DENOISING METHOD SELECTION BY COMPARISON-BASED IMAGE QUALITY ASSESSMENT

Haoyi Liang and Daniel S. Weller
University of Virginia, Department of ECE, Charlottesville, VA, 22904, USA

***Abstract*—Based on a diverse range of priors on natural scene images and noise, numerous denoising algorithms have been proposed in the literature. The image quality resulting from different denoising algorithms may vary significantly across a data set. In this work, we propose a denoising algorithm selection framework that chooses among different denoising algorithms using comparison-based image quality assessment. Extensive experiments on two databases show that the proposed comparison-based selection framework consistently selects the SSIM-optimal denoising algorithm without a reference image. The proposed selection method effectively removes the burden of selecting a denoising method for applications involving processing large data sets automatically.**

Index Terms—denoising, image quality, method selection

I. Introduction

Image denoising endures as an active image processing topic. Various denoising algorithms are proposed on different priors, such as the sparsity of natural scene images [1]–[3], the gradient distribution prior [4]–[6], learned patch priors from clean images [1], [7], [8] and nonlocal (NL) self-similarity [7], [9]–[11]. Despite steady research on image denoising, an universally optimal method remains an open problem. The reason for this is that denoising algorithms are proposed from different viewpoints. On certain distorted images, some priors are more suitable than others. This suitability is mainly decided by noise characteristics and image content. To select the best denoising algorithm for a given noisy image, one considers all available priors and chooses the most suitable one. For instance, the weighted encoding method [3] implicitly incorporates impulse pixel detection into the algorithm and thus is preferable when removing impulsive noise, such as salt-and-pepper noise. The Improved NL-Means filter in [10] effectively removes correlated noise by replacing the Euclidean distance with Mahalanobis distance with the noise covariance to calculate the patch similarity. For numerous algorithms based on NL self-similarity [9], [10], the performance is better at lower noise levels, when the patch similarity these algorithms are based on holds well. An external prior guided approach [7] aims to handle higher noise levels through combining NL self-similarity and priors learned from clean images. Characteristics of the image contents, like the texture complexity, also influence denoising algorithm performance. By enforcing the gradient distribution of the denoised image to be close to the estimated gradient distribution of the original image, [4] is particularly good at removing noise from images with fine textures, like carpets.

The various advantages of denoising algorithms inspire our work: by comparing the outputs of different denoising algorithms, and selecting the best one, we can outperform any single denoising algorithm. For applications where a large set of heterogeneous noisy images to be processed, it is preferable to optimize the denoising algorithm for each image individually. However, we cannot expect a user to manually select a denoising algorithm for each image. To automatically determine which denoising algorithm yields the perceptually most satisfying result, we turn to automatic image quality assessment (IQA).

According to the availability of a reference image, IQA algorithms are classified into three categories: full reference (FR) IQA, restrict reference (RR) IQA and no reference (NR) IQA. FR-IQA algorithms, like peak signal to noise ration (PSNR) and structural similarity (SSIM) index [12], are widely used to evaluate the performance of image enhancement algorithms. RR-IQA algorithms [13]–[15] do not assume the reference image is available, but rely on some features extracted from that image, like the statistical model of image wavelet coefficients. Finally, NR-IQA only uses the distorted image. For denoising method selection, only NR-IQA algorithms are practical in the absence of a noise-free ground truth. However, the information we have for denoising method selection is more than what traditional NR-IQA algorithms require. For traditional NR-IQA algorithms, only the image to be evaluated is needed; while for denoising method selection, we have multiple denoised versions of a single image. By comparing two images, our previous work on comparison-based IQA [16] is particularly suitable for this denoised image ranking problem.

The rest of the paper is organized as follows. Section II first discusses related NR-IQA methods, and then introduces comparison-based IQA and how it is used for denoising method selection. Experiments in Section III show different advantages of denoising algorithms, and verify that with denoising method selection, denoised image quality is significantly improved compared against single denoising algorithms.

II. Comparison-based Image Quality Assessment

Evaluating the perceptual quality of a given image is difficult especially when the reference image is not available. Some popular approaches, such as DIIVINE [17] and BRISQUE [18], are based on natural scene statistics (NSS). Because the distributions of NSS share certain common characteristics among distortion free images, DIIVINE and BRISQUE evaluate distortions by measuring the change of NSS distributions. In Anisotropy [19] and LPSI [20], image quality is measured by how much information diversity an image has. The MetricQ [21] method proposes the concept of true image content. A local patch contains true image content if the gradients is organized in a structured way. The more true image content an image has, the better the image is.

All these traditional NR-IQA algorithms take one image as the input and output a quality index. However, such an absolute quality index is difficult to measure even for human beings. In subjective image quality evaluation experiments, volunteers are asked to score an image by comparing it with other images [22]. Comparison facilitates image quality assessment when multiple distorted images are available. In [16], we propose a novel comparison-based image quality assessment (C-IQA) algorithm and show its capacity for reconstruction parameter selection by comparing the reconstructed images with different parameters.

C-IQA takes two images as inputs, and outputs a scalar number indicating the relative quality of the first image based on the second. The framework of C-IQA contains three steps: Distortion Detection, Contribution and Texture Compensation. Assume I_1 and I_2 are two images to be compared, and P_1 and P_2 are two local patches from those images at the same position. Since the overall relative quality is the average of local scores, we introduce three components in C-IQA based on P_1 and P_2. In Distortion Detection, the differential patch, $D_p = P_1 - P_2$, is classified into one of two classes: structured difference or random difference. The classification basically measures gradient directions in D_p: a differential patch with highly concentrated gradient directions is classified into structured difference, otherwise the differential patch is classified into random difference. Once obtaining the type of differential patch, the second step, Contribution, quantifies the contributions from two input patches to the differential patch using covariance. Then a straightforward philosophy combines the first two modules together: the relative quality of two input patches is determined by the type of the differential patch and contributions from input patches to the differential patch. If the differential patch shows a random pattern, the input patch that mainly contributes to the differential patch is worse; otherwise, the input patch is better if it mainly contributes to a structured differential patch. At last, Texture Compensation adjusts the weighting of different image patches because images with different texture complexities have different sensitivities to the same distortion. For example, images with flat content are more sensitive to noise than images with fine texture.

For denoising method selection, multiple denoised images need to be ranked but C-IQA is designed for comparing two images. The bubble sort algorithm is adopted to extend C-IQA to compare multiple images. Each relational operation in the traditional bubble sort algorithm is replaced by image comparison with C-IQA .

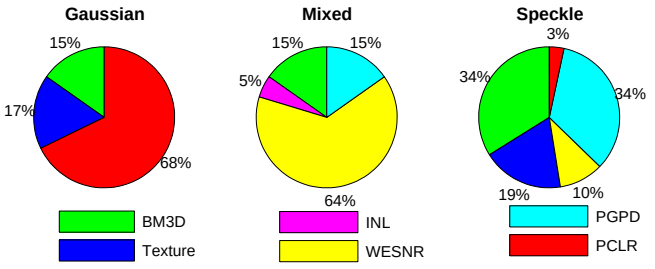


Fig. 1: The best denoising algorithms for each type of noise. Each pie chart represents 59 distorted images for one kind of noise.

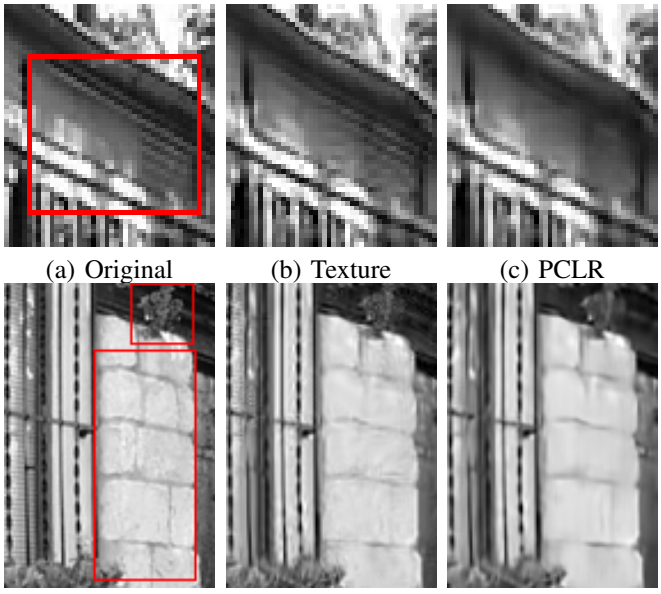
III. Experiments

The proposed denoising algorithm selection framework is tested on two public image datasets for quality assessment, LIVE [23] and CSIQ [22]. LIVE has 29 distortion-free images and CSIQ has 30 distortion-free images. Because the images in LIVE and CSIQ are all high quality natural scene images, we combine the two databases together in the following experiments. Six state-of-the-art denoising algorithms included in experiments are BM3D [9], INL [10], Texture [4], WESNR [3], PGPD [8], PCLR [7]. These algorithms are selected because BM3D and INL are widely used as the benchmark for denoising algorithms; Texture and WESNR are good at denoising images with highly textured content and contained by impulsive noise respectively; PGPD and PCLR are two of the latest algorithms showing promising results. On each original image, three distorted images are created by different kinds of noise: Gaussian noise, Gaussian noise mixed with salt-and-pepper noise (referred as Mixed noise in the follows) and speckle noise. The original images are first turned into gray images scaled between 0 and 255. The standard deviation of Gaussian noise is 25; Mixed noise is generated by adding salt-and-pepper noise to Gaussian noise with the standard deviation 25 and the density of salt-and-pepper noise is 2.5%; the standard deviation of the multiply factor is 0.01 for speckle noise. Therefore, we have 177 noisy images generated by 59 original images and 1062 denoised images in total in our experiments.

In the next three parts, we first verify the claim that the performance of denoising algorithms varies with noise type and image content in Section III-A. Section III-B and III-C illustrate that the proposed denoising algorithm selection outperforms other single denoising algorithms and C-IQA is particularly suitable for image quality assessment when multiple images are available.

(a) "building2" (b) "sailing2"

Fig. 2: Original image samples. Red patches in (a) and their denoised versions are shown in Fig. 3. The red patch in (b) and its denoised versions are shown in Fig. 5.



(a) Original (b) Texture (c) PCLR

(d) Original (e) Texture (f) PCLR

Fig. 3: Detailed Patches from "building2". (a) and (d) are patches from the original image; (b) and (e) are denoised results by Texture [4] after adding Gaussian noise; (c) and (f) are denoised results by PCLR [7] after adding Gaussian noise. The SSIM index for the denoised image by Texture is 0.7834 and 0.7804 by PCLR.

A. Performance Variation of Denoising Algorithms

As mentioned in Section I, some denoising algorithms are suitable for certain kinds of noise but do not always perform the best. This experiment illustrates the necessity of denoising method selection: the best denoising algorithms vary with the noise type and the image content. In Fig. 1, the distribution of the best denoising algorithms according to SSIM for each kind of noise are shown. It is clear that PCLR and WESNR are the best denoising algorithms for Gaussian noise and Mixed noise respectively, and BM3D and PGPD are two preferable denoising algorithms for speckle noise. But due to the variation

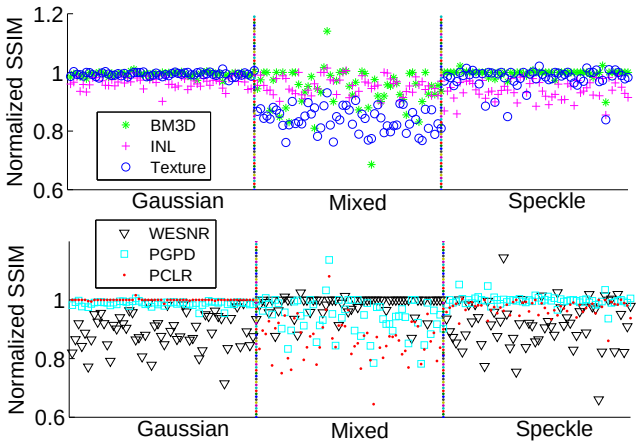


Fig. 4: Normalized SSIM indexes of all the denoised images. The original SSIM indexes of six denoised images from the same noisy image are normalized by the one selected by C-IQA. The constant horizontal line of one represents the performance of the selected algorithm using C-IQA.

of image contents, there are exceptions for each kind of noise.

One case of Gaussian noise removal is shown in Fig. 3 and the whole original image is shown in Fig. 2 (a). The SSIM index of image denoised by Texture is 0.7834 and 0.7804 by PCLR. These quantitative scores are supported by the details in Fig. 3. The grids on the roof and the texture on the marble are well preserved by Texture but blurred by PCLR. The reason is that "building2" is an image with remarkably fine textures and Texture is particularly designed to preserve these.

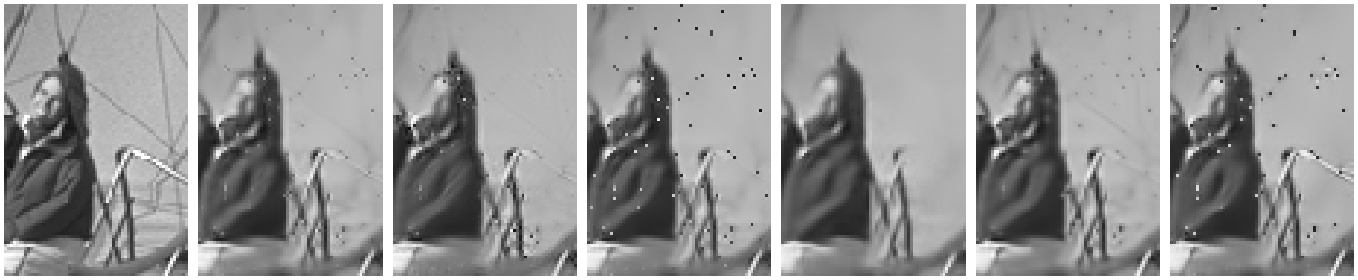
B. Denoising Algorithm Selection with C-IQA

Fig. 4 shows normalized SSIM indexes of all the 1062 denoised images. Markers at the same vertical position stand for six denoised images by different denoising algorithms. For each noisy image, SSIM indexes of six denoised images are normalized by the one selected by C-IQA. Therefore, the constant horizontal line equal to one is the relative score of the denoising algorithm selected by C-IQA. It is clear that the quality of images denoised using the algorithms selected by C-IQA behaves like an envelope for the quality attained by the other single denoising algorithms. The average SSIM indexes of different single denoising algorithms and ones selected by different IQAs are listed in Table I. Because PCLR is the optimal or suboptimal denoising algorithm for Gaussian noise in most cases, the average performance of selection by C-IQA is the same as the best single denoising algorithm, PCLR. For the other two kinds of noise, the average performance of selection by C-IQA is better than the best single denoising algorithms. Compared to selection results by other IQAs, C-IQA demonstrates its advantage when evaluating multiply images with the same content.

One example of denoising algorithm selection for "sailing2" distorted with Mixed noise is shown in Fig. 5. The quality difference among six denoised images is substantial and denoising algorithm selection ensures the best result is used. Fig. 6 further shows that accumulated relative quality

TABLE I: Average SSIM indexes of different denoising algorithms (or selected) for different noise

	Existing Denoising Algorithms						Selected by IQAs					
	BM3D	INL	Texture	WESNR	PGPD	PCLR	C-IQA	Metric Q	DII	BRI	ANI	LPSI
Gaussian	0.8534	0.8302	0.8526	0.7703	0.8482	**0.8582**	**0.8582**	0.8467	0.8417	0.8309	0.8549	0.8253
Mixed	0.6973	0.7076	0.6202	**0.7295**	0.6955	0.6365	**0.7408**	0.7139	0.6250	0.7086	0.6868	0.7054
Speckle	0.8182	0.7716	0.8045	0.7454	**0.8192**	0.7989	**0.8212**	0.8035	0.7770	0.7883	0.8161	0.7698
Overall	**0.7896**	0.7698	0.7591	0.7515	0.7876	0.7645	**0.8068**	0.7880	0.7479	0.7759	0.7859	0.7669



(a) Original SSIM: 1 (b) BM3D SSIM: 0.7900 (c) INL SSIM: 0.7738 (d) Texture SSIM: 0.6576 (e) WESNR SSIM: 0.8043 (f) PGPD SSIM: 0.7588 (g) PCLR SSIM: 0.7072

Fig. 5: Local patches from denoised images by different denoising algorithms. WESNR shows the best denoising result because it considers impulsive noise removal.

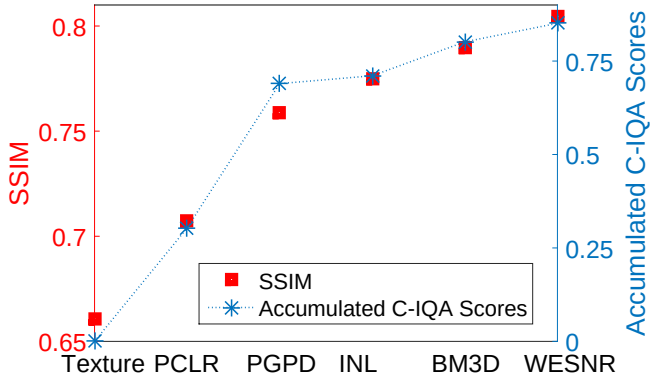


Fig. 6: The correlation between SSIM indexes of six denoised images and accumulated C-IQA scores is 0.92.

scores between neighboring rank images by C-IQA are highly correlated with SSIM indexes.

C. Discussions on Other IQAs

In the previous experiments, selected denoising algorithms by C-IQA show better performance than six state-of-the-art denoising algorithms. However, this does not mean improving denoising performance by selecting an existing algorithm is a trivial task. Only image quality assessment algorithms that fully make use of the available information can improve the denoising performance further. In Table I, the selected denoising algorithms by other traditional single-input NR-IQA metrics do not yield an improvement on the original six denoising algorithms.

To further reveal the ability of C-IQA evaluating images with the same content, we use the weighted inversion number to measure the ranking difference between NR-IQA algorithms ans SSIM. Assume $I_1, \cdots, I_6$ are the denoised images by six denoising algorithms and $(I_{s(1)}, \cdots, I_{s(6)})$ is the ranking sequence according to a NR-IQA metric from low quality to high quality. The weighted inversion number is defined as

$$Inv = \sum_{i=1:N} \sum_{j=i+1:N} max(0, SSIM(I_{s(i)}) - SSIM(I_{s(j)})).$$

TABLE II: The average weighted inversion numbers

	C-IQA	Metric Q	DII	BRI	Ani	LPSI
Gaussian	0.0273	0.1099	0.1568	0.1813	**0.0212**	0.4134
Mixed	**0.0260**	0.2929	0.8395	0.1449	0.4373	0.4225
Speckle	**0.0500**	0.1565	0.2984	0.2698	0.0825	0.3747
Overall	**0.0344**	0.1864	0.4315	0.1986	0.1803	0.4035

Table II shows the average weighted inversion numbers between different NR-IQA metrics and SSIM. C-IQA is much better than the other NR-IQA metrics and is only slightly outperformed by Anisotropy when the noise type is Gaussian.

IV. Conclusion and Future Work

In this paper, we first analyzed the advantages of different denoising algorithms and showed the performance of an image denoising algorithm depends on the noise and image content characteristics. Extensive experiments verify our analysis about denoising algorithms and show that with denoising method selection, different denoising algorithms together obtain more stable and better denoising performance.

Based on the comparison-based image quality assessment, we plan to extend the current C-IQA to a multiple-comparison version. Currently, bubble sort is adopted to rank multiple images but the relational operation only uses information of two images each time. In the multiple-comparison version, we hope to design an IQA algorithm that takes multiple images into account at the same time and facilitates parameter or method selection for other image processing problems.

REFERENCES

[1] M. Elad and M. Aharon, "Image denoising via sparse and redundant representation over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, 2006.

[2] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proc. ICCV*, 2009, pp. 2272–2279.

[3] J. Liang, L. Zhang, and J. Yang, "Mixed noise removal by weighted encoding with sparse nonlocal regularization," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2651–2662, Apr. 2014.

[4] W. Zuo, L. Zhang, C. Song, and D. Zhang, "Texture enhanced image denoising via gradient histogram preservation," in *Proc. CVPR*, 2013, pp. 1203–1210.

[5] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-laplacian priors," in *Proc. NIPS*, 2009, pp. 1033–1041.

[6] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithm," *Physical D: Nonlinear Phenomena*, pp. 259–268, 1992.

[7] F. Chen, L. Zhang, and H. Yu, "External patch prior guided internal clustering for image denoising," in *Proc. ICCV*, 2015, pp. 603–611.

[8] J. Xu, L. Zhang, W. Zuo, D. Zhang, and X. Feng, "Patch group based nonlocal self-similarity prior learning for image denoising," in *Proc. ICCV*, 2015, pp. 244 – 252.

[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, no. 8, pp. 2080–2096, Aug. 2007.

[10] B. Goossens, H. Luong, A. Pizurica, and W. Philips, "An improved non-local denoising algorithm," in *Proc. Int. Workshop on Local and Non-Local Approximation in Image Processing (LNLA)*, 2008, pp. 143–157.

[11] T. Dai, C. Song, J. Zhang, and S. Xia, "PMPA: A patch-based multiscale products algorithm for image denoising," in *Proc. ICIP*, 2015, pp. 4406–4410.

[12] 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.

[13] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE J. Sel. Topics Signal Processing*, vol. 3, pp. 202–211, Apr. 2009.

[14] 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, Feb. 2011.

[15] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E. Yang, and A. C. Bovik, "Quality-aware images," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1680–1689, Jun. 2006.

[16] H. Liang and D. S. Weller, "Comparison-based image quality assessment for parameter selection," *arXiv*, , no. 1601.04619, 2016.

[17] A. M. Krishna 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, Dec. 2011.

[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, Dec. 2012.

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

[20] Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," in *Proc. ICIP*, 2015, pp. 339–343.

[21] X. Zhu and P. Milanfar, "Automatic parameter selection for denoising algorithms using a no-reference measure of image content," *IEEE Trans. Image Processing*, vol. 19, no. 12, pp. 3116–3132, Dec. 2010.

[22] 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, pp. 011006–1–011006–20, Mar. 2010.

[23] H. R. Sheikh, A. C. Bovik, L. Cormack, and Z. Wang, "LIVE image quality assessment database release 2," 2005, http://live.ece.utexas.edu/research/quality.