PERCEPTUAL EVALUATION OF MULTI-EXPOSURE IMAGE FUSION ALGORITHMS

Kai Zeng, Kede Ma, Rania Hassen and Zhou Wang

Dept. of Electrical & Computer Engineering, University of Waterloo, Waterloo, ON, Canada
Email: kzeng@uwaterloo.ca, k29ma@uwaterloo.ca, rhassen@uwaterloo.ca, zhouwang@ieee.org

ABSTRACT

Multi-exposure image fusion is considered an effective and efficient quality enhancement technique widely adopted in consumer electronics products. Nevertheless, little work has been dedicated to the quality assessment of fused images created from natural images captured at multiple exposure levels. In this work, we first build a database that contains source input images with multiple exposure levels ($\geq$ 3) together with fused images generated by both classical and state-of-the-art image fusion algorithms. We then carry out a subjective user study using a multi-stimulus scoring approach to evaluate and compare the quality of the fused images. Considerable agreement between human subjects has been observed. Our results also show that existing objective image quality models developed for image fusion applications either poorly or only moderately correlate with subjective opinions.

Index Terms— subjective image quality assessment, multi-exposure images, image fusion, objective image quality assessment

1. INTRODUCTION

An effective and efficient approach to obtain images of enhanced quality is to acquire multiple images at different exposure levels followed by multi-exposure image fusion (MEF), which fills the gap between high dynamic range (HDR) natural scenes and low dynamic range (LDR) pictures captured by normal digital cameras. MEF combines multiple input images at different exposure levels and synthesizes an output LDR image that is more informative and perceptually appealing than any of the input images [1, 2].

The problem of MEF can be generally formulated as

$$F(x,y) = \sum_{k=1}^{K} W_k(x,y) I_k(x,y) \ , \tag{1}$$

where K is the number of input images in the source sequence, $I_k(x,y)$ and $W_k(x,y)$ represent the intensity value (or coefficient amplitude in transform domain) and the weight at the pixel located at (x,y) in the k_{th} exposure image, respectively, and F denotes the fused image. The weight factor $W_k(x,y)$ is often spatially adaptive and bears information regarding the relative structural details and perceptual importance of different exposures. Depending on the specific models for structural information and perceptual importance, MEF algorithms differ in the computation of $W_k(x,y)$.

A significant number of MEF algorithms have been proposed, ranging from simple weighted average to sophisticated methods based on advanced statistical image models. Local and global energy weighting approaches are the simplest ones, which employ the local or global energy in the image to determine W_k. Dated back to 1984, Burt [1] first employed Laplacian pyramid decomposition for binocular image fusion. Later in 1994, Burt and Kolczynski applied this decomposition to MEF, where they selected the local energy of pyramid coefficients and the correlation between pyramids within the neighborhood as quality measures. Goshtasby [3] partitioned each source image into several non-overlapping blocks and selected the block with the highest entropy to construct the fused image. Mertens *et al.* [4] adopted proper contrast, high saturation and well exposure as quality measures to guide the fusion process in a multiresolution fashion. Bilateral filter is used in [5] to calculate edge information, which is subsequently employed to compute the weights. Song et al. [6] first estimated the initial image by maximizing the visual contrast and scene gradient and synthesized the fused image by suppressing reversals in image gradients. Zhang *et al.* [7] constructed visibility and consistency measures from gradient information and used them as the weighting factors. A similar gradient-based MEF method is proposed in [8]. Based on [4], Li *et al.* [9] enhanced the details of a given fused image by solving a quadratic optimization problem. A median filter and recursive filter based MEF method is developed in [10] by taking local contrast, brightness and color dissimilarity into consideration. More recently, Li *et al.* [11] proposed a guided filter to control the roles of pixel saliency and spatial consistency when constructing W_k. Shen *et al.* [12] embedded perceived local contrast and color saturation into a conditional random field and derived W_k based on maximum a posteriori (MAP) estimation.

With multiple fusion algorithms available, a natural question that follows is which one delivers the best performance. In the literature, there has been substantial effort on developing objective image quality assessment (IQA) models for image fusion applications. Qu *et al.* [13] combined the mutual information between the fused and multiple input images to evaluate image quality. Xydeas and Petrovic [14] extracted edge information using Sobel operator and employed edge strength as the main feature in assessing the quality of fused images. A similar idea was employed in [15], where Wang and Liu retrieved edge strength using a two-scale Haar wavelet. Zheng *et al.* [16] computed spatial frequency using multi-directional gradient filters and estimate the quality of fused images based on activity levels. Inspired by the structural similarity (SSIM) index [17] for general purpose IQA, Piella and Heijmans [18] developed three models to predict fused image quality based on the universal quality index (UQI) [19]. Cvejie *et al.* [20] and Yang *et al.* [21] also built their quality measures upon structural information theory. Chen and Varshney [22] estimated local saliency based on edge intensities and combined saliency with global contrast sensitive function. Chen and Blum [23] applied contrast sensitivity filter in the frequency domain and then pool local information preservation scores to produce a global quality measure.

Despite the increasing interests in developing fusion and objective IQA models for various image fusion applications, systematic and comprehensive evaluation and comparison of these models has been largely lacking. In most cases, the performance of existing

978-1-4799-6536-6/14/$31.00 ©2014 IEEE

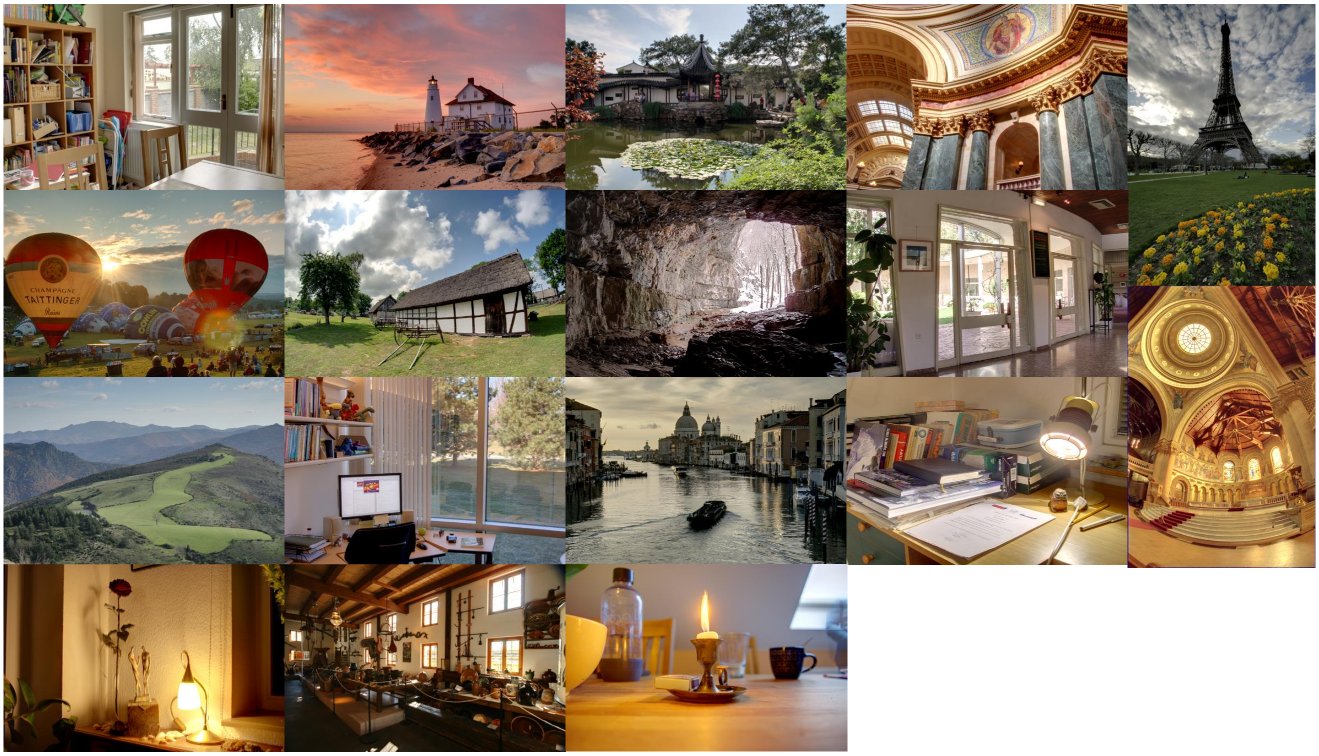

Fig. 1. Source images in the database.

methods were demonstrated using limited examples only. Because human eyes are the ultimate receivers in most applications, subjective user study is considered as the most reliable approach to evaluate the quality of fused images and the performance of objective IQA approaches. Toet and Franken [24] examined the perceptual quality of multi-scale image fusion schemes, where only night-time outdoor scenes and very simple fusion methods were included in the study. Vladimir [25] reported subjective assessment results for multi-sensor image fusion algorithms. However, the number of input images was limited to 2 and most test images were monochrome aerial pictures. Moreover, state-of-the-art image fusion algorithms are missing from the experiment. To demonstrate the effectiveness of their fusion algorithm, Song *et al.* [6] conducted two groups of paired comparison tests through both on-site and a Web platform, where the subjective experimental results only include few examples. Shen *et al.* [12] reported subjective evaluation results in terms of global contrast, details, colors, and overall appearance, which all appeared to be important contributing factors of perceptual quality. The paper also suggested that specific fusion parameters should be adapted to individual applications. Nevertheless, the main purpose of the user-study is still for demonstrating the performance of specific fusion algorithms. To the best of our knowledge, comprehensive studies that compare a wide variety of image fusion algorithms and fusion IQA models have not been reported in the literature.

In this work, we first build a database that contains images of multiple exposure levels, together with fused images produced by different MEF algorithms. A subjective user study is then conducted using the database. The significance of the database and the subjective experiment is three-fold. First, it provides useful data to study human behaviors in evaluating fused image quality; Second, it supplies a test set to evaluate and compare the relative performance of classical and state-of-the-art MEF algorithms; Third, it is useful to validate and compare the performance of existing objective IQA algorithms in predicting the subjective quality of fused images. This will in turn provide insights on potential ways to improve them.

2. SUBJECTIVE QUALITY ASSESSMENT

2.1. MEF Image Database

Seventeen high-quality natural images of maximum size of 512 × 768 are selected to cover diverse image content including natural sceneries, indoor and outdoor views, and man-made architectures. All source images are shown in Fig. 1. Note that the multi-exposure source image sequences typically contain more than 3 input images that are either underexposed or overexposed. For visualization purpose, in Fig. 1, we selected the best quality fused image in terms of subjective evaluations to represent each source image.

Eight fusion algorithms are selected, which include simple operators such as 1) local energy weighted linear combination and 2) global energy weighted linear combination, as well as advanced MEF algorithms such as 3) Raman09 [5], 4) Gu12 [8], 5) ShutaoLi12 [10], 6) ShutaoLi13 [11], 7) Li12 [9], and 8) Mertens07 [4]. These algorithms are chosen to cover a diverse types of MEF methods in terms of methodology and behavior. In all cases, default parameter settings are adopted without tuning for better quality. Eventually, a total of 136 fused images are generated, which are divided into 17 image sets of 8 images each, where the images in the same set are created from the same source image sequence. An example is shown in Fig. 2, which includes a source image sequence at three exposure levels Fig. 2(a1-a3) and the fused images generated by eight fusion algorithms Fig. 2(b-i).

2.2. Subjective Study

The subjective testing environment was setup as a normal indoor office workspace with ordinary illumination level, with no reflecting ceiling walls and floor. All image are displayed on an LCD monitor at a resolution of 2560×1600 pixel with Truecolor (32bit) at 60Hz. The monitor was calibrated in accordance with the recommendations of ITU-T BT.500 [26]. The display is controlled by a desktop PC with Intel(R) Core(TM) i7-2600 dual 3.40GHz CPU. A customized Matlab figure window was used to render the images on the screen. During the test, all 8 fused images are shown to the subject at the same time on one computer screen at actual pixel resolution but in random spatial order. The study adopted a multi-stimulus quality scoring strategy without showing the reference image. A total of 25 naïve observers, including 15 male and 10 female subjects aged between 22 and 30, participated in the subjective experiment. The subjects are allowed to move their positions to get closer or further away from the screen for better observation. All subject ratings were recorded with pen and paper during the study. To minimize the influence of fatigue effect, the length of a session was limited to a maximum of 30 minutes.

For each image set, the subject was asked to give an integer score that best reflects the perceptual quality of each fused image. The score ranges from 1 to 10, where 1 denotes the worst quality and 10 is the best. Compared with paired-comparison and ranking-based testing strategies, the advantages of this method is manifold. First, it has high efficiency because multiple images are shown on the same screen and multiple scores are collected at one time. Second, it reduces memory effect because a full set of images are evaluated on one screen, making it easier for the subjects to apply the same scoring strategy to all images, as opposed to the case when images from the same set are shown on different screen views at different times in a test session, and the subjects may forget their scoring strategies used previously. Third, the results have broad usage in performance evaluation, because the absolute category ratings being collected also inherently contain ranking information. As a result, both linear and rank-order correlation evaluations can be directly applied in data analysis stage. Finally, the results also have broad usage in algorithm development, because quality comparison across source images of different content is more meaningful, which is helpful in the development of objective IQA models to test and improve their generalization capabilities.

3. ANALYSIS AND DISCUSSION

3.1. Subjective Data Analysis

After the subjective user study, 2 outlier subjects were removed based on the outlier removal scheme in [26], resulting in 23 valid subjects. The final quality score for each individual image is computed as the average of subjective scores, namely the mean opinion score (MOS), from all valid subjects. Considering the MOS as the "ground truth", the performance of individual subjects can be evaluated by 1) comparing their quality measure with the "ground truth" for all test images, and 2) calculating the correlation coefficient between individual subject ratings and MOS values for each image set, and then averaging the correlation coefficients of all image sets.

Pearson linear correlation coefficient (PLCC) and Spearman's rand-order correlation coefficient (SRCC) are employed as the evaluation criteria. Both criteria range from 0 to 1, where higher values indicate better performance. Table 2 listed the PLCC and SRCC results for all individual subjects. Although the behaviors of individual subjects varies, there is generally a considerable agreement between them on the quality of fused images.

Table 1. Consistency between individual and average subject scores

Subject	PLCC	SRCC	Subject	PLCC	SRCC
1	0.8743	0.8631	13	0.8411	0.7989
2	0.8245	0.7984	14	0.8781	0.8743
3	0.7102	0.6735	15	0.8988	0.8924
4	0.8093	0.8182	16	0.7413	0.7313
5	0.6785	0.6649	17	0.7347	0.6488
6	0.6544	0.6567	18	0.7797	0.7486
7	0.8198	0.8030	19	0.6732	0.6814
8	0.8951	0.8849	20	0.7854	0.7643
9	0.7961	0.7835	21	0.6045	0.5638
10	0.6924	0.6826	22	0.6213	0.6121
11	0.8298	0.8275	23	0.7976	0.7558
12	0.6145	0.5795	Average	**0.7633**	**0.7438**

To further investigate the performance of individual subjects, we compute PLCC and SRCC values for each image set. As such, for each individual subject, we obtain their PLCC and SRCC results for 17 image sets. The mean and standard deviation (std) of these results are depicted in Fig. 3. It can be seen that each individual subject performs quite consistently with relatively low variations for different image content. The average performance across all individual subjects is also given in the rightmost column of Fig. 3. This provides a general idea about the performance of an average subject (Here an "average subject" should not be confused with the MOS values of all subjects. An "average subject" is used to summarize the behavior of a typical subject, whose behavior is expected to deviate from the average behavior of all subjects).

3.2. Performance of MEF Algorithms

We use the MOS values given to the 8 MEF algorithms described in Section 2.1 to evaluate and compare their performance. The mean and std of MOS values over all 17 image sets are summarized in Fig. 4. It is worth mentioning that this only provides a rough comparison of the relative performance of the MEF algorithms, where default parameters are used without fine tuning. Besides, computational complexity is not a factor under consideration.

From the subjective test results, we have several observations. First, from the sizes of the error bars, we observe that subjects agree with each other to a significant extent on the performance of any individual MEF algorithm, but the performance difference between different MEF algorithms is sometimes small (when compared with the error bars). Second, Mertens's method [4] achieves the best performance on average, while Li's method [9], which is the second best on average, is actually a detail-enhanced algorithm built upon Mertens's method [4]. It has very similar average performance and a larger error-bar than Mertens's method [4]. This suggests that detail enhancement might be useful to create perceptually appealing results on some images, but may also create unwanted artifacts in some other images, and the overall performance gain is not reliable in the current approaches. Third, comparing local energy weighting with global energy weighting approaches, the former focuses

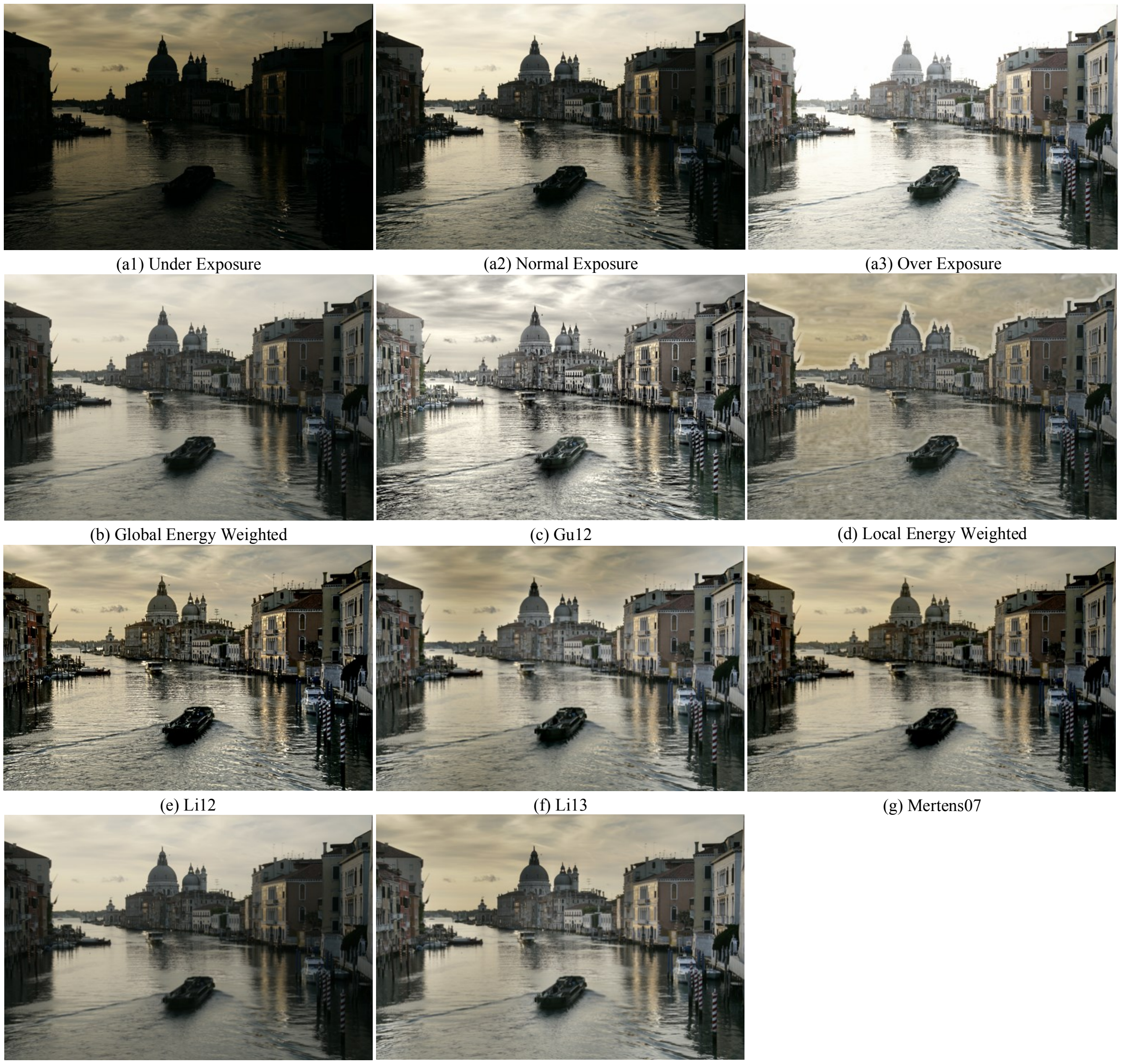

Fig. 2. An example of multi-exposure input images (a1, a2, a3) and fused images (b)-(i) created by different MEF algorithms.

more on enhancing local structures while the latter emphasizes more on global spatial consistency. The large performance gap between them indicates that maintaining spatial consistency may be an indispensable factor in determining the quality of fused image. Fourth, it is somewhat surprising that some of the advanced algorithms, such as Raman09 [5] and Gu12 [8], perform similarly to simple global energy-based weighting. Fifth, not a single algorithm produces the fused images with the best perceptual quality for all image sets. This suggests that there is still room for future improvement, and proper combination of the ideas used in different MEF algorithms has the potential to further improve the performance.

3.3. Performance of Objective IQA Models

We test 9 objective IQA models for image fusion, for which Section 1 only provides a rough introduction. An excellent survey can be found in [27]. All models tested here are designed for general-purpose image fusion, not specifically for MEF. The algorithms were elaborated with the source sequence containing two input images only. Fortunately, most of these algorithms can be directly extended to the cases of multiple input images. Models that cannot be extended such as [20, 21] are excluded. For the purpose of fairness, all models are tested using their default parameter settings. Note that to obtain a reasonable result, we take the absolute value of the objective score in [16].

Table 2 summarizes the evaluation results, which is somewhat disappointing because state-of-the-art IQA models do not seem to provide adequate predictions of perceived quality of fused images. Even the models with the best performance, such as Xydeas's [14] and Wang's [15] methods, are only moderately correlated with subjective scores. Somewhat surprisingly, some models even give negative correlations. The scatter plots of MOS versus the four objec-

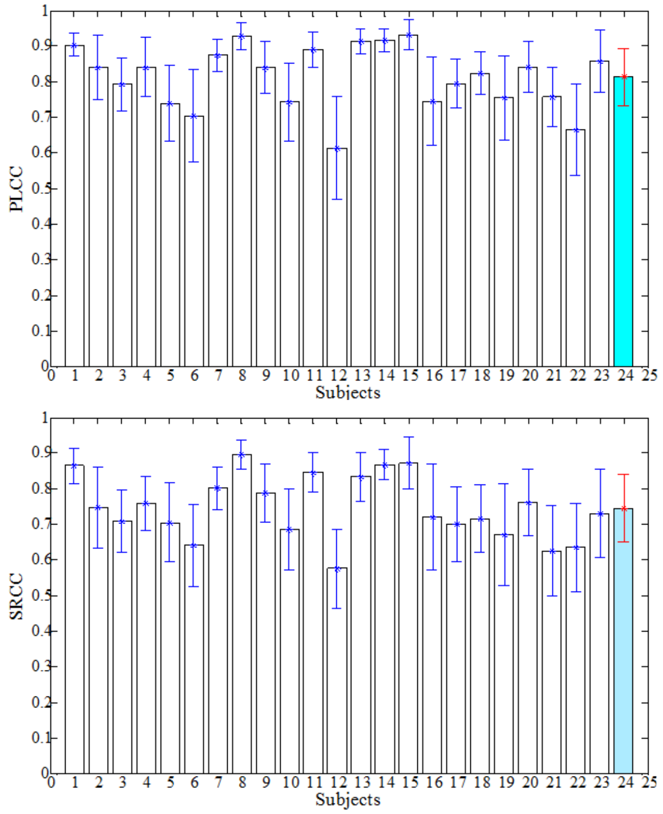

Fig. 3. PLCC and SRCC between individual subject and MOS. Rightmost column: performance of an average subject.

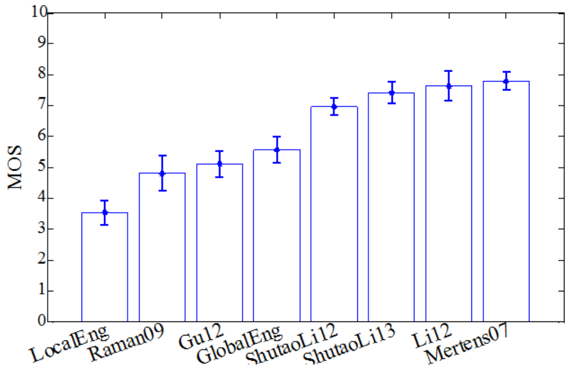

Fig. 4. Mean and std of subjective rankings of individual image fusion algorithms across all image sets.

tive models associated with the best performance are given in Fig 5, where each point denotes one test image. The widespread of the scatter plots suggests that there is still a long way to go in the development of IQA models that are useful in image fusion applications.

The above test results also provide some useful insights regarding the general approaches used in IQA models. First, models based on entropy computations of pixel intensity values and transform coefficients [28, 29] have poor correlation with perceptual quality. The reason may be that the quality of fused images is highly content dependent and only entropy of image intensity/coefficient histogram is insufficient in capturing the perceptual distortions introduced by MEF processes. Second, local structure-preservation based models, such as SSIM and gradient based approaches applied in spatial or transform domain [14, 15, 16, 18], provide the most promising results so far. However, they are often unsuccessful in capturing the degradations of spatial consistency across the image space. This suggests that more accurate objective IQA models may be developed by achieving a good compromise between assessing local structure preservation and evaluating global spatial consistency.

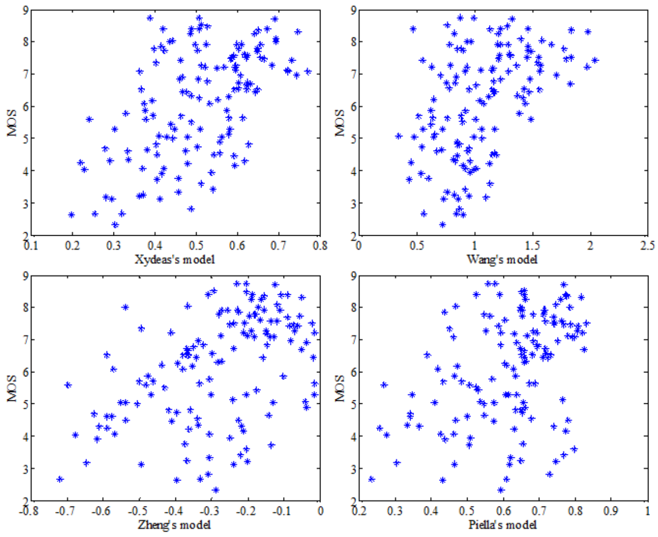

Fig. 5. MOS versus the objective IQA models (Xydeas' [14], Wang's [15], Zheng's [16], Piella's [18]) of the best performance.

Table 2. Performance evaluation of objective IQA models

IQA model	PLCC		SRCC	
	μ	σ	μ	σ
Hossny's [28]	-0.2939	0.2054	-0.2784	0.2803
Cvejic's [29]	0.0604	0.4311	0.0590	0.4968
Wang's [30]	-0.2992	0.2008	-0.2524	0.2976
Xydeas's [14]	0.6949	0.1655	0.6198	0.2452
Wang's [15]	0.6356	0.1634	0.5771	0.1761
Zheng's [16]	0.4332	0.2317	0.4614	0.1820
Piella's [18]	0.3798	0.2409	0.4131	0.1725
Chen's [22]	-0.5544	0.4089	-0.5611	0.4640
Chen's [23]	0.2667	0.4830	0.3274	0.4628

4. CONCLUSION

Image fusion has been an active research topic in the past decade, and a significant number of image fusion and objective IQA methods have been proposed. However, comprehensive validation and comparison of these algorithms are lacking. In this study, we made one of the first attempts dedicated to the evaluation and comparison of classical and state-of-the-art MEF and relevant IQA algorithms. A new MEF image database is established and subjective tests are conducted. Our results suggest that human subjects generally have significant agreement with each other. The subjective scores are used to test the performance of existing MEF algorithms and provide useful insights on the perceptual relevance of the specific approaches used in these algorithms. Perhaps the most important finding of the current work is that none of the classical and state-of-the-art objective IQA models developed for image fusion achieves good correlation with subjective opinions. This motivates us to design advanced objective quality models for image fusion, for which we learned from this study that a good balance between global spatial consistency and local structure preservation is desirable.

5. REFERENCES

[1] P. J. Burt, *The pyramid as a structure for efficient computation*. Springer, 1984.

[2] P. J. Burt and R. J. Kolczynski, "Enhanced image capture through fusion," in *Computer Vision, 1993. Proceedings., Fourth International Conference on*, pp. 173–182, IEEE, 1993.

[3] A. A. Goshtasby, "Fusion of multi-exposure images," *Image and Vision Computing*, vol. 23, no. 6, pp. 611–618, 2005.

[4] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion: A simple and practical alternative to high dynamic range photography," in *Computer Graphics Forum*, vol. 28, pp. 161–171, Wiley Online Library, 2009.

[5] S. Raman and S. Chaudhuri, "Bilateral filter based compositing for variable exposure photography," in *Proc. Eurographics*, pp. 1–4, 2009.

[6] M. Song, D. Tao, C. Chen, J. Bu, J. Luo, and C. Zhang, "Probabilistic exposure fusion," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 341–357, 2012.

[7] W. Zhang and W.-K. Cham, "Gradient-directed multiexposure composition," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2318–2323, 2012.

[8] B. Gu, W. Li, J. Wong, M. Zhu, and M. Wang, "Gradient field multi-exposure images fusion for high dynamic range image visualization," *Journal of Visual Communication and Image Representation*, vol. 23, no. 4, pp. 604–610, 2012.

[9] Z. Li, J. Zheng, and S. Rahardja, "Detail-enhanced exposure fusion," *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4672–4676, 2012.

[10] 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.

[11] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2864–2875, 2013.

[12] R. Shen, I. Cheng, and A. Basu, "Qoe-based multi-exposure fusion in hierarchical multivariate gaussian crf," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2469–2478, 2013.

[13] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electronics letters*, vol. 38, no. 7, pp. 313–315, 2002.

[14] C. S. Xydeas and V. S. Petrovic, "Objective pixel-level image fusion performance measure," in *AeroSense 2000*, pp. 89–98, International Society for Optics and Photonics, 2000.

[15] P.-w. Wang and B. Liu, "A novel image fusion metric based on multi-scale analysis," in *Signal Processing, 2008. ICSP 2008. 9th International Conference on*, pp. 965–968, IEEE, 2008.

[16] Y. Zheng, E. A. Essock, B. C. Hansen, and A. M. Haun, "A new metric based on extended spatial frequency and its application to dwt based fusion algorithms," *Information Fusion*, vol. 8, no. 2, pp. 177–192, 2007.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[18] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 3, pp. III–173, IEEE, 2003.

[19] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[20] N. Cvejic, A. Loza, D. Bull, and N. Canagarajah, "A similarity metric for assessment of image fusion algorithms," *International journal of signal processing*, vol. 2, no. 3, pp. 178–182, 2005.

[21] C. Yang, J.-Q. Zhang, X.-R. Wang, and X. Liu, "A novel similarity based quality metric for image fusion," *Information Fusion*, vol. 9, no. 2, pp. 156–160, 2008.

[22] H. Chen and P. K. Varshney, "A human perception inspired quality metric for image fusion based on regional information," *Information fusion*, vol. 8, no. 2, pp. 193–207, 2007.

[23] Y. Chen and R. S. Blum, "A new automated quality assessment algorithm for image fusion," *Image and Vision Computing*, vol. 27, no. 10, pp. 1421–1432, 2009.

[24] A. Toet and E. M Franken, "Perceptual evaluation of different image fusion schemes," *Displays*, vol. 24, no. 1, pp. 25–37, 2003.

[25] V. Petrović, "Subjective tests for image fusion evaluation and objective metric validation," *Information Fusion*, vol. 8, pp. 208–216, Apr. 2007.

[26] I.-R. BT.500-12, "Recommendation: Methodology for the subjective assessment of the quality of television pictures," Nov. 1993.

[27] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 94–109, 2012.

[28] M. Hossny, S. Nahavandi, and D. Creighton, "Comments on information measure for performance of image fusion," *Electronics letters*, vol. 44, no. 18, pp. 1066–1067, 2008.

[29] N. Cvejic, C. Canagarajah, and D. Bull, "Image fusion metric based on mutual information and tsallis entropy," *Electronics letters*, vol. 42, no. 11, pp. 626–627, 2006.

[30] Q. Wang, Y. Shen, and J. Jin, "Performance evaluation of image fusion techniques," *Image Fusion: Algorithms and Applications*, pp. 469–492, 2008.