IEEE International Conference on Image Processing, Quebec City, Canada, Sep. 2015.

PERCEPTUAL QUALITY ASSESSMENT OF DENOISED IMAGES

Kai Zeng and Zhou Wang

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

ABSTRACT

Image denoising has been an extensively investigated problem in the field of image processing, but little research has been dedicated to the development and validation of image quality assessment (IQA) approaches for denoised images. Without such IQA methods, fair comparison is difficult and further improvement is aimless. In this study, we first create a denoised image database and conduct a subjective experiment to compare the quality of these images. We find widely used IQA measures only have moderate correlations with subjective opinions. Furthermore, we propose a novel objective IQA approach that combines the full-reference SSIM approach with natural scene statistics (NSS) based reduced-reference IQA methods. Experimental results show that the proposed scheme outperforms state-of-the-art IQA models.

Index Terms— Image Quality Assessment, Denoised Image, Structural Similarity, Naturalness

1. INTRODUCTION

Image denoising has been an extensively investigated problem in the field of image processing. It not only creates visually appealing pictures, but also helps facilitate other image processing operations, such as compression, recognition, and resizing. A large number of denoising algorithms have been proposed in the past decades, but little work has been dedicated to quality evaluation of denoised images. In practice, researchers often use common IQA measures such as peak signal-to-noise-ratio (PSNR) and the structural similarity index (SSIM) [1] to compare images and denoising algorithms, but proper validations of these measures are missing.

Both subjective and objective IQA methods can be employed to assess the quality of denoised images. In a subjective experiment, multiple human subjects are asked to rate or rank the quality of denoised images for mean opinion score (MOS) collection. Subjective methods are highly valuable in comparing image denoising algorithms and in validating objective IQA methods, but they are often expensive and slow. Depending on the accessibility to the original reference image that is assumed to have perfect quality, objective IQA measures may be classified into full-reference (FR), reduced-reference (RR) and no-reference (NR) methods. Objective models can be employed to evaluate image quality automatically, and can also be embedded into the design and optimization of various image processing algorithms and systems. Notable success has been achieved in all three categories, especially in the FR case, where a number of state-of-the-art algorithms, including the SSIM family [1, 2, 3], the visual information fidelity (VIF) method [4], the visual signal-to-noise (VSNR) approach [5], the feature similarity (FSIM) algorithm [6], have been shown to have good correlations with subjective quality ratings when tested using many large-scale image databases that include a variety of distortion types and levels [3, 6].

In this work, we focus on perceptual quality assessment of denoised images. We first create a database that contains denoised images and carry out a subjective test using the database. We find that state-of-the-art IQA models only moderately correlate with subjective opinions. Closer observation reveals that popular deterministic IQA approaches such as PSNR and SSIM lack appropriate considerations on the statistical naturalness of images. This motivates us to incorporate the philosophy behind natural scene statistics (NSS) based models [3] into the framework. Therefore, we propose a novel objective IQA approach that combines FR multi-scale SSIM with RR distortion measures based on NSS features. Experimental validations show that the proposed approach outperforms state-of-the-art IQA models in predicting subjective rankings of denoised images.

2. SUBJECTIVE QUALITY ASSESSMENT

To the best of our knowledge, currently the only publicly available database that contains an image denoising dataset is TID2013 [7]. Unfortunately, the dataset includes images denoised using BM3D [8] only, and the number of samples is too limited to fully validate an IQA model. Therefore, our first goal is to develop a dedicated database for IQA of denoised images.

Ten original high-quality natural images of size 512×512 are chosen to cover diverse natural image content. Independent white Gaussian noise of three levels is added to each image with standard deviations σ_n equaling 15, 30, and 50, respectively. Eight algorithms are selected to denoise the images. These include simple noise-removal operators such as linear Gaussian filter and locally adaptive Wiener filter (MATLAB Wiener2D function), as well as state-of-the-art denoising algorithms, such as BLS-GSM [9], SURE-LET [10], BM3D [8], K-SVD [11], SADCT [12], and CSR [13]. These methods are chosen to cover a diverse types of denoisers in terms of both methodology and performance. Default parameter settings are adopted for all denoising algorithms without any tuning for better quality. With all images and denoising algorithms combined, 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. A group of sample noisy images, together with their corresponding denoised images are shown in Fig. 1.

In the subjective experiment, all 8 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 test method conforms with ITU-T BT.500 [14]. For each image set, the subject is asked to rank the perceptual quality of the 8 images from "the best" to "the worst". A total of 20 naïve observers participated in the subjective experiment. 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 "ground truth", we can observe the performance of each individual subject by comparing their rank-order with the "ground truth" for image

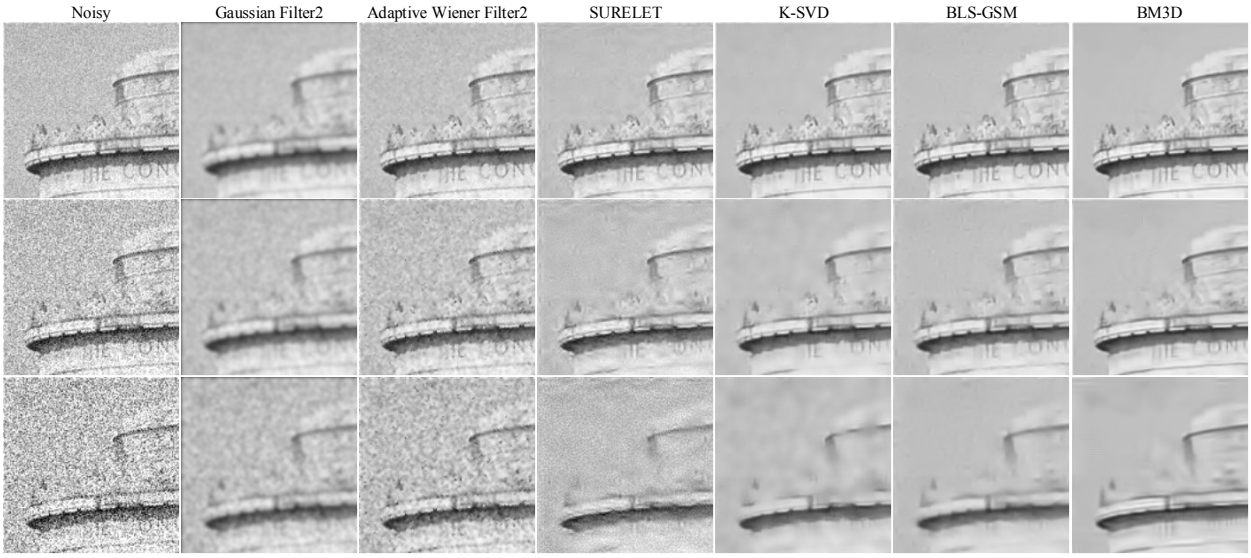

Fig. 1. Sample noisy and denoised images (enlarged and cropped for visibility). Column 1: noisy images with noise standard deviation σ_n equaling 15, 30 and 50, respectively. Columns 2-7: denoised images by 6 algorithms.

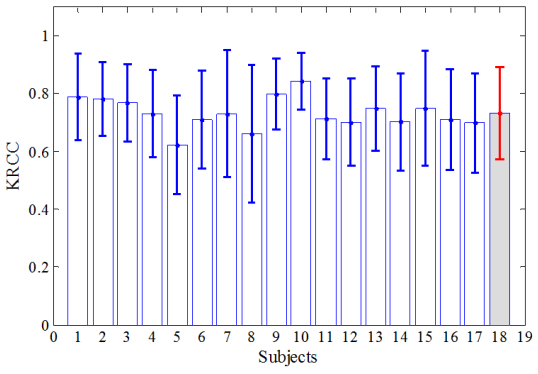

Fig. 2. Mean and error bars (±std) of KRCC values between individual subject and average subject rankings. The rightmost column is the average performance across all subjects.

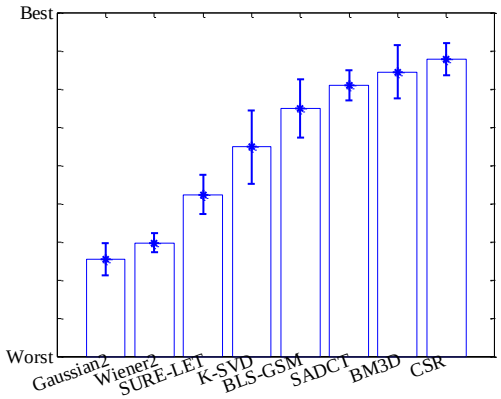

Fig. 3. Mean and error bars (±std) of subjective rankings of individual denoiser across all image sets.

set, and then average the performance over all 30 image sets. The comparison is based on Kendall's rank-order correlation coefficient (KRCC). The mean and standard deviation of KRCC values for each individual subject are depicted in Fig. 2. It can be seen that there is a considerable agreement between subjects on ranking the quality of denoised images. The average performance across all individual subjects is also given in the rightmost column in Fig. 2. This provides a general idea about the performance of an average subject.

Furthermore, we use the subjective rankings to compare the 8 denoising algorithms by computing their average and standard deviation of rankings across all image sets. The results are summarized in Fig. 3. It can be observed that state-of-the-art denoisers such as BM3D [8] and CSR [13] perform significantly better than more traditional methods. On the other hand, from the sizes of the error bars, we observe substantial variations between subject preferences of the denoisers.

3. OBJECTIVE QUALITY ASSESSMENT

Object IQA measures are highly desirable in the comparison, parameter tuning and optimal design of desnoising algorithms. Unfortunately, existing objective IQA models do not give convincing performance in our denoised image database. Details of the test results will be given in Section 4. One useful observation is that certain FR measures such as SSIM provide accurate local predictions on how image structural details are distorted, but the subjects' overall impression is often altered by whether the images look natural. This leads us to develop a new IQA measure that combines both local structural fidelity and global statistical naturalness measures.

For local structural fidelity measure, both the reference and distorted images are first transformed into a multi-scale and multi-orientation wavelet domain (in particular, the steerable pyramid [9] is employed due to its translational and rotational invariance properties). The structural distortion measure basically follows the SSIM

approach [1] but is applied in wavelet subbands. Let $\mathbf{x}$ and $\mathbf{y}$ be two sets of wavelet coefficients collected from corresponding patches from the reference and distorted subbands, respectively. The local SSIM between the patches is computed as

$$S_{local}(\mathbf{x},\mathbf{y}) = \frac{2\sigma_{xy}+C_2}{\sigma_x^2+\sigma_y^2+C_2}\,, \tag{1}$$

where σ_x^2 and σ_{xy} represent the variance and covariance of the coefficient blocks, respectively, and C_2 is a small positive constants to avoid instability when the means and variances are close to zero. Note that the luminance comparison term in the original spatial domain SSIM definition [1] is not included because the coefficients are zero-mean due to the bandpass nature of the wavelet filters. Applying the local SSIM measure across space generates a subband SSIM map, and the SSIM maps of all subbands are combined to an overall structural distortion measure given by

$$D_S = 1 - \sum_{i=1}^{M} w_i \left[\frac{1}{N_i}\sum_{j=1}^{N_i} S_{local}(\mathbf{x}_{i,j},\mathbf{y}_{i,j})\right]\,, \tag{2}$$

where $\mathbf{x}_{i,j}$ and $\mathbf{y}_{i,j}$ are the j-th coefficient patches in the i-th subband in the original and distorted images, respectively, N_i is the number of local SSIM values in the i-th subband, M is the total number of subbands, and w_i is the weight given to the i-th subband and $\sum_{i=1}^{M} w_i = 1$.

For global statistical naturalness, we propose two NSS based statistical distortion measures. The first is based on the marginal distributions of wavelet coefficients that are found previously to be heavy-tailed [9] for natural images, as exemplified in Fig.4. It can be observed that different distorted images change the distribution in different ways. In [15], the Kullback-Leibler divergence (KLD) between the distributions of the reference and distorted images was employed for RR IQA. However, KLD does not differentiate changes in the shapes of the distributions. For example, noisy images tend to make the distribution broader and blurry images may increase the peakedness of the distribution. Here we use excess kurtosis to capture such shape changes

$$K = \frac{\frac{1}{N}\sum_i (x_i-\mu_x)^4}{\left[\frac{1}{N}\sum_i (x_i-\mu_x)^2\right]^2} - 3\,, \tag{3}$$

where x_i is the i-th wavelet coefficient and μ_x is the mean value of all wavelet coefficients within the subband, respectively. The kurtosis computation is applied to each wavelet subband in the reference and distorted images, and a distortion measure is given by

$$D_K = \sum_{i=1}^{M} w_i \max\left\{1 - \frac{K_d^i}{K_r^i}, 0\right\}\,, \tag{4}$$

where K_d^i and K_r^i are the excess kurtosis of the i-th subband of the distorted and reference images, respectively.

Another important discovery in NSS literature is that the power spectrum of natural images falls with the spatial frequency approximately proportional to $1/f^p$ [16], where f is the spatial frequency and p is a content-dependent constant. Image distortions such as noise contamination and denoising operation may change the slope of such energy falloff, as exemplified in Fig. 5, where different denoised images change the energy falloff across scale in different ways. Our second statistical distortion measure is based on quantifying the changes in the slope of energy falloff. We first compute

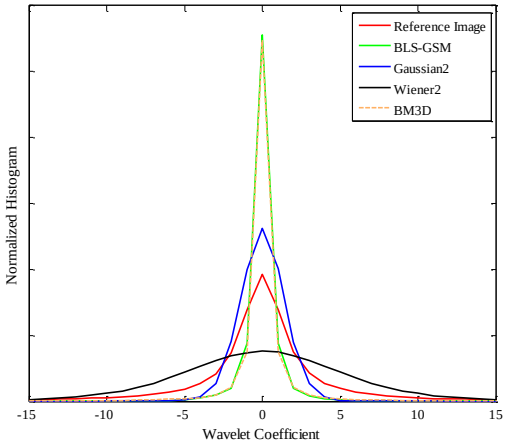

Fig. 4. The marginal wavelet coefficient distributions of reference and distorted images.

the log-energy of a wavelet subband by

$$e = \log\left(1 + \frac{\sum_i u_i x_i^2}{\sum_i u_i}\right)\,, \tag{5}$$

where the summation is over all coefficients in a subband, u_i is the weight given to the i-th coefficient, and 1 is added before the logarithm computation to avoid negative result when the subband energy is extremely low. The weight $u_i = log(1 + x_i^2/0.1)$ is determined by local log energy, so that the computation is more concentrated on high energy regions (e.g., edges) in the image.

The slope of energy falloff is evaluated between the two finest-scale wavelet subbands along the same orientation by

$$F^o = \frac{|e_L^o - e_{L-1}^o|}{C_s}\,, \tag{6}$$

where e_L^o and e_{L-1}^o are the energy of the o-th orientation at the finest and second finest scales, respectively, and C_s is a scale difference constant that has no impact on the overall energy falloff measure in Eq. (7). Only the finest two scales are employed here not only to simplify the energy falloff evaluation, but also because these are usually the scales with the strongest distortions. The overall energy falloff distortion measure is defined as

$$D_F = \frac{1}{N_o}\sum_{o=1}^{N_o} \max\left\{\frac{F_d^o}{F_r^o} - 1, 0\right\}\,, \tag{7}$$

where N_o is the number of orientations, and F_d^o and F_r^o are the slope of energy falloff evaluated at the o-th orientation for the distorted and reference images, respectively.

Finally, all three distortion components, D_S, D_K and D_F , are linearly combined to yield an overall distortion measure

$$D = w_S D_S + w_K D_K + w_F D_F\,, \tag{8}$$

where w_S, w_K and w_F are weights assigned to the three components, respectively, and $w_S + w_K + w_F = 1$. Since all three components are lower bounded by 0 which is reached when the reference and distorted images are identical, the combined distortion measure also possesses the same property.

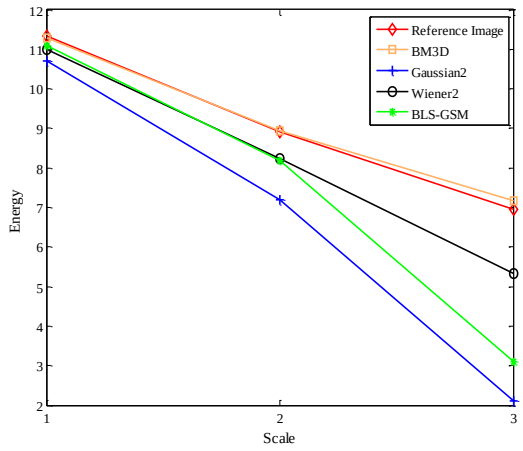

Fig. 5. The energy fall-off characteristics of reference and distorted images.

There are several parameters in the proposed algorithm. A 3-scale 4-orientation steerable pyramid is applied, thus $N_o = 4$ and $M = 12$. The weights given to each subband and to each distortion components are obtained empirically. w_i are the same for all subbands at the same scale, and different from the coarsest to the finest scales by $w_i = \{0.3, 0.6, 0.1\}$. $w_S = 0.59$, $w_K = 0.23$, and $w_F = 0.18$, respectively.

4. EXPERIMENTAL RESULTS

The proposed method is compared with 13 well-known and state-of-the-art objective IQA measures, which include 8 FR (PSNR, VSNR [5], VIF [4], VIFP [4], SSIM [1], MS-SSIM [2], IW-SSIM [3], and FSIM [6]), 2 RR (RRIQA [15] and RRED [17]), and 3 NR (BIQI [18], BRISQUE [19], and NIQE [20]) methods. In TID2013 database [7], the image denoising dataset contains 125 images, which are created by adding 5-level of independent white Gaussian noise to 25 reference images and applying BM3D [8] for denoising. KRCC is calculated between objective quality score and MOS values from database. In our developed database, for each image set, we compute the KRCC values between objective scores and average subjective rankings. The mean and standard deviation (std) of the KRCC values across all 30 image sets are used as the criteria to compare different objective IQA measures. Higher mean KRCC values indicate better correlations with subjective opinions, and lower std of KRCC values suggest better consistency or stability of the objective IQA method over different image content.

The test results based on Kendall's rank-order correlation coefficient (KRCC) are summarized in Table 1. Similar results are obtained when using Spearman rank-order correlation coefficient (SRCC) as the evaluation criterion. Somewhat surprisingly, PSNR performs quite reasonably and is slightly better than (or equivalent to) advanced FR IQA methods such as SSIM, MS-SSIM and VIF. This is in sharp contrast to the test results using other IQA databases [3, 6], where these advanced methods outperform PSNR by large margins. It can also be observed that state-of-the-art RR and NR IQA methods fail to provide useful quality predictions of denoised images. Overall, the proposed method achieves the best performance, and its improvement over SSIM and MS-SSIM demonstrates the value of including NSS-based statistical naturalness measures. Note that the performance gain of the proposed method if more pronounced on our database than TID2013 image denoising dataset. This may be because our database contains more diverse types of denoising algorithms while TID2013 only includes BM3D denoised images.

Table 1. KRCC performance comparison of objective IQA modes on our developed database and TID2013 image denoising databset

Quality/Distortion	Our database		TID2013
Model	mean	std	
PSNR	0.7587	0.1318	0.8089
VSNR [5]	0.7230	0.1789	0.7342
VIF [4]	0.4667	0.2609	0.7234
VIFP [4]	0.7372	0.1456	0.7371
SSIM [1]	0.7205	0.1497	0.7580
MS-SSIM [2]	0.7283	0.1412	0.7836
IW-SSIM [3]	0.7205	0.1598	0.7515
FSIM [6]	0.5371	0.3187	0.7895
RRIQA [15]	0.0002	0.2582	0.6655
RRED [17]	0.5369	0.1979	0.7776
BIQI [18]	0.0329	0.2953	0.2785
BRISQUE [19]	0.1336	0.3659	0.4078
NIQE [20]	0.4102	0.2954	0.4066
Proposed (D)	**0.8231**	**0.1111**	**0.8148**

Since there is no sophisticated iterative or search procedures involved in the proposed algorithm, its computational complexity remains low. On an Intel Core2 Duo E8600 computer with 4GB memory running on 64-bit OS at 3.33GHz, it takes around 1.26 second for an un-optimized MATLAB implementation of the proposed algorithm to evaluate a 512×512 grayscale image. The fast speed allows it to be easily adopted in practical applications.

5. CONCLUSION AND FUTURE WORK

The current study focuses on the quality assessment aspect of image denoising. This is an important issue in the validation and optimal design of image denoising algorithms, but has not been deeply investigated. We built one of the first databases dedicated to image denoising and carried out a subjective test to rank the quality of these images. Moreover, an objective IQA approach for denoised images is proposed that combines SSIM-based structural fidelity index with NSS-based statistical distortion measures. Experimental validation shows that the proposed algorithm outperforms state-of-the-art IQA models in terms of correlations with subjective opinions. It is worth mentioning that classical FR IQA approaches typically concentrate on predicting the visibility of local deterministic signal differences or structural distortions, but often overlook the global statistical naturalness of the distorted image. In this sense, the proposed method contributes to the general methodology of IQA by incorporating statistical naturalness measures (that are only used by RR and NR approaches before [21]) into FR IQA, and provides a demonstration of this approach. In the future, the proposed method may be improved by incorporating other statistical naturalness models. It may also be extended to the quality assessment of color image or video denoising algorithms, or to other image/video processing applications such as restoration, enhancement and compression.

6. 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, pp. 600–612, Apr. 2004.

[2] 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*, vol. 2, pp. 1398–1402, Nov. 2003.

[3] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, pp. 1185 – 1198, May 2011.

[4] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, pp. 430–444, Feburary 2006.

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

[6] Z. Lin, Z. Lei, M. Xuanqin, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, pp. 2378–2386, Aug. 2011.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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, pp. 1395–1411, May 2007.

[13] 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.

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

[15] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik, "Quality-aware images," *IEEE Trans. Image Process.*, vol. 15, pp. 1680–1689, June 2006.

[16] D. J. Field and N. Brady, "Visual sensitivity, blur and the sources of variablity in the amplitude spectra of natural scenes," *Vision Research*, vol. 37, no. 23, pp. 3367–3383, 1997.

[17] 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.

[18] 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.

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

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

[21] Z. Wang and A. C. Bovik, "Reduced- and no-reference visual quality assessment - the natural scene statistics model approach," *IEEE Signal Processing Magaine*, vol. 28, pp. 29–40, Nov. 2011.