Published in: Proceedings of the 47th IEEE Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, Nov. 3-6, 2013. ©IEEE

PERCEPTUAL EVALUATION OF IMAGE DENOISING ALGORITHMS

Kai Zeng and Zhou Wang

Dept. of Electrical & Computer Engineering, University of Waterloo, Waterloo, ON, Canada

ABSTRACT

Image denoising has been an active research topic in the past decades for its broad real-world applications, but surprisingly little work has been dedicated to the quality assessment of denoised images. In this work, we first build a database that contains noisy images at different noise levels and denoised images created by both classical and state-of-the-art denoising algorithms. We then carry out a subjective experiment using a multi-stimulus ranking approach to evaluate and compare the quality of the denoised images. Data analysis shows that there are both considerable agreement and significant variations between human subjects on their opinions of denoised images. Our results also show that state-of-the-art objective image quality models only moderately correlate with subjective opinions, and further investigations that involve both structural fidelity and naturalness measures are desirable in future development of advanced objective models.

Index Terms— image quality assessment, image denoising, human visual system

1. INTRODUCTION

Real-world images are subject to noise contaminations during acquisition, transmission, and processing. Denoising of images is highly desirable, not only to produce better perceptual quality, but also to help improve the performance of the subsequent processes such as compression, segmentation, resizing, and recognition. In the past decades, a large number of image denoising algorithms have been proposed, ranging from simple linear filtering to sophisticated methods based on advanced statistical image models. With multiple denoising algorithms available, a natural question is which one produces the best quality images. Without an appropriate quality measure, fair comparison is impossible and further improvement is aimless. Surprisingly, in the literature of image quality assessment (IQA), little work has been dedicated to the evaluation of denoised images. In practice, researchers often resort to common IQA measures such as peak signal-to-noise-ratio (PSNR) and the structural similarity index (SSIM) [1] to compare image quality and denoising algorithms, but proper validations of these measures are lacking.

Since human eyes are the ultimate receivers in most applications, subjective test is considered the most reliable approach to evaluate the quality of denoised images. There have been a number of well-known subjective database for image quality assessment (IQA) [2, 3, 4, 5, 6], but unfortunately, none of them contains denoised images as a category. Consequently, direct comparison of different denoising algorithms cannot be performed, and whether existing objective IQA measures are proper quality indicators of denoised images remains an unanswered problem.

The purpose of this work is first to build a database that contains noisy images at different noise levels and denoised images produced by different denoising algorithms. Subjective experiment is then conducted using the database, and the results can be used to 1) study the human behaviors in evaluating denoised image quality; 2) evaluate the relative performance of classical and state-of-the-art denoising algorithms; and 3) test the performance of existing objective IQA algorithms in predicting the subjective quality of denoised images and explore potential ways to improve them.

2. SUBJECTIVE QUALITY ASSESSMENT

2.1. Image Database

Ten original high-quality natural images of size 512×512 shown in Fig. 1 are selected from the CSIQ database [4] to cover diverse natural image content, including humans, animals, plants, natural sceneries, and man-made architectures. Independent white Gaussian noise of three levels is added to each image with noise standard deviation σ_n equaling 15, 30, and 50, respectively.

Eight denoising algorithms are selected, which include simple operators (1) linear Gaussian filter (with the standard deviation of the Gaussian profile equaling 11 pixels) and (2) locally adaptive Wiener filter (MATLAB Wiener2D function with sliding window size 3), as well as state-of-the-art denoising algorithms (3) BLS-GSM [7], (4) SURE-LET [8], (5) BM3D [9], (6) K-SVD [10], (7) SADCT [11], and (8) CSR [12]. These algorithms are chosen to cover a diverse types of denoisers in terms of methodology and behavior. Specifically, the linear Gaussian filter tends to over-smooth the image, resulting in blurriness; the local adaptive Wiener filter is likely to keep structures but does not cleanly remove the noise, especially at high noise levels; SURE-LET often creates non-existing structural artifacts in smooth areas; and

Fig. 1. Original reference images in the database.

BM3D and CSR demonstrate strong noise removal capability while keeping the image contrast, but meanwhile may remove subtle image details as noise. In all cases, default parameter settings are adopted without any tuning for better quality.

In the end, a total of 240 denoised images are generated, which are divided into 30 image sets of 8 images each, where the images in the same set are created from the same original image at the same noise level. For better visualization, a group of cropped and enlarged sample noisy images, together with their corresponding denoised images are shown in Fig. 2.

2.2. Subjective Experiment

The subjective experiment was conducted on a PC with Intel(R) Core(TM) i7-2600 dual 3.40GHz CPU. Images were 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 [13]. The test environment was setup as a normal indoor office workspace with ordinary illumination level. A customized Matlab figure window was used to render the images on the screen. During the test, all 8 denoised images in the same set are shown to the subject at the same time in random spatial order on one computer screen at actual pixel resolution. The order of image sets is also randomized and thus different for each subject.

The study adopted a quality ranking strategy without showing the reference image. A total of 20 naïve observers, mostly graduate students at the University of Waterloo, including 12 males and 8 females aged between 22 and 30, participated in the subjective experiment. For each image set, the subject was asked to rank the perceptual quality of the 8 images from the best to the worst with 8 levels. The subjects have the freedom to move their positions 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 30 minutes.

2.3. Discussion on Subjective Experiment Method

The strategy of subjective experiment is not a trivial issue and is worth further discussion. There could be three methods to conduct the experiment. In a paired comparison method [14], a pair of images are shown at the same time to the subject, who is asked to give preference in terms of quality. It is typically an easy task for the subject, and previous studies showed that it produces reliable results [14]. However, this approach requires a large number of paired comparisons and is very inefficiency. Meanwhile it may cause the transition problem (a judgement of $A > B, B > C, C > A$ given by the same subject). This may not be a major issue if the purpose is to obtain the average preference of many subjects, but becomes a difficult problem to handel when the behavior of an individual subject needs to be analyzed and compared with the mean subject opinions.

In a single-stimulus test method, one image is shown to the subject at one time and the subject is asked to provide a single score on a quality scale. This method directly collects absolute quality scores that are meaningful for cross-content and cross-distortion comparisons. However, the scores collected in such an experiment may be sensitive to the tasks and instructions given to the subjects. Moreover, calibrations across subjects are necessary to align the scores given by different subject.

In a multi-stimulus ranking method, multiple images are presented to the subject together and the subjects rank the images from the lowest to the highest quality. This method has many advantages $-$ It is highly efficient, without any transi-

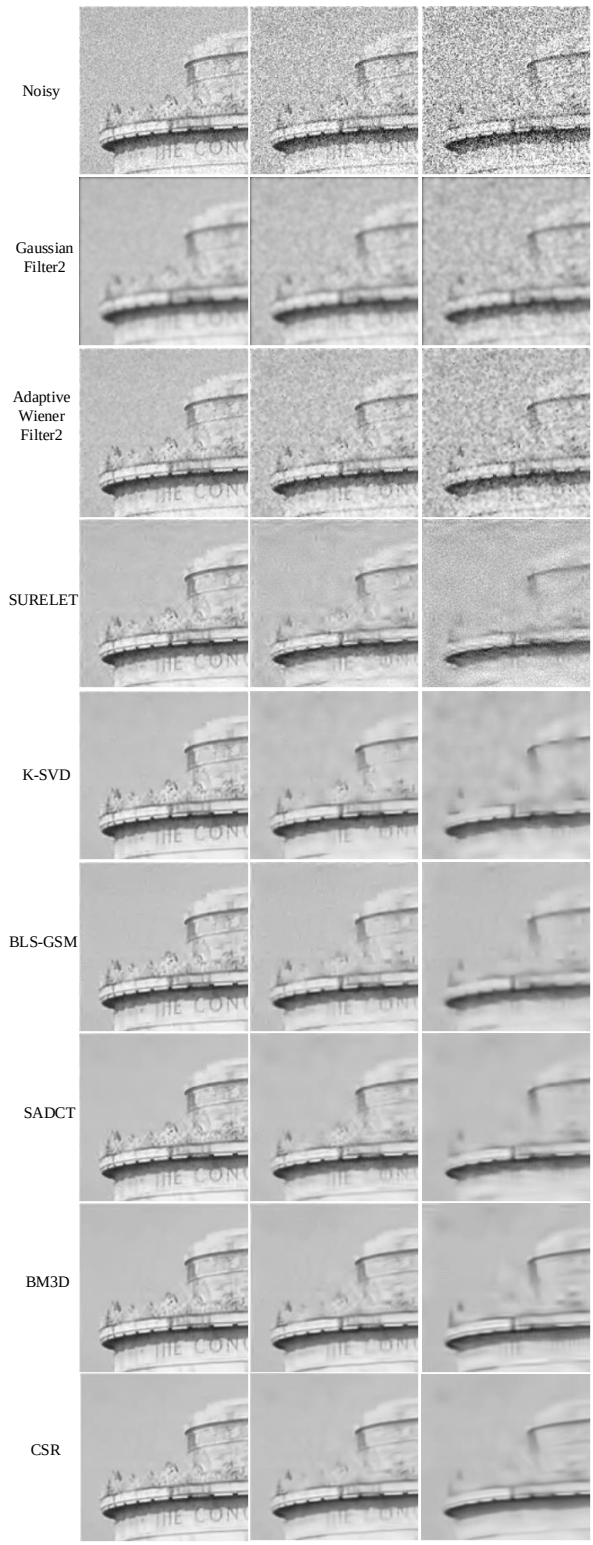

Fig. 2. Sample denoised images (cropped and enlarged for visibility). Left: $\sigma_n = 15$; Middle: $\sigma_n = 30$; Right: $\sigma_n = 50$.

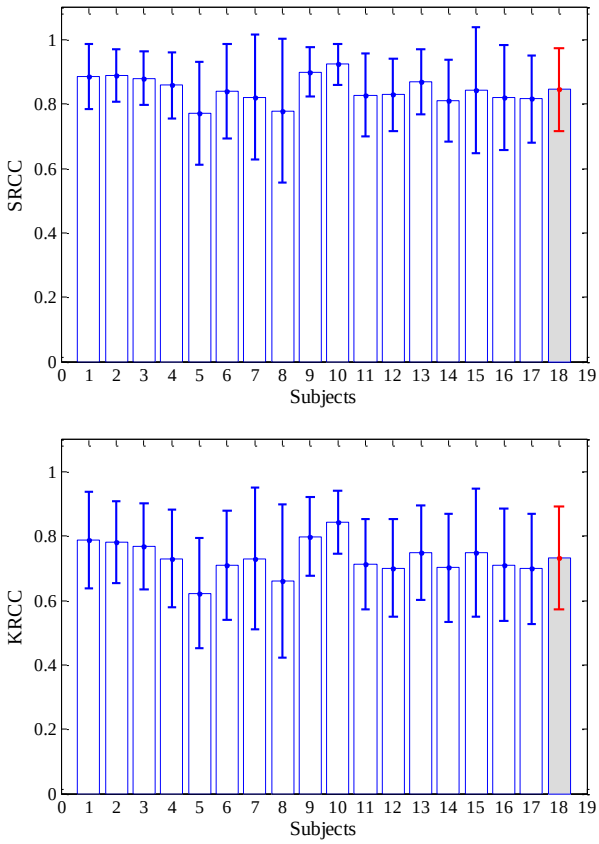

Fig. 3. Mean and std of SRCC and KRCC values between individual subject and average subject rankings. The rightmost column represents the performance of an average subject.

tion problem, and the results can be directly compared across subjects. On the other hand, this approach is often constrained by the physical testing conditions. For example, the screen size may not be enough to show multiple images, and the backlight, viewing angle and viewing distance may not be uniform across the screen.

In reality, the best subjective testing method should be determined by the nature of the experiment to achieve a compromise between effectiveness, accuracy, robustness and efficiency. In the current study, the multi-stimulus ranking method fits well with our target and is thus adopted.

3. ANALYSIS AND DISCUSSION

3.1. Analysis of Subjective Data

After the subjective test, 3 outlier subjects are removed based on the outlier removal scheme in [13], resulting in 17 valid subjects. The final rank-order within each image set is computed as the average ranking from all valid subjects. Considering these average rank-orders for all image sets as the

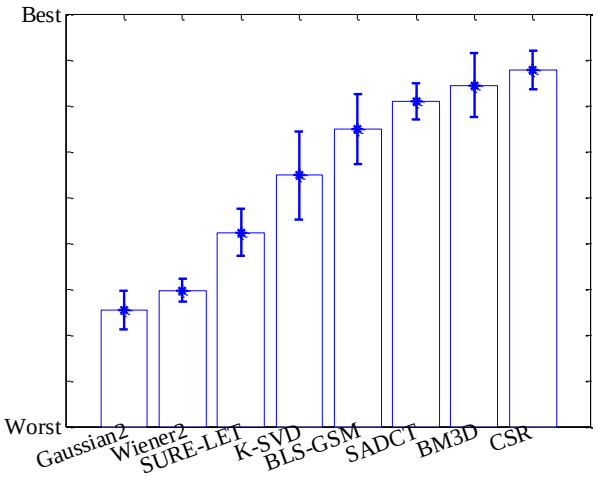

Fig. 4. Mean and std of subjective rankings of individual denoiser across all image sets.

"ground truth", we can observe the performance of each individual subject by comparing their rank-order with the "ground truth" for image set, and then average the performance over all 30 image sets. The comparison is based on Spearman's rank-order correlation coefficient (SRCC) and Kendall's rank-order correlation coefficient (KRCC). The mean and standard deviation of SRCC and KRCC values for each individual subject are depicted in Fig. 3. It can be seen that there is a quite considerable agreement between different subjects on ranking the quality of denoised images. The average performance across all individual subjects is also given in the rightmost column in Fig. 3. This provides a general idea about the performance of an average subject.

The average SRCC and KRCC results of all 8 denoising algorithms over all 30 image sets are summarized in Fig. 4. It can be observed that state-of-the-art denoisers such as BM3D [9] and CSR [12] perform significantly better than more traditional methods. On the other hand, from the sizes of the error bars, we observe significant variations between subject preference of the best denoisers. It is worth mentioning that this only provides a rough comparison of the relative performance of the denoising algorithms, where default parameters are used without fine tuning. Besides, computational complexity is not a factor under consideration.

3.2. Testing Objective IQA Algorithms

Using the database, we test 12 full-reference (FR), 2 reduced-reference (RR), and 4 no-reference (NR) objective IQA models, and the mean (μ) and standard deviation (σ) of SRCC and KRCC values across all 30 image sets are given in Table 1. Larger μ values of SRCC and KRCC indicate better consistency with subjective opinions, while smaller σ values suggest more stable performance across image sets.

Table 1. Performance of objective IQA models

IQA model		SRCC		KRCC	
		μ	σ	μ	σ
PSNR	FR	0.872	0.089	0.759	0.132
VSNR[16]	FR	0.817	0.167	0.723	0.179
WSNR[17]	FR	0.888	0.076	0.797	0.113
IW-PSNR[18]	FR	0.745	0.210	0.644	0.216
NQM[19]	FR	0.806	0.138	0.690	0.167
IFC[20]	FR	0.805	0.167	0.673	0.174
VIF[21]	FR	0.579	0.306	0.467	0.261
VIFP[21]	FR	0.857	0.107	0.737	0.146
SSIM[1]	FR	0.847	0.113	0.721	0.150
MS-SSIM[22]	FR	0.853	0.108	0.728	0.141
IW-SSIM[18]	FR	0.838	0.155	0.721	0.160
RFSIM[23]	FR	0.862	0.125	0.747	0.151
RRED[24]	RR	0.651	0.211	0.537	0.198
RRIQA[25]	RR	0.029	0.323	0.000	0.258
AniNRIQA[26]	NR	0.390	0.362	0.329	0.314
BRISQUE[27]	NR	0.188	0.430	0.134	0.366
BIQI[28]	NR	0.083	0.351	0.033	0.295
NIQE[29]	NR	0.513	0.338	0.410	0.295

It can be observed that state-of-the-art RR and NR approaches do not provide adequate predictions of denoised image quality. Several FR models (PSNR, WSNR, VIFP, SSIM, MS-SSIM, IW-SSIM, RFSIM) are moderately correlated with subjective scores, but somewhat surprisingly, these models include PSNR, which was widely criticized for its poor prediction of perceptual image quality in general [15].

The above test results imply that there is significant space for future improvement of objective models. One useful observation we have is that state-of-the-art structural fidelity measures such as SSIM and MS-SSIM can very well predict the loss or distortion of local structural details, but fail to capture the degradation in naturalness in the denoised images. This suggests that more accurate objective IQA models can be built by combining structural fidelity and naturalness measures. Our preliminary results show that by combining SSIM with naturalness models built on statistical image models, significantly improved performance can be achieved. These results will be refined and published in our future publications.

4. CONCLUSION

We make one of the first attempts dedicated to quality assessment of denoised images. A database of denoised images was created, followed by subjective experiment and data analysis. Our results are somewhat surprising, suggesting that well-known objective IQA methods are quite limited in predicting the quality of denoised images. This motivates us to develop advanced objective models, which, based on our cur-

rent understanding, should incorporate both structural fidelity and naturalness assessment. Optimal design of future image denoising algorithms based on novel quality assessment measures is another promising topic worth further investigation.

5. REFERENCES

[1] 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. 600–612, Apr. 2004.

[2] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE image quality assessment database release 2," `http://live.ece.utexas.edu/research/quality/`.

[3] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008 - a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, pp. 30–45, 2009.

[4] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, Jan. 2010.

[5] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *IEEE Asilomar Conference on Signals, Systems and Computers*, Nov. 2012, pp. 1693–1697.

[6] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo, "Color image database TID 2013: peculiarities and preliminary results," in *European Workshop on Visual Information Processing*, Jun. 2013.

[7] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, pp. 1338–1351, 2003.

[8] F. Luisier, T. Blu, and M. Unser, "SURE-LET for orthonormal wavelet-domain video denoising," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 20, no. 6, pp. 913–919, 2010.

[9] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, pp. 2080–2095, 2007.

[10] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 11, pp. 4311–4322, 2006.

[11] A. Foi, V. Katkovnik, and K. Egiazarian, "Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images," *IEEE Trans. Image Process.*, vol. 16, no. 5, pp. 1395–1411, May 2007.

[12] W. S. Dong, X. Li, L. Zhang, and G. M. Shi, "Sparsity-based image denoising vis dictionary learning and structural clustering," in *IEEE Conf. Computer Vision and Pattern Rec. (CVPR)*, 2011.

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

[14] H. A. David, *The Method of Paired Comparisons*, Oxford University Press, New York, 1988.

[15] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," vol. 26, no. 1, pp. 98–117, Jan. 2009.

[16] D.M. Chandler and S.S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, pp. 2284–2298, 2007.

[17] T. Mitsa and K. L. Varkur, "Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms," in *IEEE Inter. Conf. Acoustic, Speech, and Signal Process. (ICASSP), 1993*, april 1993, vol. 5, pp. 301 –304 vol.5.

[18] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185 – 1198, May 2011.

[19] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, Apr. 2000.

[20] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[21] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[22] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems and Computers*, Nov. 2003, vol. 2, pp. 1398–1402.

[23] Z. Lin, Z. Lei, and M. Xuanqin, "RFSIM: A feature based image quality assessment metric using Riesz transforms," in *IEEE Inter. Conf. Image Process. (ICIP), 2010*, Sept. 2010, pp. 321–324.

[24] R. Soundararajan and A. C. Bovik, "Rred indices: Reduced reference entropic differencing for image quality assessment," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 517–526, 2012.

[25] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Human Vision and Electronic Imaging X, Proc. SPIE*, San Jose, CA, January 2005, vol. 5666.

[26] S. Gabarda and G. Cristóbal, "Blind image quality assessment through anisotropy," *J. Opt. Soc. Am. A*, vol. 24, no. 12, pp. B42–B51, Dec. 2007.

[27] 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, Dec. 2012.

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

[29] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Process. Letters*, 2012.