IEEE International Conference on Image Processing, Beijing, China, Sep. 2017.

PERCEPTUAL QUALITY ASSESSMENT OF HDR DEGHOSTING ALGORITHMS

Yuming Fang[1], *Hanwei Zhu*[1], *Kede Ma*[2], *and Zhou Wang*[2]

[1]School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China
[2]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

ABSTRACT

High dynamic range (HDR) imaging techniques aim to extend the dynamic range of images that cannot be well captured using conventional camera sensors. A common practice is to take a stack of pictures with different exposure levels and fuse them to produce a final image with more details. However, a small displacement between images caused by either camera or scene motion would void the benefits and cause the so-called ghosting artifacts. Over the past decade, many HDR deghosting algorithms have been proposed, but little work has been dedicated to evaluate HDR deghosting results either subjectively or objectively. In this work, we present a comprehensive subjective study for HDR deghosting. Specifically, we create a database that contains 20 dynamic image sequences and their corresponding deghosting results by 9 deghosting algorithms. A subjective user study is then carried out to evaluate the perceptual quality of deghosted images. The experimental results demonstrate the performance and limitations of existing HDR deghosting algorithm as well as no-reference image quality assessment models. We will make the database available to the public.

Index Terms— High dynamic range imaging, HDR deghosting, subjective quality assessment.

1. INTRODUCTION

Natural scenes often span a greater dynamic range of luminance values than those captured by current imaging sensors. In many practical scenarios, it is desirable to obtain high dynamic range (HDR) illumination in the real-world [1]. During the past decade, various HDR imaging techniques have been developed. A common approach shared by them is to capture multiple pictures with different exposure levels of the same scene and then reconstruct an HDR image by inverting the camera response function. Through tone mapping operators [2], the dynamic range of HDR images is reduced to facilitate display on devices with low dynamic range (LDR). On the other hand, multi-exposure image fusion (MEF) is considered as an effective alternative for HDR imaging. Taking the same sequence as input, MEF algorithms directly synthesize an LDR image that is more informative and perceptually appealing than any of the input images [3].

(a) Source image sequence of Fabrizio Pece [9]

(b) Pece10 [9]

(c) SPD-MEF [16]

Fig. 1. Ghosting artifacts due to camera or object motion.

A major problem of most computational HDR imaging and MEF algorithms is that a small displacement between images by either camera or scene motion would void the benefits from fusion and cause the so-called ghosting artifacts, as shown in Fig. 1. In recent years, much effort has been put to HDR deghosting for dynamic scenes. With many HDR deghosting algorithms proposed, it becomes increasingly important to evaluate the visual quality of deghosting results both qualitatively and quantitatively. Since the human visual system (HVS) is the ultimate receiver of visual information in most applications, subjective evaluation is an effective approach to understand the human behaviors when viewing deghosted images. Although it is expensive and time consuming [4, 5], subjective quality assessment has several benefits. First, it provides useful data to study human behaviors in evaluating the perceived quality of deghosted images. Second, it supplies a benckmark to compare the performance of state-of-the-art HDR deghosting algorithms. Third, it is useful to validate and compare the performance of existing and future objective image quality assessment (IQA) models in predicting the perceptual quality of deghosted images. This will in turn provide insights on developing effective IQA methods for HDR deghosting.

There had been some investigations on the performance of deghosting methods [6]-[8]. Unfortunately, only a small set of images and limited HDR deghosting algorithms are involved. These results become less relevant with many new al-

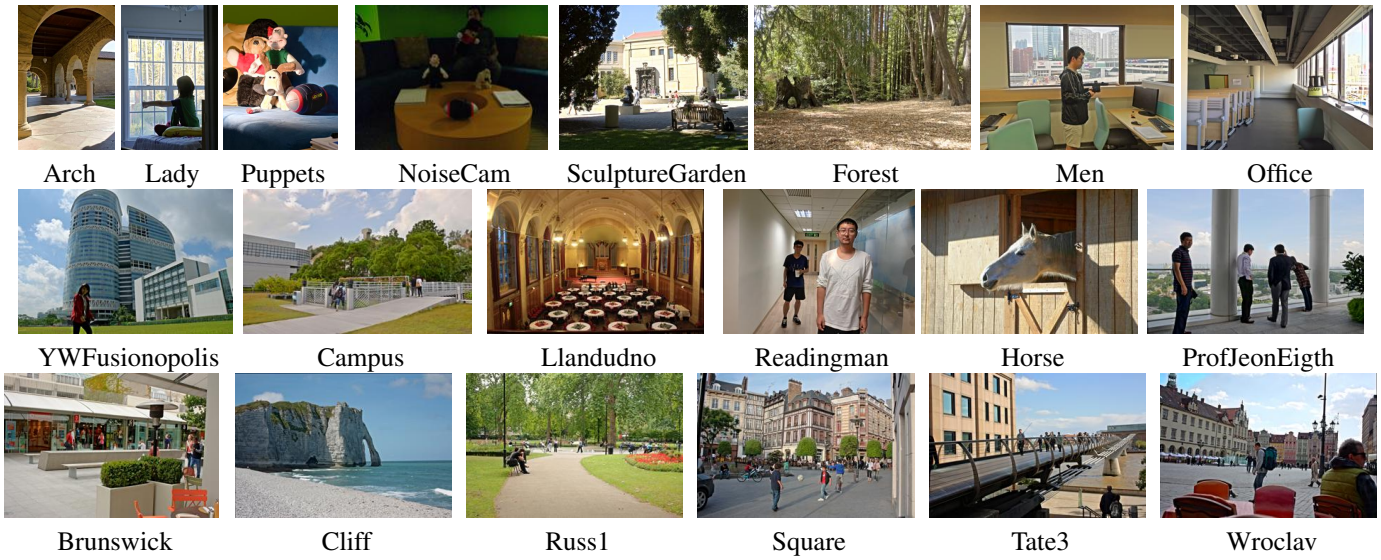

Fig. 2. Source image sequences. Each sequence is presented by one deghosted image, which has the best quality in the subjective test.

gorithms being proposed recently. This motivates us to carry out a large-scale subjective test in order to analyze classic and up-to-date HDR deghosting algorithms in depth. A database containing 20 source sequences with multiple exposure levels (≥ 3) is constructed in this study. Nine HDR deghosting algorithms are adopted to generate deghosted images, with the most recent one developed in 2016. A large-scale subjective test is carried out to compare the visual quality of deghosted images using a paired comparison methodology. We observe a considerable consensus among observers regarding the quality evaluation of deghosting results. The nine HDR deghosting algorithms are analyzed in depth based on their design philosophies. In addition, we show that state-of-the-art image quality assessment (IQA) methods are limited in predicting the perceptual quality of deghosted images. We believe all these findings are meaningful for the future development of HDR deghosting.

2. THE DATABASE

We collect a set of image sequences covering a variety of image content and motion. Since camera motion is usually small and relatively uniform in practice, we only consider various objection motion and counteract the camera motion by either setting a tripod or some image registration algorithms. In other words, all source sequences are aligned. Moreover, the database constitutes sequences with different environments, including indoor and outdoor scenes, deformable and non-deformable patterns, noisy and non-noisy scenarios, and small and large motions. Keeping these considerations in mind, we collect 20 source sequences, as shown in Fig. 2.

All of them contain at least 3 input images which represent under-exposed, over-exposed, and in-between cases.

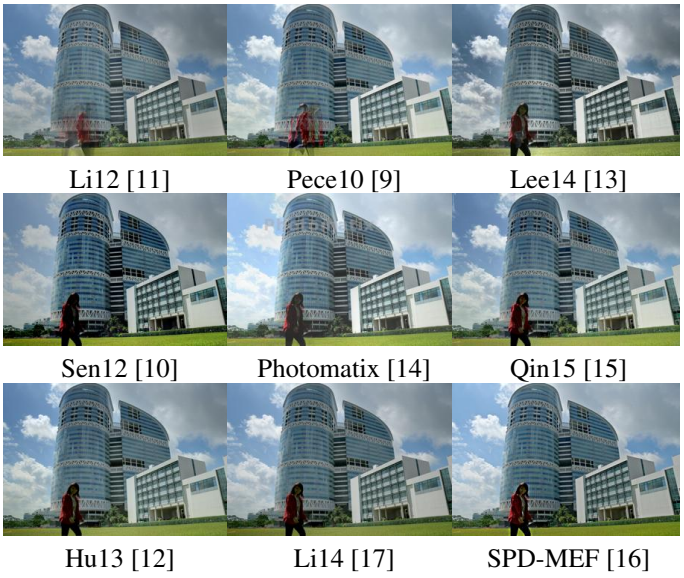

Fig. 3. Deghosting results of nine algorithms adopted in the subjective user study.

Next, we choose nine state-of-the-art ghost removal algorithms to generate deghosted images, including: Pece10 [9], Sen12 [10], Li12 [11], Hu13 [12], Lee14 [13], Li14 [17], Photomatix [14], Qin15 [15], and SPD-MEF[16] (Fig. 3). These methods are chosen to represent a variety of design philosophies, including pixel-based and path-based, HDR reconstruction followed by tone mapping and MEF, structural similarity based and low rank based algorithms. Note that Photomatix [14] is a commercial HDR software. The

deghosting results are either generated by original authors or by the code available to the public with the default settings. As a result, a total of 180 deghosted images are produced.

In the subjective test, the paired comparison methodology is employed, where participants are shown with two deghosted images at the same time, and are asked to choose the one they prefer. Since the deghosting algorithms operate on an exposure stack rather than a single image like traditional image processing algorithms, the graphical user interface (shown in Fig. 4) provides 3 representative LDR images on the top of the screen as the reference information. A pair of deghosted images are presented below the source sequence and the corresponding radio indicates their preference. Two identical 24-inch LED monitors are used in the test, calibrated according to the ITU Recommendation [18].

For each participant, the experiment starts with a training session, which includes 10 paired comparisons, using image sequences independent of the testing session. When evaluating deghosted images, the participants are instructed to mainly consider the artifacts and detailed information preservation in deghosted images.

In the testing session, we perform a complete experiment within each image sequence, which results in $20\times\binom{9}{2}$=720 pairs in total. We divide the complete experiment into 3 small sessions, within which each participant compares 240 pairs of images in 30 minutes in order to mitigate the fatigue effect. We invite 60 participants with normal or corrected visual acuity, aged from 18 to 40 to the subjective test, each of whom only conducts one session experiment. As a result, each image pair is compared exactly 20 times. The participants have no experience in image processing and quality assessment, among which 20 are female and the rest are male.

3. DATA ANALYSIS

The result for each scene is a preference matrix C, where each entry is given by

$$C_{i,j} = \begin{cases} \#\text{ times option } i \text{ preferred over option } j & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}, \quad (1)$$

which represents the number of times that each option is preferred over the other option. The final results obtained by accumulating the preference matrices for all scenes is shown in Table 1.

We adopt the statistical data analysis method by Kristi and Maya [19], which explains the experimental results under the Thurstones Case V Model [20]. In brief, it assumes that all options have equal variance and zero correlations (or less restrictively, equal correlations [21]). Furthermore, the Thurstones Case V Model sets the variances of the binary options, A and B, to one half $\sigma_A^2 = \sigma_B^2 = \frac{1}{2}$ and the covariance of A and B is 1, which simplifies the Thurstone Law to

$$\overline{\mu}_{AB} = \phi^{-1}\left(\frac{C_{A,B}}{C_{A,B} + C_{B,A}}\right), \quad (2)$$

where $\overline{\mu}_{AB}$ is the quality different from the binary choice and $\phi^{-1}(x)$ is the inverse cumulative distribution function (CDF) of the standard normal function.

We then apply four different approaches including least squares method, maximum likelihood method, maximum a posteriori estimation method, and morrissey and Gulliksen's incomplete matrix method [22] to aggregate the pairwise ranking information into a global ranking. The experimental results are shown in Fig. 5 and Fig. 6, from which we have several useful observations. First, not a single HDR deghosting algorithm performs the best for all sequences, which indicates room for further performance improvement. Second, SPD-MEF [16] performs the best in general. The success may be that SPD-MEF [16] takes advantage of both exposure invariant feature and intensity mapping function for robust inconsistent motion rejection. Third, patch-based algorithms such as SPD-MEF [16], Hu13 [12] and Sen12

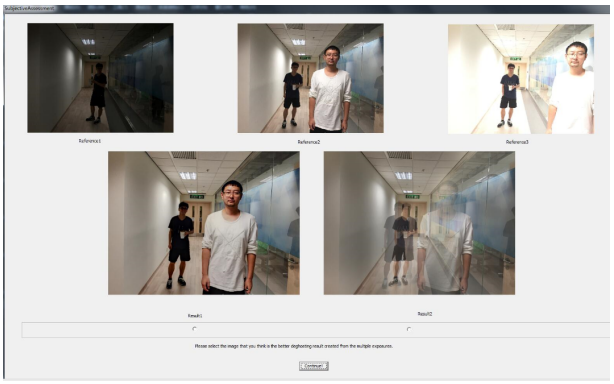

Fig. 4. Graphical user interface in the subjective test.

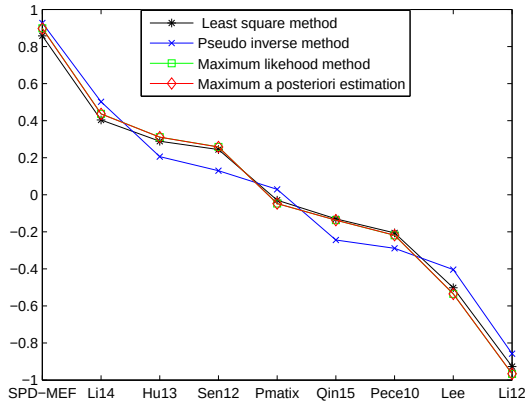

Fig. 5. The aggregation results obtained by the four methods.

Algorithm	Li12	Lee14	Photomatix	Qin15	Pece10	Sen12	Hu13	Li14	SPD-MEF	SUM
Li12	0	134	86	68	52	60	52	54	15	521
Lee14	266	0	137	119	141	71	99	73	53	959
Photomatix	314	263	0	189	184	115	108	102	59	1334
Qin15	332	281	211	0	211	143	123	109	52	1431
Pece10	348	259	216	220	0	127	160	117	94	1541
Sen12	340	329	285	257	273	0	157	177	91	1909
Hu13	348	301	292	277	240	243	0	159	111	1971
Li14	346	327	298	291	283	223	241	0	111	2120
SPD-MEF	385	347	341	348	306	309	289	289	0	2614

Table 1. Total aggregate preference matrix in the subjective user study and the total number of preferences of each method.

[10] generally perform better than pixel-based algorithms such as Pece10 [9], Li12 [11] and Lee14 [13], except for Li14 [17] that ranks the second. There is no surprise because patch-based algorithms take neighbouring information into consideration and result in more robust motion alignment. Note that Li14 [17], a pixel-based method, also considers neighbouring pixels in extreme cases to refine the motion rejection process. Fourth, the low-rank based method, Lee14 [13] is subpar in the test, resulting from the failure of preventing ghosting artifacts, especially on sequences with small motions. This may be because low rank schemes implicitly assume the static background dominates the scene, but small motions often do not follow the sparsity assumption, resulting in artifacts. Other methods that explicitly hold the assumption such as Pece10 [9] and Li12 [11] do not perform well either. Fifth, besides ghosting artifacts, we find some other types of distortions that may clearly affect the human judgements of perceptual quality. For example, halo artifacts at the boundary of the main subject with a bright background is the main problem for SPD-MEF [16]. Color speckle noise due to the inaccurate camera response function estimation would frequently appear for images generated by Lee14 [13]. In addition, blurring artifacts may be generated if there are small errors in motion estimation. Finally, in certain extreme cases, if the moving objects that appear in the final image are under-/overexposed, and their structures cannot be properly retrieved from other exposures, the ghosting artifacts would likely appear, which is a common problem to most existing deghosting schemes.

Fig. 6. Grouping and ranking for each algorithm according to the quality scores. The left, the better.

As indicated previously, the subjective quality assessment method is time consuming, laborious, and expensive. With new HDR deghosting algorithms being proposed, it is desirable to adopt objective quality models to measure how well different deghosting algorithms perform, without the need of new subjective experiments. It is difficult to develop full-reference image quality models for HDR deghosting for the lack of a well defined reference image. On the other hand, there exists no no-reference objective quality models specifically designed for HDR deghosting. Here, we use several general purpose no-reference objective quality models to conduct the experiments based on the subjective test: NIQE [23], IL-NIQE [24], and QAC [25]. In order to measure how well these models, we adopt three commonly used metrics: Pearson linear correlation coefficient (PLCC), Spearman rank-order correlation coefficient (SRCC), and Root mean squared error (RMSE). The experimental results are shown in Table. 2, where it can been seen that PLCC and SRCC values from these three metrics are relative low, and RMSE values are high, which demonstrates that these existing models cannot obtain good performance in performance evaluation of deghosted images.

	NIQE	IL-NIQE	QAC
PLCC	0.5985	0.5504	0.5161
SRCC	-0.0279	0.2551	-0.0234
RMSE	27.8045	27.8155	29.8785

Table 2. Experimental results by three no-reference image quality metrics.

4. CONCLUSION AND FUTURE WORK

We create a large-scale database for quality evaluation of deghosted images. It contains 20 dynamic image sequences, together with 720 deghosting results by nine state-of-the-art deghosting algorithms, based on which we conduct a subjective user study. The constructed database with subjective scores can be used for the benckmark of objective quality assessment for HDR deghosting in the research community. We also provide in-depth analysis for the performance of different deghosting algorithms. In the future, we will develop effective IQA models for quality assessment of deghosted images.

5. REFERENCES

[1] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski, *High Dynamic Range Imaging: Acquisition, Display, and Image-based Lighting.* Morgan Kaufmann, 2010.

[2] K. Ma, H. Yeganeh, K. Zeng, and Z. Wang, "High dynamic range image compression by optimizing tone mapped image quality index," *IEEE Transactions on Image Processing*, Vol. 24, No. 10, pp. 3086-3097, 2015.

[3] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Transactions on Image Processing,* vol. 24, no. 11, pp. 3345-3356, 2015.

[4] Y. Fang, and W. Lin, *Methods for image quality assessment*, Wiley Encyclopedia of Electrical and Electronics Engineering, 1-11, John Wiley and Sons, Inc, 2015.

[5] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo Exploration Database: New Challenges for Image Quality Assessment Models," *IEEE Transactions on Image Processing*, In Press, Nov. 2016.

[6] A. Srikantha, "Ghost detection and removal for high dynamic range images: Recent advances," *Signal Processing Image Communication,* vol. 27(6), pp. 650-662, 2012.

[7] K. K. Hadziabdic, J. H. Telalovic and R. Mantiuk, "Comparison of Deghosting Algorithms for Multi-exposure High Dynamic Range Imaging," *in Spring Conference on Computer Graphics, ACM* pp. 21-28, 2013.

[8] O. T. Tursun, A. O. Akyz, A. Erdem and E. Erdem "The State of the Art in HDR Deghosting: A Survey and Evaluation" *in Computer Graphics Forum,* pp. 683-707, 2015.

[9] F. Pece and J. Kautz, "Bitmap movement detection: HDR for dynamic scenes," *in IEEE Conference on Visual Media Production*, pp. 1-8, 2010.

[10] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust patch-based HDR reconstruction of dynamic scenes," *ACM Transactions on Graphics*, vol. 31(6), pp. 439-445, 2012.

[11] S. Li and X. Kang, "Fast multi-exposure image fusion with median filter and recursive filter," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 2, pp. 626-632, 2012.

[12] J. Hu, O. Gallo, K. Pulli, and X. Sun, "HDR deghosting: How to deal with saturation?" *in IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1163-1170, 2013.

[13] C. Lee, Y. Li, and V. Monga, "Ghost-free high dynamic range imaging via rank minimization," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1045-1049, 2014.

[14] Photomatix, Commercially-Available HDR Processing Software, Available: http://www.hdrsoft.com/, 2015.

[15] X. Qin, J. Shen, X. Mao, X. Li, and Y. Jia, "Robust match fusion using optimization," *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1549-1560, 2015.

[16] K. Ma and Z. Wang, "Multi-exposure image fusion:A patch-wise approach," *IEEE International Conference on Image Processing*, pp. 1717-1721, 2015.

[17] Z. Li, J. Zheng, Z. Zhu, and S. Wu, "Selectively detail-enhanced fusion of differently exposed images with moving objects," *IEEE Transactions on Image Processing*, vol. 23, pp. 4372-4382, 2014.

[18] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment, Syntheses Lectures on Image, Video and Multimedia Processing,* Morgan & Claypool Publishers, Mar, 2006.

[19] K. Tsukida, M. R. Gupta, "How to Analyze Paired Comparison Data," 2011.

[20] L. L. Thurstone, *A law of comparative judgment. Psychological Review,* Psychological Review, vol. 34 pp. 273-286, 1927.

[21] F. Mosteller, "Remarks on the method of paired comparisons: I.The least squares solution assuming equal standard deviations and equal correlations," *Psychometrika* , vol. 16(1), pp. 3-9, 1951.

[22] J. H. Morrissey, "New method for the assignment of psychometric scale values from incomplete paired comparisons" *Journal of the Optical Society of America*, vol. 45(5), pp. 373-378, 1955.

[23] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Process Letters,* vol. 20, no. 3, pp. 209-212, 2013.

[24] L. Zhang, L. Zhang, A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, pp 2579-2591, 2015.

[25] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," *in IEEE Conference on Computer Vision and Pattern Recognition,* pp. 995-1002, 2013.