2015 IEEE International Conference on Image Processing (ICIP'15), Quebec, Canada, Sep. 2015

PERCEPTUAL EVALUATION OF SINGLE IMAGE DEHAZING ALGORITHMS

Kede Ma, Wentao Liu and Zhou Wang

Dept. of Electrical & Computer Engineering, University of Waterloo, Waterloo, ON, Canada
Email: {k29ma,w238liu,zhou.wang}@uwaterloo.ca

ABSTRACT

Images captured in outdoor scenes often suffer from poor visibility and color shift due to the presence of haze. Although many algorithms have been proposed to remove the haze, not much effort has been made on quality assessment of dehazed images. In this paper, we first build a database that contains 25 hazy images as well as dehazed images created by eight dehazing algorithms. A subjective user study is then carried out based on the database, from which we have several useful findings. First, considerable agreement between human subjects on the perceived quality of hazy and dehazed images is observed. Second, not a single dehazing algorithm performs the best for all test images. Third, existing objective image quality assessment (IQA) models are very limited in providing proper quality predictions of dehazed images.

Index Terms— Image dehazing, subjective image quality assessment, image enhancement, perceptual image processing

1. INTRODUCTION

Images captured in outdoor scenes often suffer from poor visibility and color shift due to the presence of haze[1]. Dehazing, also referred to as haze removal, is highly desirable for both computational photography and computer vision tasks. The resulting dehazed image is expected to be more perceptually appealing in general. Moreover, many computer vision algorithms can only work well with the scene radiance that is haze-free.

In 1924, Koschmieder [1] proposed a hazy image formation model that is widely adopted later on

$$\mathbf{I}(\mathbf{x}) = \mathbf{J}(\mathbf{x})t(\mathbf{x}) + \mathbf{A}(1 - t(\mathbf{x})), \quad (1)$$

where $\mathbf{I}$ is the observed hazy image, $\mathbf{J}$ is the underlying scene radiance, $\mathbf{A}$ is the atmospheric light, $t \in [0, 1]$ is the media transmission coefficient and $\mathbf{x}$ denotes pixel coordinates. The light reflected by the object to the camera is first attenuated through transmission (first term) and then blent with the atmospheric light (second term). Assuming validity of this model, the goal of haze removal is to recover $\mathbf{J}$ with no additional information about $\mathbf{A}$ and t. Apparently, this is an ill-posed problem with 3 constraints but 7 unknowns for an RGB color image. Previously, many algorithms have been proposed with the aid of additional information such as multiple images taken under different weather conditions [2, 3] or different degrees of polarization [4, 5], and depth information from either the user inputs [6] or given 3D models [7].

Only recently has single image dehazing become an active research topic. Most single image dehazing algorithms adopt the aforementioned physical model and differ mainly in the methodologies to estimate the transmission t. Fattal [8] estimated t by assuming that the surface albedo is locally uncorrelated with the transmission. Tan [9] recovered the scene radiance by maximizing local contrast with a smooth constraint on t. Inspired by the dark object subtraction technique [10], a dark channel prior is proposed to estimate t in [11], which is highly effective in detecting the thickness of the haze. When recovering the scene radiance, it needs to be combined with soft matting [12] or a guided filter [13], followed by an exposure increase procedure to obtain reasonable results. Tarel et al. [14] proposed a fast haze removal algorithm based on median filter. Using guided joint bilateral filter [15, 16], Xiao and Gan [17] refined t in [14] to better represent depth edge information. With real time filter implementation [18], the complexity of the algorithm grows linearly with the number of pixels. Kim et al. [19] estimated t by maximizing block-wise contrast and meanwhile minimizing the information loss due to pixel truncation. The transmission is then refined using a guided filter [13] as well. By adding a temporal coherence measure, they extended the algorithm to account for video dehazing. Meng et al. [20] further proposed a boundary constraint on t. Combined with a weighted L_1 norm, the induced optimization problem has a closed form solution in each iteration using a clever trick of variable splitting. Fattal [21] estimated a coarse t using a local color-lines model [22] and refined it in a Gaussian Markov random field. Tang et al. [23] investigated several haze-revelent features using Random Forest [24] for image dehazing. Assuming the independence of image content and scene depth, they synthesized a hazy image patch with a random transmission t; the input and output of Random Forest are extracted features and t of a given patch, respectively. Noticing that all images captured in natural scenes contain some noise due to sensor error, Matlin and Milanfar [25] adopted an iterative kernel regression method [26] to denoise and dehaze a hazy image simultaneously. Another approach that handles denoising and dehazing problems jointly using a variational approach is proposed in [27]. A multiresolution fusion scheme is adopted in [28], which does not depend on the physical model in [1].

With multiple dehazing algorithms available, it becomes pivotal to compare their dehazing performance so as to find further directions for advancement. Surprisingly, not much work has been done in this aspect. To the best of our knowledge, the only subjective test reported in the literature was done by Chen et al. [29] on a database of limited dehazing algorithms proposed before 2010. A comprehensive study that compares a variety of both classical and state-of-the-art dehazing algorithms has not been reported in the literature. On the other hand, objective quality assessment of dehazing algorithms is a challenging problem since a perfect quality dehazed image is not available as a reference. General purpose no-reference IQA models [30, 31, 32, 33, 34] and no-reference models for contrast-distorted images [35] may be applicable but are never tested on dehazed images. Only a few models have been designed specifically for dehazed

[1] In this paper, we do not differentiate the phenomena similar to haze (e.g., dust, mist and fume) and use "haze" as a common term for simplicity.

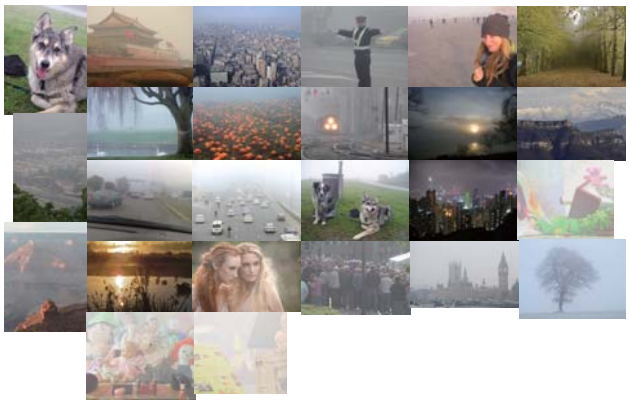


Fig. 1. Hazy images in the database.

images. In [36], the authors proposed three measures based on recovered visible edges and saturated pixels to account for the quality of a dehazed image. More recently, Chen et al. [29] exploited rank SVM to learn a quality predictor using GIST [37] and color motion features [38].

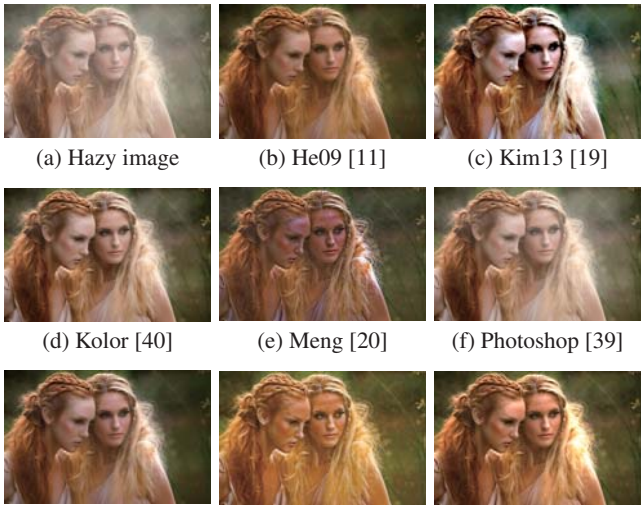
In this paper, we first build a database that contains a variety of hazy images and their corresponding dehazed images created by eight dehazing algorithms. We then conduct a subjective user study from which we have several useful observations. First, considerable agreement between human subjects on the perceived quality of hazy and dehazed images is observed. Second, not a single dehazing algorithm performs the best for all test images. Third, existing objective IQA models are very limited in predicting the quality of dehazed images.

2. SUBJECTIVE QUALITY ASSESSMENT

2.1. Image Database

We select 25 hazy images to cover diverse outdoor scenes and different degrees of haze thickness. These include humans, animals, plants, architectures, landscapes, statics, traffics and night scenes, as shown in Fig. 1. Some images are frequently used in the literature of image dehazing. Most of the images were captured in the real-world, but the haze of the three indoor static objects are simulated homogenously.

Eight dehazing algorithms are selected to cover a variety of dehazing methodologies and behaviors. These include simple operator 1) Photoshop auto-contrast [39] and advanced algorithms 2) He09 [11], 3) Kim13 [19], 4) Kolor Neutralhazer [40], 5) Meng13 [20], 6) Tang14 [23], 7) Tarel09 [14] and 8) Xiao12 [17]. The dehazing results of Kim13 [19] and Meng13 [20] are generated by the authors. In other cases, default parameter settings are adopted without tuning for better quality. Eventually, we have a total of 225 images (25 hazy ones and 200 dehazed ones) and divide them into 25 image sets with 9 images each. The images from the same set are from the same source image content, as exemplified in Fig. 2, where we observe that different dehazed images have substantially different perceptual appearance. Interestingly, the quality of some dehazed images are even worse than the hazy one at the first glance. This motivates us to conduct a comprehensive subjective user study to evaluate the performance of those dehazing algorithms quantitatively.



(a) Hazy image (b) He09 [11] (c) Kim13 [19]

(d) Kolor [40] (e) Meng [20] (f) Photoshop [39]

(g) Tang14 [23] (h) Tarel09 [14] (i) Xiao12 [17]

Fig. 2. Sample hazy and dehazed images from one image set.

2.2. Subjective User Study

The subjective user study is conducted at the University of Waterloo in the Image and Vision Computing (IVC) laboratory, which has a normal lighting condition without reflecting ceiling walls and floor. A Truecolor (32 bits) LCD monitor resolution of 2560×1600 pixel is used to display all images. The monitor is calibrated in accordance with the recommendations of ITU-R BT.500 [41]. A customized MATLAB figure window is built to render all 9 images to the subject simultaneously at their original pixel resolution but in random spatial order. A total of 24 naive observers, including 12 male and 12 female subjects aged between 22 and 28, participated in the subjective experiment. The subjects are allowed to adjust their positions for better observation. The length of a session is limited to a maximum of 30 minutes in order to minimize the influence of fatigue effect.

The subjects scored each image with an integer from 1 to 10 that best reflects its perceptual quality. 1 denotes the worst quality and 10 the best. The instructors neither provided the subjects backgrounds of this study nor instructed them with other sample images. In other words, the subjects gave opinions completely based on their own prior knowledge and preference. The reasons to adopt this testing strategy are as follows [42]. First, it is highly efficient since multiple scores are collected at one shot. Second, it reduces memory effect because a full set of images are evaluated at the same time, making it easier for the subjects to apply the same scoring strategy to all images. Third, because the absolute category ratings being collected also inherently contain ranking information, both linear and rank-order correlation evaluations can be directly applied in the data analysis stage. Finally, quality comparison across source images of different content is meaningful in the development of objective IQA models to test and improve their generalization capabilities.

3. ANALYSIS AND DISCUSSION

3.1. Subjective Data Analysis

Based on the outlier detection and removal scheme in [41], we find that all 24 subjects are valid. The final quality score, namely the

Table 1. PLCC and SRCC evaluations between the subjective scores of individual observers and MOSs across the database

Subject	PLCC	SRCC	Subject	PLCC	SRCC
1	0.6239	0.5871	14	0.6349	0.6340
2	0.7578	0.7342	15	0.6673	0.6570
3	0.8301	0.8273	16	0.6975	0.6739
4	0.6502	0.6147	17	0.6698	0.6571
5	0.5537	0.5293	18	0.6538	0.6182
6	0.6626	0.6332	19	0.5148	0.4706
7	0.7137	0.7189	20	0.6210	0.6258
8	0.5657	0.5570	21	0.6047	0.5868
9	0.7824	0.7793	22	0.7264	0.7107
10	0.6569	0.6661	23	0.5820	0.5351
11	0.6535	0.6424	24	0.7345	0.7339
12	0.6577	0.6615	Average	**0.6620**	**0.6462**
13	0.6733	0.6553			

mean opinion score (MOS), is computed by averaging subjective scores from all subjects. We can evaluate the performance of individual subjects against the MOSs (the "ground truth") in the following two ways: 1) compare their scores with the MOSs for all test images, where data across different content are analyzed together; 2) compute the correlation for each image set and then take the average. We employ Pearson linear correlation coefficient (PLCC) and Spearman's rand-order correlation coefficient (SRCC) [43] as evaluation criteria. Both criteria lie in $[0, 1]$ with a higher value standing for better performance. Table 1 lists the PLCC and SRCC values for all subjects across the database, from which considerable agreement on the quality of hazy and dehazed images can be observed. We also compute PLCC and SRCC values of individual subjects for each image set. By averaging the results of all 24 subjects, we obtain the general performance of an average observer, which can be served as a baseline to test objective IQA models. The statistics are summarized in Fig. 3, where the performance of an average subject is given at the rightmost column.

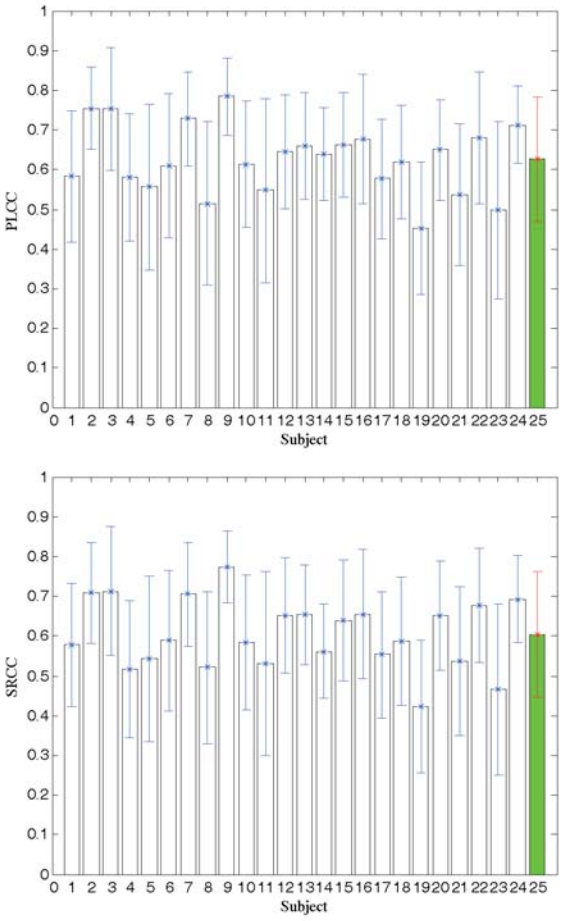


Fig. 3. PLCC and SRCC values of each individual subject ratings against MOSs. Rightmost column: average subject performance.

3.2. Performance of Dehazing Algorithms

As described in Section 3.1, the MOS of a dehazed image created by a certain dehazing algorithm is a good indicator of its performance. Therefore, we compute the mean and standard deviation (std) of the MOSs for each algorithm as well as the hazy image across all image sets, as shown in Fig. 4. Furthermore, we test a hypothesis based on t-statistics [44] to evaluate the statistical significance of the subjective experimental results. The null hypothesis is that the MOSs of one dehazing algorithm is statistically indistinguishable (with 95% confidence) from those of another algorithm. The test is carried out for all possible combinations of pairs of algorithms (including hazy image set), and the results are summarised in Table 2. Note that this is only a rough comparison of the dehazing performance of those algorithms. Many algorithms work in their default parameter settings without fine tuning. Besides, computational complexity is not considered in this study.

From Fig 4 and Table 2, we have several useful and somewhat surprising observations. First, all dehazing algorithms have a relatively large error bar, which indicates the performance gain is not reliable for all the algorithms. In fact, not a single algorithm creates the best dehazing results for all test images. Second, the commercialized software Kolor Neutralhazer [40] performs the best on average, although it is statistically indistinguishable from the naive Photoshop auto-contrast method [39]. Generally, Kolor Neutralhazer [40] keeps a good balance between the degree of haze removal and structure recovery, and thus looks more natural with less artifacts and haze compared with other dehazed images as well as the hazy images. Third, the quality of the dehazed images by some algorithms such as He09 [11] and Meng13 [20] are statistically indistinguishable from the hazy images. This is mainly because a variety of distortions are introduced during the extensive haze removal process. For example, in distant areas with heavy haze, the underlying structures may not be captured by the camera; instead the sensor noise of those areas can easily be amplified after haze removal. For another example, some hazy images may be JPEG compressed immediately after being acquired. This may not cause visual degradation in hazy images, but the blocking artifacts become clearly visible after haze removal because they are mistakenly considered as underlying structures by the dehazing algorithms. Other distortions that seriously decrease the perceived quality of dehazed images include halo artifacts in the background and near edges due to inaccuracy

Table 2. Statistical significance matrix based on the hypothesis testing. A symbol "1" means that the performance of the row algorithm is statistically better than that of the column algorithm, a symbol "0" means that the row algorithm is statistically worse, and a symbol "-" means that the row and column algorithms are statistically indistinguishable.

	Hazy	He09 [11]	Kim13 [19]	Kolor [40]	Meng13 [20]	PS [39]	Tang14 [23]	Tarel09 [14]	Xiao12 [17]
Hazy	-	-	0	0	-	0	-	1	0
He09 [11]	-	-	0	0	0	0	0	1	0
Kim13 [19]	1	1	-	-	1	-	1	1	-
Kolor [40]	1	1	-	-	1	-	1	1	1
Meng13 [20]	-	1	0	0	-	0	-	1	0
PS [39]	1	1	-	-	1	-	1	1	-
Tang14 [23]	-	1	0	0	-	0	-	1	0
Tarel09 [14]	0	0	0	0	0	0	0	-	0
Xiao12 [17]	1	1	-	0	1	-	1	1	-

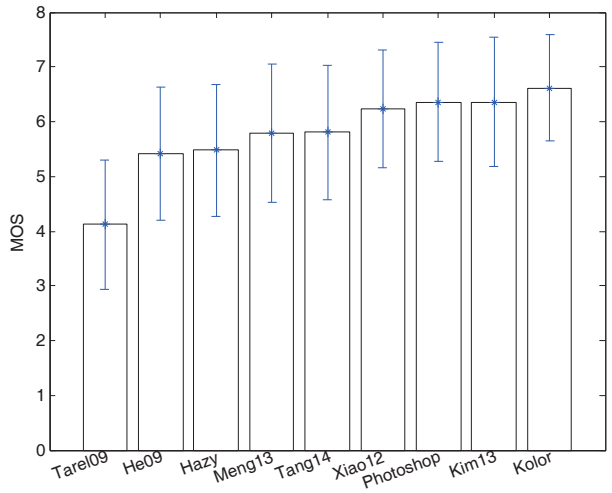


Fig. 4. Mean and std of subjective ratings of dehazing algorithms along with the hazy image across all image sets.

of transmission estimation, artificial edges, color saturation due to over-contrast enhancement, and unnatural appearance due to color distortion or overall dark luminance. These distortions may counteract the effort of haze removal and cause new problems in subsequent computer vision tasks. In summary, single image dehazing is a very challenge problem and still has much room for further improvement. For example, most physical model based dehazing algorithms put a lot of effort on the evaluation of the transmission coefficient t in Eq. (1), but simply estimate the atmospheric light $\mathbf{A}$ using somewhat ad hoc methods. It turns out that the selection of $\mathbf{A}$ is critical in improving the perceptual quality of dehazed images [20, 23] and needs to be seriously investigated.

3.3. Performance of Existing Objective IQA Models

We test five general purpose no-reference objective IQA models [30, 31, 32, 33, 34], one no-reference model for contrast-distorted images [35] and one model [36] specific for dehazing images. The results are tabulated in Table 3. As we can see, none of the IQA models properly predicts the perceived quality of dehazed images. This is mainly because most existing no-reference IQA models are designed to work with typical distortions such as luminance, blur and blocking artifacts, and do not have the generalization capabilities to account for many distortions introduced during the dehazing process discussed previously. The model in [36] only focuses on the recovered structures and saturated pixels while ignoring the naturalness of the dehazed images. This may not be appropriate since the recovered structures may be created from amplified background noise and blocking artifacts. In summary, our study suggests that natural scene statistics as well as distortion specific features may be combined to yield a more accurate objective IQA model in the future.

Table 3. Performance evaluation of objective IQA models

IQA model	PLCC	SRCC
BIQI [30]	0.1688	0.1557
BRISQUE [31]	0.1749	0.1674
NIQE [32]	0.2051	0.1732
DILT [34]	0.0304	-0.0353
BLINDS-II [33]	0.0604	-0.0204
NCDQI [35]	0.2748	0.2765
Hautière e [36]	-0.1442	-0.0876
Hautière r [36]	-0.0405	-0.0301
Hautière N_s [36]	-0.0478	-0.2416

4. CONCLUSIONS AND DISCUSSION

Single image dehazing has been an active research topic recently, but little work has been dedicated to subjective and objective quality assessment of dehazed images. We make one of the first attempts to evaluate both classic and advanced dehazing algorithms, as well as related IQA models. A new image database is established and subjective tests are conducted. Data analysis shows that not a single dehazing algorithm is reliable enough to create high quality dehazing results for all test images. The quality of dehazed images generated by some state-of-the-art dehazing algorithms are statistically equivalent to the hazy images. Our study also shows that none of the existing objective IQA models gives proper quality predictions of dehazed images. Careful observations suggest that future dehazing algorithms need to keep a balance between the natural appearance and the degree of haze removal in a dehazed image. Furthermore, objective IQA models that incorporate natural scene statistics and distortion specific features may have the potential to better predict the perceived quality of dehazed images.

5. REFERENCES

[1] H. Koschmieder, *Theorie der horizontalen Sichtweite: Kontrast und Sichtweite*. Beiträge zur Physik der freien Atmosphäre, Keim & Nemnich, 1924.

[2] S. K. Nayar and S. G. Narasimhan, "Vision in bad weather," in *ICCV*, 1999.

[3] S. G. Narasimhan and S. K. Nayar, "Chromatic framework for vision in bad weather," in *CVPR*, 2000.

[4] Y. Y. Schechner, S. G. Narasimhan, and S. K. Nayar, "Instant dehazing of images using polarization," in *CVPR*, 2001.

[5] S. Shwartz, E. Namer, and Y. Y. Schechner, "Blind haze separation," in *CVPR*, 2006.

[6] S. G. Narasimhan and S. K. Nayar, "Interactive (de) weathering of an image using physical models," in *IEEE Workshop on Color and Photometric Methods in Computer Vision*, 2003.

[7] J. Kopf, B. Neubert, B. Chen, M. Cohen, D. Cohen-Or, O. Deussen, M. Uyttendaele, and D. Lischinski, "Deep photo: Model-based photograph enhancement and viewing," *ACM TOG*, 2008.

[8] R. Fattal, "Single image dehazing," *ACM TOG*, 2008.

[9] R. T. Tan, "Visibility in bad weather from a single image," in *CVPR*, 2008.

[10] P. S. Chavez Jr, "An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data," *Remote sensing of environment*, 1988.

[11] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in *CVPR*, 2009.

[12] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE TPAMI*, 2008.

[13] K. He, J. Sun, and X. Tang, "Guided image filtering," in *ECCV*, 2010.

[14] J.-P. Tarel and N. Hautiere, "Fast visibility restoration from a single color or gray level image," in *ICCV*, 2009.

[15] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *ICCV*, 1998.

[16] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM TOG*, 2007.

[17] C. Xiao and J. Gan, "Fast image dehazing using guided joint bilateral filter," *The Visual Computer*, vol. 28, no. 6-8, pp. 713–721, 2012.

[18] Q. Yang, K.-H. Tan, and N. Ahuja, "Real-time o (1) bilateral filtering," in *CVPR*, 2009.

[19] J.-H. Kim, W.-D. Jang, J.-Y. Sim, and C.-S. Kim, "Optimized contrast enhancement for real-time image and video dehazing," *Journal of Visual Communication and Image Representation*, vol. 24, no. 3, pp. 410–425, 2013.

[20] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in *ICCV*, 2013.

[21] R. Fattal, "Dehazing using color-lines," *ACM TOG*, 2014.

[22] I. Omer and M. Werman, "Color lines: Image specific color representation," in *CVPR*, 2004.

[23] K. Tang, J. Yang, and J. Wang, "Investigating haze-relevant features in a learning framework for image dehazing," in *CVPR*, 2014.

[24] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[25] E. Matlin and P. Milanfar, "Removal of haze and noise from a single image," in *IS&T/SPIE Electronic Imaging*, International Society for Optics and Photonics, 2012.

[26] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE TIP*, 2007.

[27] F. Fang, F. Li, and T. Zeng, "Single image dehazing and denoising: A fast variational approach," *SIAM Journal on Imaging Sciences*, vol. 7, no. 2, pp. 969–996, 2014.

[28] C. O. Ancuti and C. Ancuti, "Single image dehazing by multi-scale fusion," *IEEE TIP*, 2013.

[29] Z. Chen, T. Jiang, and Y. Tian, "Quality assessment for comparing image enhancement algorithms," in *CVPR*, 2014.

[30] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE SPL*, 2010.

[31] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE TIP*, 2012.

[32] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE SPL*, 2013.

[33] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE TIP*, 2012.

[34] Q. Wu, H. Li, K. N. Ngan, B. Zeng, and M. Gabbouj, "No reference image quality metric via distortion identification and multi-channel label transfer," in *IEEE ISCAS*, 2014.

[35] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, and G. Zhai, "No-reference quality assessment of contrast-distorted images based on natural scene statistics," *IEEE SPL*, 2015.

[36] N. Hautière, J.-P. Tarel, D. Aubert, E. Dumont, *et al.*, "Blind contrast enhancement assessment by gradient ratioing at visible edges," *Image Analysis & Stereology Journal*, 2008.

[37] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, 2001.

[38] M. A. Stricker and M. Orengo, "Similarity of color images," in *IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology*, 1995.

[39] Photoshop. `http://www.photoshop.com/`.

[40] Kolor neutralhazer plugin for photoshop. `http://www.kolor.com/`.

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

[42] K. Zeng, K. Ma, R. Hassen, and Z. Wang, "Perceptual evaluation of multi-exposure image fusion algorithms," in *6th International Workshop on Quality of Multimedia Experience*, 2014.

[43] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," Apr., 2000. [Online]. Available: http://www.vqeg.org.

[44] D. C. Montgomery, *Applied Statistics and Probability for Engineers 6th edition*. Wiley, 2013.