Subjective and Objective Evaluation of Multi-exposure High Dynamic Range Image Deghosting Methods

Kanita Karaduzovic-Hadziabdic, Jasminka Hasic Telalovic[1] and Rafal Mantiuk[2]

[1]International University of Sarajevo, Bosnia and Herzegovina
[2]Cambridge University, United Kingdom

Abstract

To avoid motion artefacts when merging multiple exposures into an HDR image, a number of deghosting algorithms have been proposed. These algorithms, however, do not work equally well on all types of scenes, and some may even introduce additional artefacts. Even though subjective methods of evaluation provide reliable means of testing, they need to be repeated for each new proposed method or even its slight modification and are cumbersome to perform. In this work, we evaluate several computational approaches of quantitative evaluation of multi-exposure HDR deghosting algorithms and demonstrate their results on five state-of-the-art algorithms. The quality of HDR images produced by deghosting methods is measured in a subjective experiment, and then evaluated using five objective metrics. The most reliable metric is then selected by testing correlation between subjective and objective metric scores.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: HDR image deghosting methods—subjective and objective evaluation

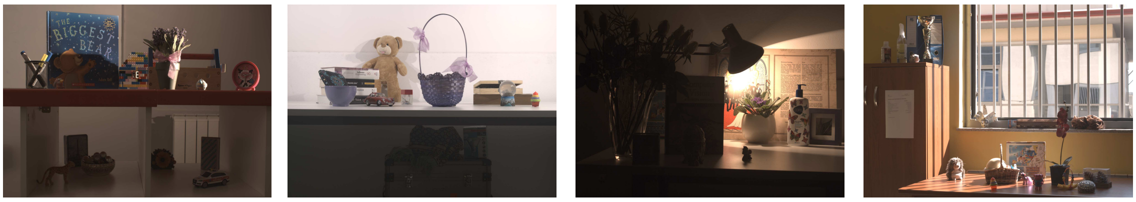
1. Introduction

There are various methods for capturing high dynamic range illumination present in most real world scenes. However, the most popular method of generating HDR images is multi-exposure technique [DM08, MN99, RBS99], where a sequence of differently exposed low dynamic range (LDR) images are merged to produce an HDR image. Multi-exposure technique works well for static scenes taken on a tripod. However, most everyday images contain moving objects and are captured hand-held. To merge such photographs into HDR images, a number of multi-exposure HDR deghosting algorithms have been proposed [SKY*12, HGPS13, GKTT13]. The main goal of these algorithms is to produce a good quality HDR image without motion artefacts. This brings the need to evaluate and compare their results. Subjective quality assessments provide a reliable means of image quality evaluation. However, they are often costly and demanding to perform. Objective quality assessment methods provide computational and automated means of measuring performance of different algorithms.

The main contribution of this paper is to compare subjective and objective evaluation of multi-exposure HDR image deghosting methods with the focus on selecting the most suitable objective metric for evaluating HDR deghosting algorithms. Thus, an in-depth evaluation of 4 full-reference objective quality metrics suitable for evaluation of HDR images has been performed: (PU2SSIM) [AMS08], HDR-VDP-2 [MKRH11] and Weber root mean square error (Weber RMSE); and Liu et al.'s [LWC*13] no-reference metric designed for evaluating motion deblurring. We also tested the performance of Liu et. al's metric as a full-reference metric (using a deghosted and a reference images as inputs to the metric). The success of objective metrics was measured by a subjective evaluation of five state-of-the-art algorithms (Sen et al. [SKY*12], Silk and Lang [SL12], Hu et al. [HGPS13], Photomatix Pro (version 4.2.6) and Photoshop CS5 Extended (version 12.0), and then by computing Spearman and Pearson correlation between the two scores to select the most reliable metric for objective evaluation. An HDR image generated by merging a sequence of RAW images using Robertson et al. method [RBS99] without deghosting is also included in the evaluation as a control condition. Please refer to the state-of-the-art report [TAEE15] for a comprehensive survey of approximately 50 HDR deghosting algorithms. A dataset of carefully selected test and reference images was also created in order to be used in the evaluation.

2. Dataset

Since algorithm performance may be scene dependent, we created a dataset particularly designed to provide a comprehensive set of challenging real life scenes for evaluating deghosting algorithms. The dataset contains 36 scenes organized into 4 different *image sets*. Each image set refers to a specific lighting condition under which 9 different scenes, each scene with different type of motion, have been captured in a controlled environment. The first three sets were captured in a dark room where for *set 1*, the only source



Example scene, image set 1 Example scene, image set 2 Example scene, image set 3 Example scene, image set 4

Figure 1: *Four example scenes used in the experiments, one scene from each image set.*

of illumination was coming from a Halogen 300 Watt spot light, positioned at 45 degrees to the table containing objects in motion and two 60 Watt light bulbs-white positioned at 45 degrees on the other side of the table; for *set 2*, the light source was coming from a 2×300 Watt Halogen spot photographic light positioned at 45 degrees to the table containing objects in motion on both sides; for *set 3*, the light source was coming from a table lamp with 60 Watt light bulb; *set 4* was captured in a room where the camera was pointing towards a large window. Figure 1 shows 4 representative scenes used in the experiments. Complete dataset is available for the research community [deg]. For each scene, both *test* and *reference* multi-exposure sequence was captured. Test exposure sequence refers to the sequence of multi-exposures that contain either objects or camera motion. Reference exposure sequence refers to the sequence of multi-exposures where all pixels are perfectly aligned (i.e. ground truth sequence). For exposure sequences with objects in motion, the position in the middle of the motion was selected for the reference exposure sequence. Each image set contained exactly 9 test and reference sequences, each with different type of motion [KHM14]: 1) complex motion*, 2) handheld, 3) large object displacement with large motion (lolm*), 4) large object displacement with small motion (losm*), 5) multiview, 6) non-rigid motion (nrm*), 7) occlusion*, 8) small object displacement with large motion (solm*), and 9) small object displacement with small motion (sosm*). Scenes marked with * indicate dynamic scenes captured on a tripod were objects were moved between LDR image capture to simulate motion. For each scene, five exposures of RAW and JPG images with one f-stop exposure time difference were captured. In order to avoid any camera motion, camera was remotely controlled by gPhoto2 (version 2.5.5. http://gphoto.sourceforge.net) and mounted on a tripod. To get the best algorithm performance, we used linear 16-bit images as inputs to the algorithms. For subjective experiments, generated HDR images were tonemapped by applying a customized tone mapping operator (TMO) [KHM13] based on the fast bilateral filter [DD02]. The main goal of this TMO is to reproduce details exactly as they were captured in HDR images.

3. Subjective Experimental Setup

20 participants with computer graphics background, aged between 22 and 41, were asked to complete a pairwise comparison experiment. Each participant was presented with randomized all possible comparison pairs of the same scene processed with a different deghosting algorithm. For 6 evaluated algorithms (5 deghosting and 1 without deghosting) and 36 scenes, a total number of comparison pairs was $36 \times \binom{6}{2} = 540$. The experiment was divided into 4 sessions, where each session contained 9 scenes from each image set (i.e. 135 comparison pairs) that were presented randomly for each participant. Each observer participated in all 4 sessions, where each session lasted maximum 30 minutes. Each image pair was displayed side by side on two $21''$ 1600×900 HP 2011x LCD monitors. Monitors were rotated 20° around the vertical axis (to be perpendicular to the viewing direction) and at an eye level of the participants, with a viewing distance of 70 cm. All experiments were performed in a dark room where the only light source was coming from a corridor light, which was constant throughout the experiments. The most common artefacts that could be introduced by a deghosting process as identified in [KHM14] are: motion artefacts, loss of dynamic range (i.e. amount of details visible), noise and color artefacts. The participants were asked to choose the preferred image based on the following criteria: firstly, select an image that has the least amount of motion artefacts. If it is not possible to make a difference between an image pair based on motion artefacts, select the image that has lower amount of any of three artefacts: loss of details in under-/over-exposed regions, color artefacts, and noise. No time limit was imposed during a selection of the preferred image. Before the start of the experiment, a short briefing on possible multi-exposure HDR deghosting artefacts was presented to the participants. A pilot study was performed to evaluate the time required for participants to perform an experiment session, and the overall clarity of the experiment.

4. Results and Analysis

The results of the subjective experiments were analyzed by estimating which portion of the population will select one algorithm over another. To do this, pairwise comparison data was scaled in Just-Noticable-Difference (JND) units (Figure 2) under Thurstone Case V assumptions, where the difference in 1 JND unit corresponds to 75% of observers selecting one algorithm over another. To scale the pairwise comparison data in JND units, we applied Bayesian method of Silverstein and Farrel [SF01]. The error bars in Figure 2 denote 95% confidence intervals computed by bootstrapping.

The results show that Sen et al.'s method outperforms all algorithms for most of the tested scenes in image sets 1, 2, and 3, followed by Hu et al.'s algorithm for image sets 1 and 2. However, for image set 4, these two algorithms do not perform as well. The reason is that for this image set, in most of the scenes, even the shortest exposure is over-saturated. This shows that these algorithms struggle with reconstructing the large saturated region of the scene. Furthermore, Hu et al.'s algorithm also produces color artefact in over

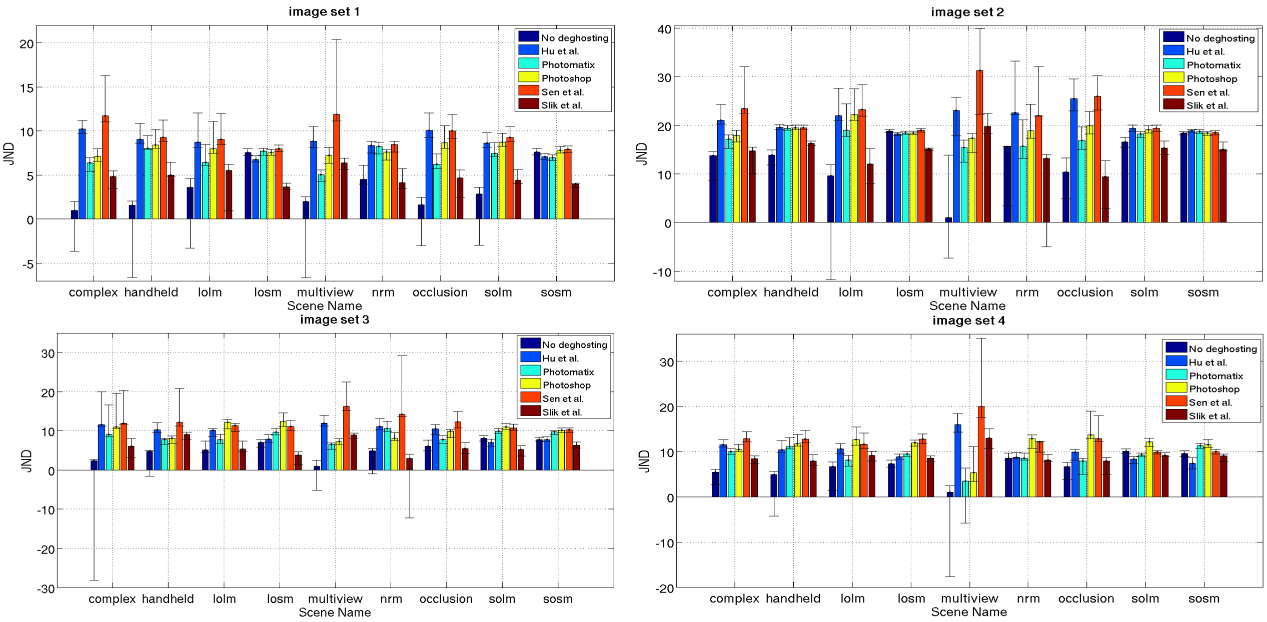


Figure 2: *The results of the subjective experiment for each image set scaled in JND units (higher the values, the better). Absolute values are arbitrary and only the relative differences are relevant. The error bars denote* 95% *confidence intervals computed by bootstrapping.*

saturated regions. These artefacts are mostly visible in images generated from image set 3, where for most scenes, this algorithm produces color artefact in the area close to the lamp's light bulb. Subjective results also show that for almost all scenes, Photoshop outperforms Photomatix. Silk et al.'s algorithm has lowest score from 5 evaluated algorithms (not considering the non-deghosted image) for most scenes. Exception is the *multiview* scene where this algorithm was found to outperform Photomatix (in all 4 image sets) and Photoshop (in 3 image sets). Even though when this algorithm performs well in region where motion is present, it usually generates a 'washed out' faded trail in the regions where there was object movement resulting in a reduced dynamic range.

Five objective metrics were tested whether they can predict deghosting artefacts. Because each evaluated method produces slightly different HDR pixel values (in terms of both contrast and absolute values), test and reference HDR images were generated individually by each method. To minimize possible small pixel misalignments, each reference image was aligned to the test image by homographic transformation found from SURF key-point matching (*pfsalign* command from *pfstools* [MKMS07]). The metric prediction error was determined by Spearman (ρ) and Pearson (*r*) correlation coefficients computed between subjective experiment results scaled in JND units and objective quality metric predictions. Before computing Pearson correlation, a non-linear regression using a logistic function was applied to the objective scores (for all metrics except Liu et. al's due to it's low correlation). An example of logistic function fitting is displayed in Figure 3. Values highlighted in bold in Table 1 represent statistically significant Spearman and Pearson correlation scores at $\alpha = 0.05$ using a t-test distributed as Student's distribution with 18 degrees of freedom.

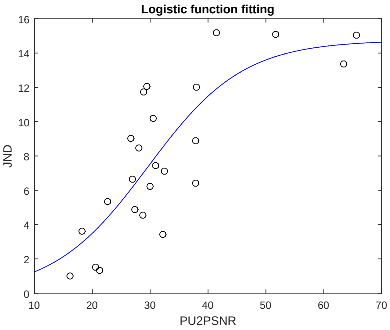


Figure 3: *Applying non-linear regression to PU2PSNR metric for multiview scene using logistic function.*

The results show that HDR-VDP-2 metric has the highest correlation scores for all scenes except for the *handheld* scene where PU2SSIM score is slightly better. In addition, correlation scores show that there is a low correlation for Liu et al.'s metric, which implies that this metric is not suitable for evaluating HDR deghosting algorithms. One of the emerging patterns is that all metrics except the HDR-VDP-2, show weak correlation for the small-object-small-motion (*sosm*). Even HDR-VDP-2 metric has the lowest correlation score for this scene, when compared to the correlation scores of other scenes. HDR-VDP-2 can be considered more suitable for evaluating deghosted HDR images with large motion displacement than for images with small motion displacement. Possible reason for this could be that for small motion displacements,

Table 1: *Spearman's (ρ) and Pearson's (r.a. and r.b) correlation coefficients for relation between objective metric predictions and subjective evaluation scores: values averaged across all image sets. Pearson correlation was computed before (r.b) and after (r.a.) fitting the logistic function. Values highlighted in bold represent statistically significant correlation scores.*

	PU2PSNR			PU2SSIM			HDR-VDP2-Q			Weber RMSE			Liu no-ref.		Liu full-ref.	
	ρ	*r.b*	*r.a*	ρ	*r.b*	*r.a*	ρ	*r.b*	*r.a*	ρ	*r.b*	*r.a*	ρ	*r*	ρ	*r*
complex	**0.89**	**0.72**	**0.85**	**0.83**	**0.69**	**0.71**	**0.97**	**0.82**	**0.86**	**0.90**	**0.80**	**0.82**	0.07	0.08	-0.01	0.21
handheld	**0.79**	**0.63**	**0.83**	**0.94**	**0.82**	**0.83**	**0.90**	**0.74**	**0.87**	**0.80**	**0.74**	**0.78**	0.30	0.15	0.03	0.15
lolm	**0.89**	**0.63**	**0.74**	**0.80**	0.39	**0.46**	**0.94**	**0.83**	**0.88**	**0.83**	**0.69**	**0.72**	-0.01	0.06	-0.10	-0.16
losm	**0.64**	0.32	0.42	**0.56**	0.10	0.10	**0.67**	0.33	0.42	**0.61**	0.39	0.39	-0.01	0.18	0.01	-0.06
multiview	**0.81**	**0.77**	**0.81**	**0.81**	**0.70**	**0.81**	**0.97**	**0.91**	**0.91**	**0.77**	**0.69**	**0.75**	-0.20	-0.13	0.39	0.44
nrm	**0.53**	**0.47**	**0.48**	**0.54**	0.22	0.23	**0.76**	**0.64**	**0.64**	**0.47**	0.33	0.33	-0.03	0.05	-0.20	-0.44
occlusion	**0.74**	**0.65**	**0.71**	**0.63**	0.35	0.43	**0.89**	**0.83**	**0.92**	**0.80**	**0.63**	**0.64**	-0.03	-0.10	-0.46	-0.34
solm	**0.49**	0.38	0.39	0.30	0.08	0.08	**0.81**	**0.63**	**0.75**	0.43	0.26	0.26	0.03	0.00	0.04	-0.07
sosm	0.24	0.16	0.40	0.24	0.00	0.00	**0.56**	0.32	0.44	0.22	0.26	0.27	0.09	0.18	0.10	0.26
Average	**0.67**	**0.53**	**0.62**	**0.63**	0.37	0.41	**0.83**	**0.67**	**0.74**	**0.65**	**0.53**	**0.55**	0.02	0.05	-0.02	0.00

the human eye may not be as sensitive to these small pixel changes as computational metrics.

5. Conclusion

The paper presented a subjective and objective evaluation of five state-of-the-art HDR deghosting algorithms. Initially, a comprehensive set of 36 test and reference multi-exposure images were captured. Then, subjective experiments were performed based on the most common HDR deghosting algorithms' artefacts. Afterwards, a set of 5 suitable objective metrics were tested to assess whether they can be used in objective evaluation of HDR deghosting algorithms. The most reliable metric was selected by performing Spearman and Pearson correlation between the two scores. The results showed that from the tested metrics, HDR-VDP-2 is the most suitable objective metric (at least for the tested scenes) for evaluating HDR deghosting algorithms. Even though test and reference images are made publicly available [deg] for future evaluations of HDR deghosting algorithms, reference images may be hard to produce for additional scenes. Therefore, future studies may include design of a no-reference objective metric.

References

[AMS08] Aydin T. O., Mantiuk R., Seidel H.-P.: Extending quality metrics to full luminance range images. In *Proc. of SPIE* (2008). 1

[DD02] Durand F., Dorsey J.: Fast bilateral filtering for the display of high-dynamic-range images. *ACM Trans. Graph. 21*, 3 (2002). 2

[deg] http://projects.ius.edu.ba/computergraphics/hdr/. 2, 4

[DM08] Debevec P. E., Malik J.: Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes* (2008), SIGGRAPH '08, ACM. 1

[GKTT13] Granados M., Kim K. I., Tompkin J., Theobalt C.: Automatic noise modeling for ghost-free hdr reconstruction. *ACM Trans. Graph. 32*, 6 (2013). 1

[HGPS13] Hu J., Gallo O., Pulli K., Sun X.: Hdr deghosting: How to deal with saturation ? In *CVPR* (2013). 1

[KHM13] Karaduzovic K., Hasic J., Mantiuk R.: Comparison of deghosting algoirthms for multi-exposure high dynamic range imaging. In *In: Proc. of Spring Conference in Computer Graphics* (2013). 2

[KHM14] Karaduzovic K., Hasic J., Mantiuk R.: Expert evaluation of deghosting algorithms for multi-exposure high dynamic range imaging. In *Proc. of HDRi2014 - Second International Conference and SME Workshop on HDR imaging* (2014). 2

[LWC*13] Liu Y., Wang J., Cho S., Finkelstein A., Rusinkiewicz S.: A no-reference metric for evaluating the quality of motion deblurring. vol. 32. 1

[MKMS07] Mantiuk R., Krawczyk G., Mantiuk R., Seidel H.-p.: High dynamic range imaging pipeline: perception-motivated representation of visual content. In *Human Vision and Electronic Imaging* (2007). 3

[MKRH11] Mantiuk R., Kim K. J., Rempel A. G., Heidrich W.: Hdr-vdp-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph (Proc. SIGGRAPH) 30*, 4 (2011). 1

[MN99] Mitsunaga T., Nayar S.: Radiometric self calibration. In *IEEE Conference on CVPR* (1999), vol. 1. 1

[RBS99] Robertson M., Borman S., Stevenson R.: Dynamic range improvement through multiple exposures. In *Proc. of ICIP* (1999). 1

[SF01] Silverstein D., Farrell J.: Efficient method for paired comparison. *Journal of Electronic Imaging 10*, 6 (2001). 2

[SKY*12] Sen P., Kalantari N. K., Yaesoubi M., Darabi S., Goldman D. B., Shechtman E.: Robust patch-based hdr reconstruction of dynamic scenes. *ACM Transactions on Graphics (TOG) (Proceedings of SIGGRAPH Asia 2012) 31*, 6 (2012). 1

[SL12] Silk S., Lang J.: Fast high dynamic range image deghosting for arbitrary scene motion. In *Proc. of Graphics Interface* (2012). 1

[TAEE15] Tursun O. T., Akyuz A. O., Erdem A., Erdem E.: The state of the art in hdr deghosting: A survey and evaluation. *Computer Graphics Forum* (2015). 1