intra-observer variability for Experiment 1, the number of wrong decision (WD) were calculated from the repeatability data included in the experiment. If an observer selected different choice for the same image when it was repeated, it was called WD. To calculate observer variability in terms of WD, number of WDs were counted and divided by the total number of decisions. The worst observer had 0.30 and the best observer had 0.14 variability with a mean of 0.19 and standard deviation (SD) of 0.028. For inter-observer variability the coefficient of variation (CV) was calculated. The worst and best values were 0.21 and 0.30 CV respectively with mean of 0.26 CV. The mean and SD are not large which shows that the observer were reliable and the collected data can be used with confidence.

The Inter-observer variability and intra-observer variability for Experiment 2 were calculated using the coefficient of variation to represent the consistency and repeatability of observers. For inter observer variability the worst and best values were 0.20 and 0.28 CV respectively with mean of 0.28 CV. For intra-observer variability, the worst observer has 0.16 CV and the best observer had 0.13 CV with a mean of 0.15 CV.

Selection of Reference Images

The raw data form Experiment 1 was converted to Z-score. Each image had 125 Z-scores and total were 1250 Z-scores. The Z-score had positive and negative values. If more than 50% observer gave high quality rating to a rendered image then its value was positive and if less than 50% observers gave ratings then Z-score was negative. The rendered images with positive Z-score were considered as high quality image and images with negative Z-score were discarded considering low quality images. On the basis of this criteria, only 158 out of 1250 rendered images were found high quality. This set of 158 images was sorted image wise and divided into 10 non-equal similar image bins. The image from each bin with highest Z-score was selected as reference image. Figure 3 shows four reference and tone mapped images including the Reinhard's local images (without renditions). It can be seen that the reference images had higher contrast, sharpness and colorfulness as compared to their original tone mapped images from various TMOs including Reinahrd's local TMO. Looking at the tone mapped images in Figure 3, it is clear that contrast and sharpness of the original images were very low. The reference images shown in Figure 3 were more colorful as compare to the original images. Furthermore, it was observed from the Z-scores that any original image was not selected as high quality image. This indicates that observers did not liked the straightforward tone mapped images.

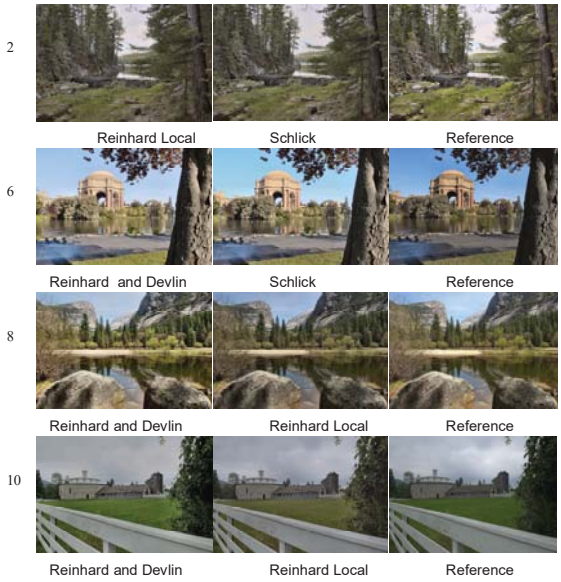

Figure 3: *Tone mapped images compared with reference images. The numbers on the left shows the image numbers.*

Tone Mapping Operators Rankings

For Experiment 2, the raw data were used to calculate image rankings using 6 point categorical judgment scale. Figure 3 compares various tone mapped images with reference images. Figure 4, presents the rankings of TMOs included in studies, for which Figures 4(a), 4(b) and 4(c) compares the TMOs for each image and Figures 4(d), 4(e) and 4(f) represent overall rankings of the TMOs. Figures 4(a) and 4(b) represents the rankings based on the visual data. In Figures 4(a) and 4(d), the lower value corresponds to the better rankings and higher values corresponds to the lower rankings. Figures 4(b), 4(c), 4(e) and 4(f) show image differences in terms of CIELAB ΔE^*_{ab} lower values corresponds to the higher rankings.

Figure 4(a) shows that when the reference images were used as test images**,** the mean categorical scale of all reference images are close to 2 which corresponds to "Just Noticeable Difference". According to Figure 4(d) the mean rank for the reference images is 1.88, giving the highest ranking.

Figures 4(a) and 4(b) represent the rankings calculated with the results of Experiment 2. From Figure 4(a), it is clear that for Image 2, mean image ranking of Reinhard local has minimum and Schlick has maximum. This is shown in top row of Figure 3. Mean image ranking for Drago's TMO is minimum for Image 4 and highest for Image 2. Similarly the results for other TMOs were predicted, Reinhard Devlin and Schlick had very high mean rank for Image 6. This can be seen in Figure 3, the results were very poor for this image by Reinhard and Devlin and Schlick TMOs. Mean rank for Image 4 and Image 5 were low in all cases except Reinhard Devlin. Reinhard Devlin performs worse in most cases. The mean rank is almost same for all TMOs for Image 2.

Figures 4(a) and 4(d) showed that Reinhard local operator gave best performance. There are two reasons: firstly, in Experiment 1, the images used for rendering purpose were produced using this TMO. Secondly, the TMO gave more texture details therefore its quality is closest to that of the reference image. Drago's operator produces higher contrast images as compared to the other TMOs. Also, it is a global operator therefore it gave less details as compared to the Reinhard's local operator. Therefore Drago's operator ranked second. Reinhard's'

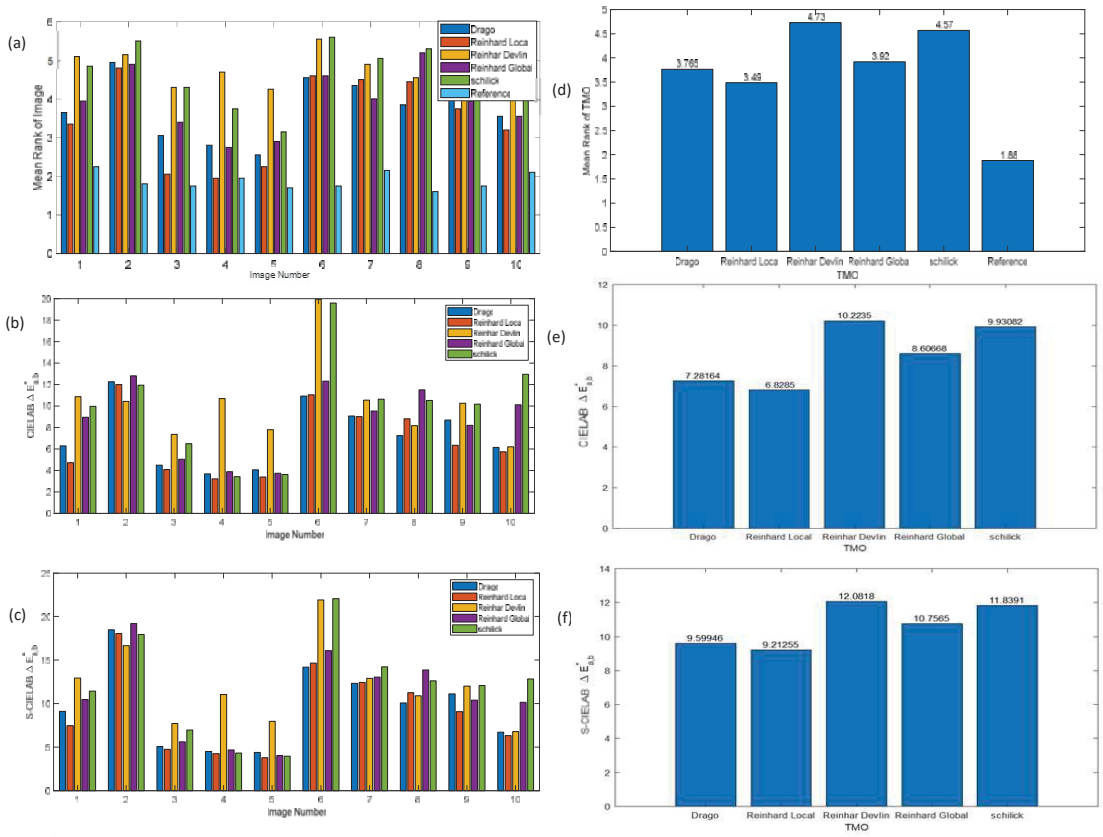

Figure 4: *TMOs ranking a) Mean image ranking based on subjective assessment. b) CIELAB(2:1) ΔE^*_{ab} for each image. c) S-CIELAB ΔE^*_{ab} for each image. D) Mean rank of each TMO based on subjective assessment. E) Average CIELAB (2:1) ΔE^*_{ab} for each TMO. f) Average S-CIELAB ΔE^*_{ab} for each TMO.*

global operates produces less details and its tone mapped images are not high contrast therefore it ranked third. The Schilik's operator ranked the 4th and Reinhard and Devlin's operator ranked the 5th, the worst.

Finally, for objective evaluation of the TMOs, full reference image quality metrics CIELAB(2:1) was used as recommended by CIE. The input for this metric is the reference and the test images in CIELAB L^*, a^* and b^* attributes. The Euclidean distance between the corresponding pixel values were calculated. In CIELAB (2:1), the ΔL^* is divided by two. Because lightness difference is less weighted, CIELAB ΔE^*_{ab} is not well correlated with perceived image difference, the spatial extension of CIELAB called S-CIELAB is used as well. In the opponent colour space the red-green and yellow-blue planes were strongly blurred with Gaussian 2-D filter but the luminance was slightly blurred. Then CIELAB ΔE^*_{ab} was applied.

Figures 4(b) and 4(c) show each image CIELAB ΔE^*_{ab} produced by calculating the image difference using CIELAB (2:1) and S-CIELAB respectively. CIELAB(2:1) and S-CIELAB both gave similar ranking as the subjective assessment. In both cases few images do not agree with subjective assessment e.g. for Image 10, CIELAB(2:1) and S-CIELAB both showed that Reinhard and Devlin's operator had less difference as compared to the Reinhard global operator. The Image 10 is illustrated in Figure 3 in comparison to the reference image. It appears to the authors that the colorfulness of Reinhard and Devlin is close to the colorfulness of the reference image which gives lower image difference consequently higher rank as compared to Reinhard's

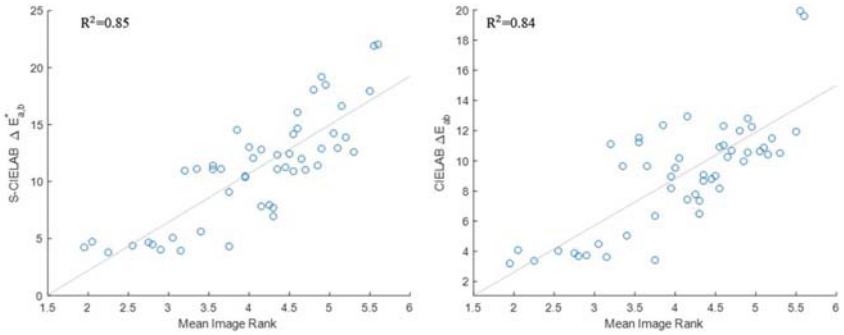

Figure 5: *a) Correlation between S-CIELABΔE^*_{ab} and mean image ranking. b) Correlation between CIELAB(2:1) ΔE^*_{ab} and mean image ranking.*

global operator. However, the subjective assessment for image 10 describe that Reinhard and Devlin's operator has lower ranking as compared to the other TMOs. The same is the case for image 8, CIELAB (2:1) and S-CIELAB both describe that Reinhard and Devlin's operator is better than Reinhard's local operator while its opposite in the subjective assessments.

Figures 4(e) and 4(f) show the average ΔE^*_{ab} calculated by CIELAB(2:1) and S-CIELAB respectively. It can be seen that both image quality metrics agreed with the subjective assessments. However, the image difference in terms of S-CIELAB had a greater magnitude than that of CIELAB (2:1).

To find out which metric has better correlation with subjective assessment, the S- CIELAB and CIELAB (2:1) were plotted vs mean image ranking respectively. The scatter plots are shown in Figure 5(a) and 5(b). These plots showed that both metrics strongly agree with the subjective assessment with R^2=0.85 and R^2=0.84 respectively.

Conclusion

In this research work, two psychophysical experiments were conducted. In the first experiment, high quality reference images were obtained. They were further used for the evaluation of five tone mapping operators in the second experiment. The results showed that Reinhard's local operator ranked first, followed by Drago's adaptive algorithmic operator, Reinhard's global operator, Schilik's quantization method and the Reinhard Devlin operator ranked the worst. Two full reference image quality metrics CIELAB(2:1) and S-CIELAB were also evaluated for evaluation of these TMOs. It was found that both metrics strongly agreed with the subjective assessments. The overall goal has been achieved to obtain reference images and to evaluate TMOs.

Future Work

A new tone mapping model is under development. The reference image produced by these experiment are used to test the quality of the tone mapped results by calculating image difference using CIELAB(2:1) and S-CIELAB ΔE^*_{ab} as it is already demonstrated that CIELAB image difference agrees with the visual assessments. Further we plan to test these TMOs with other reference images based quality metrics to study how close there results agree with the visual results.

References

[1] Mantiuk, R., Daly, S., & Kerofsky, L. "Display adaptive tone mapping" ACM Transactions on Graphics (TOG). 27(3), 68 (2008).

[2] Drago, F., Myszkowski, K., Annen, T., & Chiba, N., "Adaptive logarithmic mapping for displaying high contrast scenes" Computer graphics forum. 22(3), 419 (2003).

[3] Reinhard, E., Stark, M., Shirley, P., & Ferwerda, J., "Photographic tone reproduction for digital images" ACM transactions on graphics (TOG), 21(3), 267 (2002).

[4] Larson, G. W., Rushmeier, H., & Piatko, C., "A visibility matching tone reproduction operator for high dynamic range scenes" IEEE Transactions on Visualization & Computer Graphics, 4, 291 (1997).

[5] Schlick, C., "Quantization Techniques for the Visualization of High Dynamic Range Pictures" P. Shirley, G. Sakas, and S. Müller (eds.), Photorealistic Rendering Techniques, New York: Springer-Verlag, 7 (1994).

[6] Debevec, P. E., & Malik, J., Recovering high dynamic range radiance maps from photographs, Proceedings of the 24th annual conference on Computer graphics and interactive techniques, pg. 369. (2008).

[7] Duan, J., & Qiu, G., Fast tone mapping for high dynamic range images. ICPR'04 Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04), vol. 2, pg. 847. (2004).

[8] Reinhard, E., & Devlin, K., "Dynamic range reduction inspired by photoreceptor physiology" IEEE Transactions on Visualization & Computer Graphics, 13. (2005).

[9] Kuang, J., Yamaguchi, H., Johnson, G. M., & Fairchild, M. D., Testing HDR image rendering algorithms, Colour and Imaging Conference, Society for Imaging Science and Technology, vol. 2004, no. 1, pg. 315. (2004).

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

[11] Čadík, M., Wimmer, M., Neumann, L., & Artusi, A., Image attributes and quality for evaluation of tone mapping operators, National Taiwan University, (2006).

[12] Barkowsky, M., & Le Callet, P., On the perceptual similarity of realistic looking tone mapped high dynamic range images, IEEE International Conference on Image Processing, pg. 3245 (2010).

[13] Weber, A. G., The USC-SIPI image database version 5, USC-SIPI Report, 315, 1. (1997).

[14] Franzen, R., Kodak lossless true color image suite. source: http://r0k. us/graphics/kodak, 4. (1999).

[15] Luo, M. R., CIE Division 8: a servant for the imaging industry, Colour Science and Imaging Technologies, International Society for Optics and Photonics, vol. 4922, pg. 51. (2002).

[16] Zhang, X., & Wandell, B. A., "A spatial extension of CIELAB for digital color image reproduction" SID international symposium digest of technical papers, Society for Information Display, 27, 731 (1996).

[17] Olson, J. A., & Krinsky, N. I., "Introduction: the colourful, fascinating world of the carotenoids: important physiologic modulators" The FASEB Journal, 9(15) 1547 (1995).

[18] Kaur, M., Kaur, J., & Kaur, J., "Survey of contrast enhancement techniques based on histogram equalization" International Journal of Advanced Computer Science and Applications, 2(7) (2011).

[19] Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T. & Zuiderveld, K., "Adaptive histogram equalization and its variations", Computer vision, graphics, and image processing, 39(3), 355 (1987).

[20] Reza, A. M., "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement" Journal of VLSI signal processing systems for signal, image and video technology, 38(1), 35 (2004).

[21] Fairchild, M. D., The HDR photographic survey, Colour and imaging conference, Society for Imaging Science and Technology, vol. 2007, no. 1, pg. 233. (2007).

[22] Berns, R. S., "Methods for Characterizing CRT Displays" Displays 16(4), 173 (1996).

Author Biography

Imran Mehmood received his M.Sc. and M.Phil. degrees from Department of Electronics, Quaid-i-Azam University, Islamabad, Pakistan. He is currently doing his Ph.D. in Optical Science and Engineering from Zhejiang University Hangzhou, China, under the supervision of Professor Ming Ronnier Luo. He is working on testing and developing generic tone mapping operators in HDR imaging for real time applications. His research interest includes HDR imaging, tone mapping, image quality and colour science etc.