PERCEPTUAL EVALUATION OF SINGLE-IMAGE SUPER-RESOLUTION RECONSTRUCTION

Guangcheng Wang[1], *Leida Li*[1], *Qiaohong Li*[2], *Ke Gu*[3], *Zhaolin Lu*[4] *and Jiansheng Qian*[1]

[1]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, 221116, China
[2]School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore
[3]BJUT Faculty of Information Technology, Beijing University of Technology, 100124, China
[4]Advanced Analysis and Computation Center, China University of Mining and Technology, 221116, China

ABSTRACT

In recent years, single-image super-resolution (SR) reconstruction has aroused wide attention. Massive SR enhancement algorithms have been proposed. However, much less work has been down on the perceptual evaluation of SR enhanced images and the corresponding enhancement algorithms. In this work, we create a Super-resolution Reconstructed Image Database (SRID), which consists of images produced by two interpolation methods and six popular S-R image enhancement algorithms at different amplification factors. Then, subjective experiment is conducted to collect the subjective scores by using the single-stimulus method. The performances of the SR image enhancement algorithms are then evaluated by the obtained subjective scores. Finally, the performances of the general-purpose no-reference (NR) image quality metrics are investigated on the SRID database. This study shows that it is difficult for the state-of-the-art NR image quality metrics to predict the quality of SR enhanced images.

Index Terms— Super-resolution reconstruction, Image quality assessment, Database, No-reference.

1. INTRODUCTION

Image super-resolution (SR) is to estimate a high-resolution (HR) image using one or several low-resolution (LR) images from a real scene [1]. The technique has a wide range of applications, including computer vision, medical and remote sensing imaging, video surveillance, and entertainment. Therefore, it has attracted much attention in recent two decades. With substantial SR image enhancement algorithms proposed, a problem that how to find out the best SR image enhancement algorithms is presented.

With the rapid advances of SR image enhancement, the relevant quality assessment of SR enhanced images should also be taken into consideration. In practice, the classical peak signal-to-noise-ratio (PSNR) and the structural similarity (S-SIM) index [2] are usually used to evaluate the quality of SR enhanced images. However, the difficulty lies in the fact that a perfect quality HR image is unavailable to compare with in real-world scenarios. As a result, the common full-reference (FR) approaches [2, 3, 4, 5] are not readily applicable. With this consideration, it is necessary to study the performance of the existing NR image quality metrics on SR enhanced images.

To answer the aforementioned problems systematically, we establish a Super-resolution Reconstructed Image Database (SRID), which includes images produced by two interpolation methods and six popular SR image enhancement algorithms at different amplification factors. The performances of two interpolation methods and six popular SR image enhancement algorithms are evaluated based on the result of the subjective experiment, which is performed using the single-stimulus method. Finally, the performances of the state-of-the-art NR image quality metrics are evaluated based on the SRID database. The experimental results show that the existing metrics are only moderately correlated with the subjective ratings.

2. SUPER-RESOLUTION RECONSTRUCTED IMAGE DATABASE (SRID)

A SRID database is built to evaluate the performances of two interpolation methods, six popular SR image enhancement algorithms and the existing NR image quality metrics.

2.1. Selection of Images and SR Algorithms

Twenty LR natural images with diversified contents are selected, which include animal, natural scenery, building, human, etc. These LR images are shown in Fig. 1. Two interpolation algorithms and six SR image enhancement algorithms are used to generate the HR images, including Nearest Interpolation, Bicbuic Interpolation [6], Iterative

This work is supported by the National Natural Science Foundation of China (61379143, 51604217), National Key Research and Development Program of China (2016YFC0801808) and the Qing Lan Project. (*Corresponding author: Leida Li, reader1104@hotmail.com)*

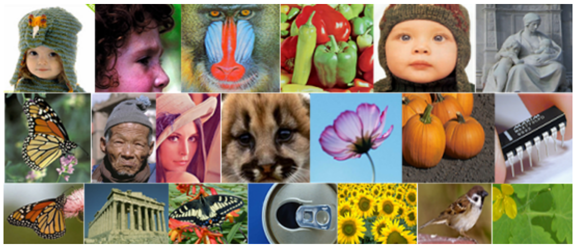

Fig. 1. Twenty LR images used to build the database.

Curvature-based Interpolation (ICBI) [7], Coupled Dictionary Training for Image Super-Resolution (SCSR) [8], Gaussian Process Regression for Super-Resolution (GPRSR) [9], Adaptive Gradient Magnitude Self-Interpolation for Edge-Directed Single-Image Super-Resolution (AGMSSR) [10], Local Fractal Analysis of Gradient for Single-Image Super-Resolution (LFAGSR) [11], and Fuzzy-Rule-Based Approach for Single-Frame Super-Resolution (FRBSR) [12]. In order to generate HR images with different distortion levels, the LR images are processed using the two interpolation methods and six SR enhancement algorithms with three different amplification factors, namely 2, 4 and 8. Finally, the HR images constitutes the SRID [1] database.

The SR enhanced lena images shown in Fig. 2 are enhanced by Bicbuic Interpolation [6], GPRSR [9] and FRBSR [12]. It can be seen from the figure that the quality of the SR enhanced images degrades in accordance with the increase of amplification factors. The SR reconstructed images are usually corrupted by multiple distortions, which include blur, ringing, and local texture unnaturalness. Therefore, it is difficult for distortion-specific image quality evaluation models to evaluate the quality of SR reconstructed images [13, 14].

2.2. Subjective Experiment

Due to the lack of ground truth as the reference image in real applications, the single-stimulus (SS) has been employed according to International Telecommunications Union (ITU) recommendation [15]. The SS method has been recently used in building databases of deblocked images [16]. Twenty non-expert subjects participated in the subjective experiment (all without visual impairment). Their ages range from 20 to 30. All SR enhanced images displayed in their original resolutions are shown to the participants in random order. At the beginning of the subjective experiment, the range of quality levels was illustrated through a set of training examples. A MATLAB graphical user interface (GUI) is developed to perform the subjective test, which is shown in Fig. 3. The subjective score can be automatically saved in a sheet after pressing the *Submit* button.

[1]We will make the database freely public to the research community.

Fig. 2. An example of SR reconstructed images. (a), (b) and (c) are reconstructed via Bicbuic Interpolation [6]. (d), (e) and (f) are reconstructed via GPRSR [9]. (g), (h) and (i) are reconstructed via FRBSR [12].

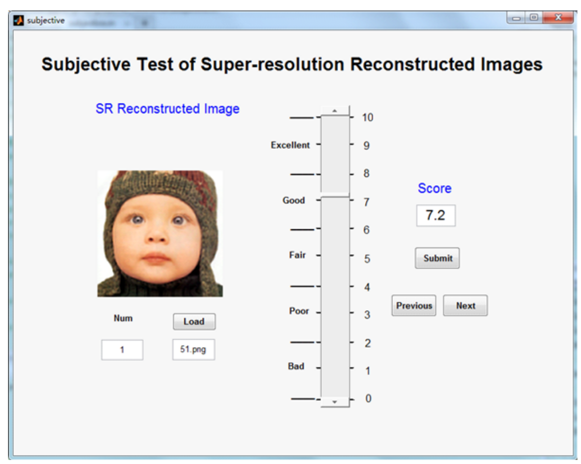

Fig. 3. Graphical user interface used to rate image quality.

2.3. Processing and Analysis of Subjective Scores

In order to obtain accurate subjective ratings, we remove the maximum and minimum scores for each image. Then, the mean opinion score (MOS) is computed as the final ground truth of image quality.

The histogram distribution of the MOS values are shown in Fig. 4. It can be found from the figure that the MOS values cover the whole range. This indicates that the SRID contains SR reconstructed images with different distortion levels.

Considering the effect of subjective tests, two common metrics are employed to evaluate the performance of the sub-

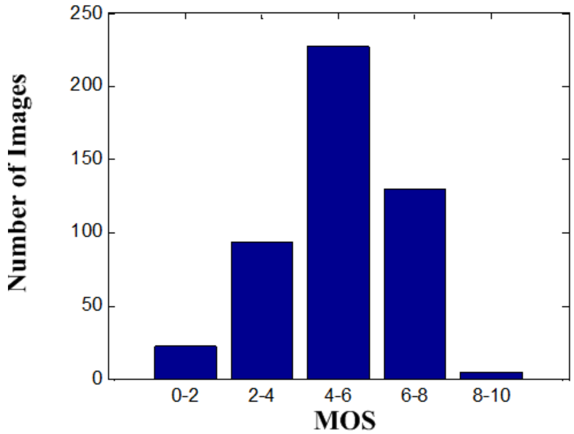

Fig. 4. Histogram distribution of the MOS values.

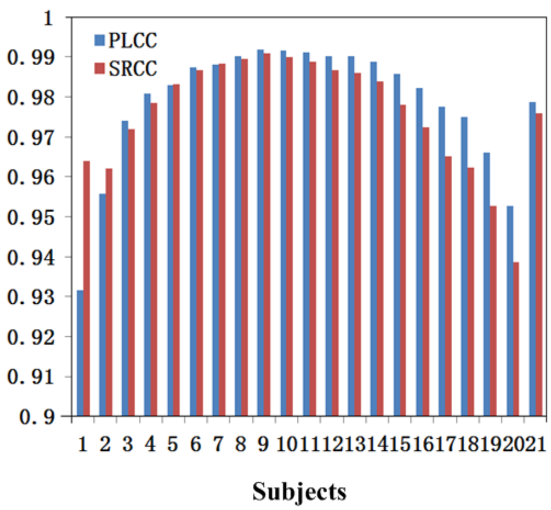

Fig. 5. PLCC and SRCC between the subjective scores and the MOS values.

jective experiment. The first is the Spearmans rand-order correalation coefficient (SRCC) between the subjective scores and the MOS values. SRCC measures the prediction monotonicity. The second metric is the Pearson linear correlation coefficient (PLCC) between the subjective scores and the MOS values following a nonlinear regression. The nonlinear regression used is as follows [17]:

$$f(x) = \tau_1(\frac{1}{2} - \frac{1}{1+e^{\tau_2(x-\tau_3)}}) + \tau_4 x + \tau_5, \qquad (1)$$

where $\tau_i, i = 1, 2, 3, 4, 5$ are the parameters to be fitted.

Fig. 5 shows the PLCC and SRCC values between the subjective scores and the MOS values. The PLCC and SRCC in the rightmost column are computed between the average of all subjective scores and the MOS values. It is worth noting that all PLCC and SRCC values are higher than 0.9, which

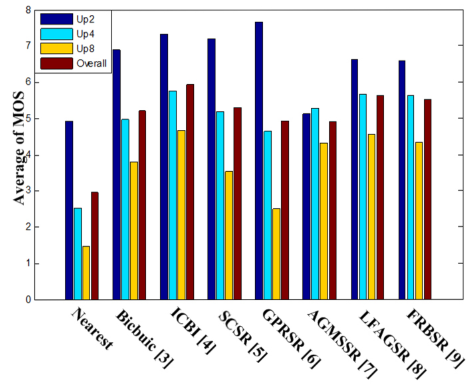

Fig. 6. Average MOS values for two interpolation methods and six popular SR image enhancement algorithms.

indicates that the subjects are quite consistent in rating the qualities of the images.

3. PERFORMANCE EVALUATION OF SR IMAGE ENHANCEMENT ALGORITHMS

Each image is processed by two interpolation methods and six popular SR image enhancement algorithms at different amplification factors. If the SR reconstructed images have higher subjective scores, the corresponding SR image enhancement algorithm has better performance, vice versa. Therefore, the performances of the SR image enhancement algorithms can be evaluated by systematacially analyzing the results of the subjective test. Since the SRID contains images with three different amplification factors, i.e. 2, 4 and 8, we calculate the average MOS values for each amplification factor and across all amplification factors to evaluate the performances of the eight approaches. The results are shown in Fig. 6.

It is observed from Fig. 6 that for two interpolation methods and six popular SR image enhancement algorithms, with the rise of the amplification factors, the average MOS values of the SR reconstructed images decrease accordingly.

In order to explicitly know the performance differences at each amplification factor, we further rank the performances of the two interpolation methods and six popular SR image enhancement algorithms for the amplification factors of 2, 4 and 8, respectively. Table 1 lists the performance rankings of two interpolation methods and six popular SR image enhancement algorithms in three different amplification factors. The overall performance rankings are also given. It is observed from Table 1 that for all different amplifying factors, nearest interpolation produces the worst results. Apart from the scaling factor 2, ICBI [7] produces the best results. In most cases, the overall performance rankings are similar to the results for the amplification factors of 4 and 8.

Table 1. Performance rankings of two interpolation methods and six popular SR image enhancement algorithms.

Algorithm	2*	4*	8*	Overall
Nearest	8	8	8	8
Bicbuic [6]	4	6	5	5
ICBI [7]	2	1	1	1
SCSR [8]	3	5	6	4
GPRSR [9]	1	7	7	6
AGMSSR [10]	7	4	4	7
LFAGSR [11]	5	2	2	2
FRBSR [12]	6	3	3	3

Table 2. Performances of state-of-the-art NR image quality metrics on SRID database.

Metrics	PLCC	SRCC	RMSE
BRISQUE [18]	0.6738	**0.6666**	1.1953
NIQE [19]	0.5247	0.4759	1.3769
NFERM [20]	0.6011	0.6177	1.2927
BIQI [21]	0.4253	0.4336	1.2682
BLIINDS2[22]	0.3783	0.3687	1.4973
CORNIA [23]	**0.6767**	0.5985	**1.1909**
DESIQUE [24]	0.5253	0.5453	1.3763
DIIVINE [25]	0.4286	0.4826	1.4614
ILNIQE [26]	0.4136	0.4233	1.4729
SISBLIM [27]	0.6223	0.5965	1.2661

4. EVALUATION OF EXISTING NR-IQA METRICS

Since in real applications there is no ground truth that could be used as a reference, only the existing state-of-the-art NR image quality metrics are taken into consideration in this study. The tested NR image quality metrics include Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [18], Natural Image Quality Evaluator (NIQE) [19], NR Free Energy based Robust Metric (NFERM) [20], Blind Image Quality Index (BIQI) [21], Blind Image Integrity Notator using DCT Statistics (BLIINDS2) [22], COdebook Representation for No-reference Image Assessment (CORNIA) [23], Derivative Statistics-based Image QUality Evaluator (DESIQUE) [24], Distortion Identification-based Image Verity and Integrity Evaluation (DIIVINE) Index [25], Integrated Local Natural Image Quality Evaluator (ILNIQE) [26] and SIx-Step BLInd Metric (SISBLIM) [27]. Three commonly used criteria are used to estimate the performances of the quality metrics, which are SRCC, PLCC and root mean square error (RMSE). The SRCC, PLCC and RMSE are computed after a nonlinear mapping [17] between the objective and subjective scores. The experimental results are summarized in Table 2, where the best results are marked in boldface. It is easily observed that the existing NR image quality metrics are quite limited in evaluating the quality of SR reconstructed images. Although BRISQUE [18] and CORNIA [23] deliver the best monotonicity and accuracy, the SRCC and PLCC values are only around 0.6666 and 0.6767 respectively, which are far from ideal. Quality models specifically designed for SR reconstructed images are still needed.

5. CONCLUSION

In this paper, a database of SR reconstructed images has been built. Then, the subjective experiment is carried out using the single-stimulus method. The obtained MOS values are then used to evaluate the performances of the two interpolation methods and six popular SR image enhancement algorithms. Finally, the performances of the popular NR image quality metrics are evaluated on the SRID database. The experimental results shows that it is still quite limited for the state-of-the-art quality metrics to predict the quality of SR reconstructed images. This indicates that quality models specifically designed for SR reconstructed images are highly needed, which will be our future work.

6. REFERENCES

[1] T. S. Huang and P. Y. Tsai, "Multi-frame image restoration and registration," *Adv. Comput. Vis. Image Process.*, vol. 1, no. 2, pp. 317-339, 1984.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 35-44, 2004.

[3] L. Zhang, L. Zhang, X. Q. Mou and D. Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378-2386, 2011.

[4] L. Li, H. Cai, Y. Zhang, W. Lin, A. C. Kot and X. Sun, "Sparse Representation-Based Image Quality Index With Adaptive Sub-Dictionaries," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3775-3786, 2016.

[5] K. Gu, L. Li, H. Lu, X. Min and W. Lin, "A Fast Reliable Image Quality Predictor by Fusing Micro- and Macro-Structures," *IEEE Trans. Industrial Electronics.*, vol. 64, no. 5, pp. 3903-3912, 2017.

[6] H. S. Hou and H. C. Andrews, "Cubic splines for image interpolation and digital filtering," *IEEE Trans. Signal Process.*, vol. 26, no. 6, pp. 508-517, 1978.

[7] A. Giachetti and N. Asuni, "Real-time artifact-free image upscaling," *IEEE Trans. Image Process.*, vol. 20, no. 10, pp. 2760-2768, 2011.

[8] J. Yang, Z. Wang, Z. Lin, S. T. Cohen and T. S. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3467-3478, 2012.

[9] H. H and W. C. Siu, "Single image super-resolution using Gaussian process regression," *in Proc. Int. Conf. Comput. Vis. Pattern Recognit. (CVPR).*, pp. 449-456, 2011.

[10] L. Wang, S. Xiang, G. Meng, H. Wu and C. H. Pan, "single-image super-resolution via adaptive gradient magnitude self-interpolation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 8, pp. 1289-1299, 2013.

[11] H. Xu, G. Zhai and X. K. Yang, "Single image super-resolution with detail enhancement based on local fractal analysis of gradient," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 10, pp. 1740-1299, 2013.

[12] P. Purkait, N. R. Pal and B. Chanda, "A fuzzy-rule-based approach for single frame super resolution," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2277-2290, 2014.

[13] L. Li, W. Lin, X. Wang, G. Yang, K. Bahrami and A. C. Kot, "No-reference image blur assessment based on discrete orthogonal moments," *IEEE Trans. Cybernetics.*, vol. 46, no. 1, pp. 39-50, 2016.

[14] L. Li, H. Zhu, G. Yang and J. Qian, "Referenceless Measure of Blocking Artifacts by Tchebichef Kernel Analysis," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 122-125, 2014.

[15] ITU, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation, International Telecommunication Union/ITU Radio communication Sector*, 2009.

[16] L. Li, Y. Zhou, W. Lin, J. Wu, X. Zhang and B. Chen, "No-reference quality assessment of deblocked images," *Neurocomputing*, vol. 177, pp. 572-584, 2016.

[17] (Aug. 2003). *Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II.* [Online]. Available: http://www.vqeg.org

[18] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695-4708, 2012.

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

[20] K. Gu, G. Zhai, X. K. Yang and W. J. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Mulultimedia.*, vol. 17, no. 1, pp. 50-63, 2015.

[21] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513-516, 2010.

[22] M. A. Saad and A. C. Bovik, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339-3352, 2012.

[23] P. Ye, J. Kumar, L. Kang and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," *in Proc. Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1098-1105, 2012.

[24] Y. Zhang and D. M. Chandler, "No-reference image quality assessment based on log derivative statistics of natural scenes," *Journal of Electronic Imaging.*, vol. 22, no. 4, pp. 451-459, 2013.

[25] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350-3364, 2011.

[26] L. Zhang, L. Zhang and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579-2591, 2015.

[27] K. Gu, G. Zhai, X. Yang and W. Zhang, "Hybrid No-Reference Quality Metric for Singly and Multiply Distorted Images," *IEEE Trans. Broadcasting.*, vol. 60, no. 3, pp. 555-567, 2014.