A COMPREHENSIVE PERFORMANCE EVALUATION OF OBJECTIVE QUALITY METRICS FOR CONTRAST ENHANCEMENT TECHNIQUES

Muhammad Ali Qureshi[1], *Azeddine Beghdadi*[2], *Bilel Sdiri*[2,3], *Mohamed Deriche*[1], *Faouzi Alaya-Cheikh*[3]

[1]King Fahd University of Petroleum and Minerals, Saudi Arabia
[2]L2TI, Institut Galilée, Université Paris 13, Sorbonne Paris Cité, France
[3] NTNU, Color and Visual Computing Lab, Gjϕvik, Norway

ABSTRACT

In this paper, we present a comprehensive analysis and comparison of state-of-the-art Contrast Enhancement (CE) evaluation metrics. In this work, we developed a new database consisting of 182 images. The subjective experiments were performed to obtain a preference rating from different observers for the enhanced images using six CE methods selected from different representative categories. The quality of the enhanced images is measured with most commonly used CE evaluation metrics. We provide the ranking of images based on perceptual preference as well as objective quality metrics scores. We show that some of the metrics used for the enhancement evaluation are not consistent with the human subjective scores. This new database, named as Contrast Enhancement Evaluation Database (CEED2016), is made publicly available to the research community at http://www-l2ti.univ-paris13.fr/site/index.php/en/CEED2016/ and is expected to be a contribution to the area of Image Quality Assessment (IQA) and particularly for image quality enhancement evaluation.

Index Terms— image enhancement, image quality assessment, subjective experiments, contrast enhancement database, contrast enhancement evaluation metrics

1. INTRODUCTION

The contrast is one of the most important factors that affects the perceptual quality of images. This psychophysical factor is highly related to the acquisition environment and the sensors limitations. The Human Visual System (HVS) is also very sensitive to the image contrast. Due to the context-dependent nature of the contrast enhancement, there is no single, universally accepted definition of image contrast. To improve the quality of low or poor contrast images, different enhancement techniques have already been proposed in the literature. The main goal of a CE algorithm is to improve the perceptual quality of images by enhancing salient features and without altering global appearance (pleasantness). Contrast enhancement evaluation is also complex due to the effects of numerous parameters and factors, such as image pleasantness, naturalness, illumination, colorfulness, etc. The quality of enhancement is typically evaluated using objective metrics. In the literature, many objective metrics have been proposed and are used for quality evaluation of enhancement algorithms. But there is no common benchmark for the evaluation of the different CE evaluation metrics themselves. Very little work can be seen in this regard. Simone et al. [1] proposed a scheme to measure the perceptual contrast of digital images. They performed subjective experiments on a set of images and gathered quality scores from different observers. The correlation results between the subjective and the objective scores were not encouraging. Damon Chandler et al. [2], proposed a database consisting of retouched images. The main goal of this database was to validate the performance of existing IQA metrics (designed to estimate the quality of a reproduction, filtered or degraded image) for quality prediction of enhancement. Recently, another database CID:IQ [3] was introduced. It contains images degraded with different types of distortions and images with the variation of contrast due to gamut mapping. The main objective of these databases was to validate the performance of existing IQA metrics designed mainly for the degraded images. Unlike the available databases [4], dedicated to various common distortions evaluation (decrease in image quality), our database is designed for CE evaluation (increase in image quality). It is worth to mention that we are not comparing the CE in terms of most highly contrasted-images. However, we are searching for the CE which enhances the image contrast without introducing artifacts. This will help in searching for the best metric that account for the CE artifacts. This new framework and the results for the database will be very useful for the research community focusing on image enhancement techniques as well as CE evaluation metrics.

The main objectives of our study are: (1) to introduce a new database for assessment of different CE evaluation metrics, (2) to perform a psychophysical experiment to rank CE algorithms, (3) performance analysis of different CE evaluation metrics by computing the correlation between objective scores and subjective ranking of each set of enhanced images. Therefore, with the subjective ratings of the enhanced images

and the corresponding objective scores, we can comment on the consistency of a given CE evaluation metric with the perceptual human judgment of CE quality. This database may be also useful for the evaluation of newly introduced CE evaluation metrics.

In the next sections, we will introduce briefly the enhancement methods used for the database, the evaluation metrics under comparisons, and the protocol for the subjective experiments followed by a discussion of the results.

2. IMAGE ENHANCEMENT TECHNIQUES

Image contrast enhancement is usually subjective as it relates to the way, we perceive details of some desired features in an image. There is a plethora of image CE methods. The CE algorithms can be grouped into histogram-based, pixel-based, transform-based, mathematical morphology-based, and HVS-based, etc. In the literature, we can find numerous methods designed for different applications. In this paper, we selected six CE methods as representatives from the categories mentioned above. These methods are: Adaptive Edge Based Contrast Enhancement (AEBCE) [5], Contrast Limited Adaptive Histogram Equalization (CLAHE) [6], Discrete Cosine Transform based (DCT) [7], Global Histogram Equalization (GHE), Top Hat Transformation based (TOPHAT) [8], and Multiscale Retinex (MRETINEX) [9]. We have used the codes of these methods provided by the authors. For GHE and CLAHE, we have used the Matlab built-in functions `histeq` and `adapthisteq` respectively. The main goal of the study is to compare CE evaluation metrics instead of CE methods, therefore, for our experiments, we have used the CE algorithms with their default parameters without tuning the algorithms for performance improvement.

3. CONTRAST ENHANCEMENT EVALUATION METRICS

Although considerable efforts have been devoted to the development of image CE algorithms, there is no universal measure to evaluate enhancement quality due to its dependence on the applications and the subjective nature of image perception. Several measures that have been proposed for CE evaluation, are not, however, consistent with the visual appearance of the processed images. One of the first studies on contrast enhancement evaluation based on the bimodality analysis of the gray-level distribution was proposed in [10]. Among different CE measures, we have selected the EME (Measure of Enhancement) [11], AME (Absolute Measure of Enhancement) [11], SDME (Second Derivative like Measure of Enhancement) [11], EC (Edge Content) [12], IEM (Image Enhancement Measure) [13], and Absolute Mean Brightness Error (AMBE) [14] in this work (see Table 1).

The metrics EME and AME are inspired from Weber contrast and Michelson contrast respectively. Note, both Weber contrast and Michelson contrast measures are not applicable to natural images. To simulate these contrast measures on natural images, both EME and AME are calculated on small image blocks and the final measure is estimated on all these blocks. Both EME and AME are also sensitive to noise due to the use of maximum and minimum operator in their expressions. By incorporating the center pixel value within the block in the expression of SDME, the measure becomes less sensitive to noise. The default values for the block size are either 3×3 or 5×5. The EME and the AME are computed on non-overlapping blocks of size 8×8, whereas, SDME is computed for non-overlapping blocks of size 5×5. The metric EC is based on the local gradient of the image intensity. In the expression of EC, ΔI represents the gradient image intensity by using Sobel operator. Note that, higher values of EC correspond to images with higher contrast.

For contrast enhancement, it is desirable to preserve the original brightness of the original image. The AMBE metric represents the deviation of mean intensity of the enhanced image from the original image. It is used to measure brightness preservation of CE algorithms. In its expression, $E(.)$ represents the statistical expectation. Here, brightness preservation does not mean that the image natural look (quality) is also preserved. The lower value of AMBE means that the enhanced image has good brightness preservation. A median value of AMBE implies better brightness preservation. Either a very low or very high value of AMBE also indicates poor performance in case of contrast enhancement. In the IEM, both original and enhanced images are divided into non-overlapping blocks (3×3). The measure is calculated as the ratio of the sum of absolute values of differences of the center pixel from its eight neighbors in all blocks of an enhanced im-

Table 1. The expressions of the CE evaluation measures used in our experiments

No.	CE evaluation metrics expressions
1.	$EME = \frac{1}{B_1 \times B_2} \sum_{i=1}^{B_1} \sum_{j=1}^{B_2} 20 \ln \left(\frac{I_{ij}^{max}}{I_{ij}^{min}+c} \right)$
2.	$AME = \frac{-1}{B_1 \times B_2} \sum_{i=1}^{B_1} \sum_{j=1}^{B_2} 20 \ln \left(\frac{I_{ij}^{max}-I_{ij}^{min}}{I_{ij}^{max}+I_{ij}^{min}} \right)$
3.	$SDME = \frac{-1}{B_1 \times B_2} \sum_{i=1}^{B_1} \sum_{j=1}^{B_2} 20 \ln \left\| \frac{I_{ij}^{max}-2I_{ij}^{cen}+I_{ij}^{min}}{I_{ij}^{max}+2I_{ij}^{cen}+I_{ij}^{min}} \right\|$
4.	$EC = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \|\Delta I(i,j)\|$
5.	$IEM = \frac{\sum_{i=1}^{B_1} \sum_{j=1}^{B_2} \sum_{n=1}^{8} \|I_{ij}^{e,cen}-I_{ij}^{e,n}\|}{\sum_{i=1}^{B_1} \sum_{j=1}^{B_2} \sum_{n=1}^{8} \|I_{ij}^{r,cen}-I_{ij}^{r,n}\|}$
6.	$AMBE = \|E(I_r) - E(I_e)\|$

I_{ij}^{max}, I_{ij}^{min}, and I_{ij}^{cen} are the maximum, minimum, and center pixel intensity within the block (i,j), respectively.

age and the corresponding blocks in the original image. The absolute intensity difference between a pixel and its neighbors corresponding to the reference and enhanced images are used to represent a change in contrast and sharpness.

In an enhanced image, high values of EME, EC, IEM, represents better contrast, whereas for AME and SDME, the opposite is true. For the purpose of completeness, we list the mathematical expressions used for the calculation of the above-mentioned metrics in Table 1. Here, in the expressions, the original and enhanced images are represented by I_r and I_e respectively. The variables M and N represent image height and width respectively, L is the number of gray levels, b is the block size and c is the denominator constant to avoid division by zero (usually 0.0001).

In our earlier study [15], we have compared different CE evaluation metrics to show how these metrics may help in evaluating the artifacts in CE process. A new metric based on mutual information computed from the gray level co-occurrence matrix was proposed and tested on two classical neighborhood-based CE methods. Here, we extend this study and propose a database, to see how different CE evaluation metrics are consistent with the human judgement of image quality enhancement.

4. EXPERIMENTAL SETUP

To assess the correlation between the HVS perception and the different CE evaluation measures, we performed an intensive psychophysical experiment. Initially, we constructed a database consisting of 30 original color images with size of 512×512 pixels. Most of the images are captured by lab members, while we have also included some classical images commonly used by the signal processing community. It is well understood that the human perception of image quality is highly dependent upon the scene under observation. For this reason, we selected images with different contents, color distribution, and contrast variations. We have used three quantitative measures for the selection of images. These measures are Spatial Information (SI) [4], Colorfulness (CF) [4], and Global Contrast Factor (GCF) [16].

The contrast enhanced images were generated from the six selected CE methods mentioned in the previous section. In Fig. 1, we show some original images with different content, and contrast variations and their enhanced versions. To obtain the ranking scores, we adopted pairwise preference based ranking protocol (Condorcet method). For this, for each original image, we randomly displayed pair of enhanced images to the observers. We displayed also the original image in the center of the screen (pair of enhanced images are to its left and right), in order to facilitate the analysis on after effects of contrast enhancement.

The subjective experiments were performed at Université Paris 13 at Laboratoire de Traitement et Transport de l'Information (L2TI). The images were displayed on EIZO ColorEdge CG242W (24.1 inches) with the screen resolution of 1920×1200. Since we deal with contrast enhancement, the experiment room was kept in low level of ambient illumination as done in [3], to allow the observers to focus more on the observed scenes. The background color of the screen was set to gray. The display screen was calibrated with an Eye-One Match 3, with sRGB color space at 119 cd/m2 white point luminance value and color temperature of 6500K. The Gamma value was set to a value of 2.2 at a display frame rate of 60Hz and contrast 80.

Fifteen observers both experts and non-experts and coming from different age groups, gender, and background, took part in the experiment. All the observers had either normal vision or corrected to normal vision. The observers were forced to perform the experiments at a fixed distance of 80 cm from the screen. The observers were not informed about the definition of contrast and were asked to give their preference about which image they feel perceptually better than the other compared to the original image by taking into consideration the artifacts introduced due to CE such as halo effects, unnaturalness etc. They were allowed to give the same rank for both images in the case of equivalent degree of quality.

We have used the Kendall's Coefficient of Concordance [17] to ensure the significance of the ranking and to ensure the agreement of all observers on the rankings. From the statistical analysis, we have found disagreement among the observers' rankings on 4 images, and we discarded the results of these images for further analysis.

5. RESULTS AND DISCUSSIONS

From the subjective experiments, we have derived the preferences scores (the number of times the algorithm is preferred over others) of the six CE algorithms. The subjective preference scores of the CE methods are shown in Table 2. From the preference scores, it is clear that CLAHE is highly preferred whereas GHE and TOPHAT are not preferred most of the time. The objective scores of the six CE evaluation metrics for the enhanced images created from the six CE techniques are also calculated. The objective scores for each metric averaged over the 26 images in the database, are also shown in Table 2.

In order to observe the consistency of the CE evaluation metrics with the human visual perception, we calculated the Spearman Rank Order Correlation Coefficient (SROCC) between the ranking scores and the objective scores of each metric and for the six enhanced versions of each original image in the database separately. The SROCC is a non-parametric measure that depends only on the ranks of the data variables ranging from -1 to $+1$. In the case of strong correlation, its value is $+1$, in case perfect disagreement, it is -1, and it is zero, when there is no correlation exists.

If I_i and $I_{i,j}$ represents original image and its enhanced image enhanced with method A_j, for $i = 1, 2, \cdots, n_I$ and

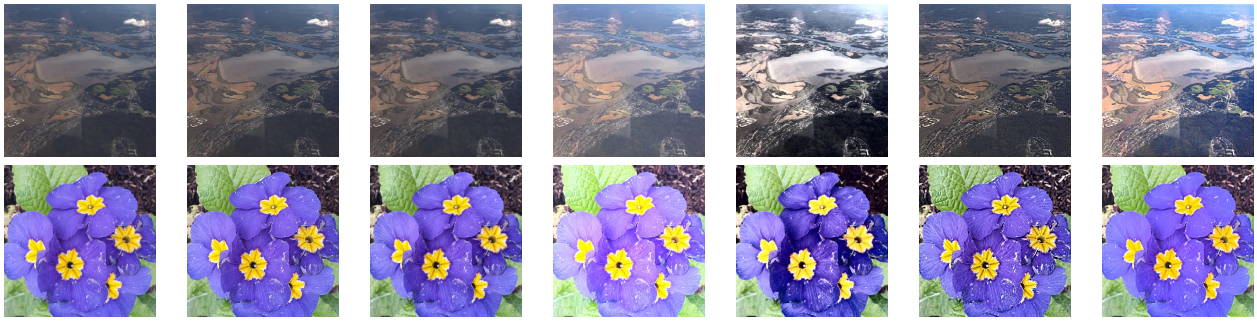


Fig. 1. Examples of two original images and the enhanced images from the database (from Top Left) original image, then enhanced images with AEBCE, CLAHE, DCT, GHE, TOPHAT, and MRETINEX based CE methods

Table 2. Summary of contrast metrics objective scores and subjective rankings for different contrast enhancement methods

	AEBCE	CLAHE	DCT	GHE	TOPHAT	MRETINEX
Subjective Rankings	1388	1422.5	779	591.50	468.5	1300.5
EME	0.4932	0.1998	-0.1443	1.3161	1.3190	0.4263
AME	-0.2529	0.0907	-0.0684	0.5746	0.6865	0.1443
EC	0.4305	0.1698	0.2685	0.7204	1.0722	0.5348
AMBE	3.7171	0.6813	0.0215	0.0631	6.3516	2.1987
SDME	0.1400	0.0639	-0.0122	0.2057	0.3253	0.1128
IEM	1.7208	1.1836	1.3125	1.8098	2.2861	1.5656

$j = 1, 2, \cdots, n_J$, for $(n_I = 26, n_J = 6)$. The SROCC is calculated as follows:

$$\rho_{i,k} = 1 - \frac{6d_{i,j}^2}{n_J(n_J^2 - 1)}, \text{ for } i = 1, 2, \cdots, n_I \qquad (1)$$

where $d_{i,j}$ represents the difference between the ranks of subjective preferences objective scores of metric k for each image (see Fig. 2). Finally, the median values of the SROCC's for each metric is calculated and are shown in Table 3.

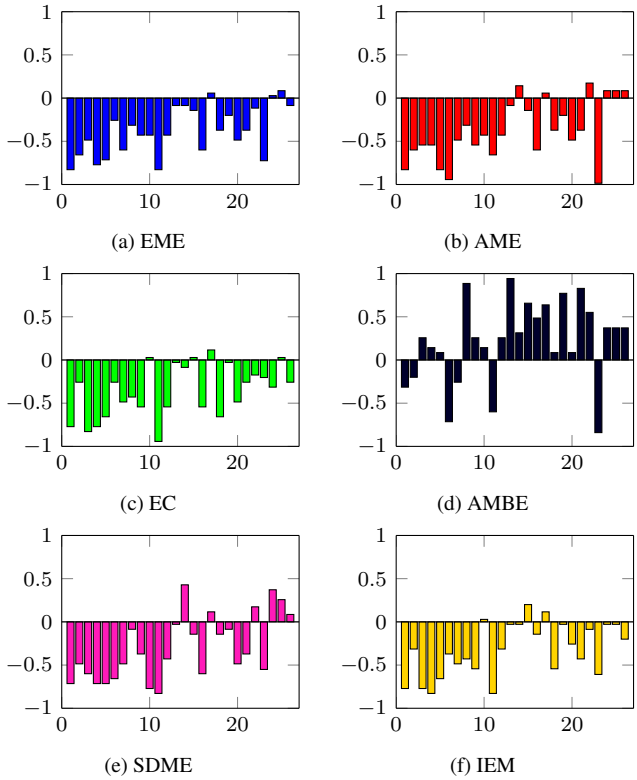
Table 3. Median SROCC results for the CE evaluation metrics under study.

Metric	SROCC (ρ)	Metric	SROCC (ρ)
EME	-0.4000	AMBE	+0.2571
AME	-0.4286	SDME	-0.4000
EC	-0.2857	IEM	-0.3143

In these experiments, the correlations are calculated between the rankings of CE methods from subjective preferences and metric scores. It is worth noted that for AMBE, AME, and SDME, the small values correspond to good image quality, whereas for other metrics in comparison, good quality corresponds to large values. From Fig. 2, it could be observed that only AMBE metric shows a positive correlation with the subjective ranking. The negative SROCC values of other metrics show strong disagreement with subjective preferences. This is mainly due to the two CE methods i.e., TOPHAT and GHE, where all metrics except AMBE are highly ranked. In contrast, these methods are least preferred in the subjective experiments. We can also draw a conclusion from these correlation values, that most of the existing CE evaluation metrics could not quantify the improvement in image quality in a consistent manner for all the CE methods.

6. CONCLUSION

In this paper, we have performed both subjective and objective comparisons of some commonly used CE evaluation metrics for the newly proposed database consisting of enhanced images from six different CE techniques. Based on the correlation of preference based ranking data gathered from subjective experiments, and the objective scores of different CE evaluation metrics, we have found that some of these metrics are not consistent with the human perceptual judgments. It is worth noticing that some of the considered metrics do not take into account the most relevant characteristics of the HVS. Therefore, HVS-inspired measure for CE evaluation is still missing. The difficulties are due to the context-dependent nature of contrast enhancement problem. The contrast enhancement evaluation is also complex due to the effect of many parameters, such as image pleasantness, naturalness, and colorfulness. Taking into account these aspects is a challenging problem that could be considered in a future work in the design of new CE evaluation metric.



(a) EME (b) AME (c) EC (d) AMBE (e) SDME (f) IEM

Fig. 2. SROCC plots between different CE evaluation metrics values and subjective rankings

7. REFERENCES

[1] G. Simone, M. Pedersen, and J. Y. Hardeberg, "Measuring perceptual contrast in digital images," *Journal of Visual Communication and Image Representation*, vol. 23, no. 3, 2012.

[2] C. T. Vu, T. D. Phan, P. S. Banga, and D. M. Chandler, "On the quality assessment of enhanced images: A database, analysis, and strategies for augmenting existing methods," in *2012 IEEE Southwest Symposium on Image Analysis and Interpretation*. apr 2012, pp. 181–184, IEEE.

[3] X. Liu, M. Pedersen, and J. Y. Hardeberg, "CID:IQ - A New Image Quality Database," in *International Conference on Image and Signal Processing (ICISP)*, Cherbourg, Normandy, France, jul 2014, vol. 8509 of *Lecture Notes in Computer Science*, pp. 193–202, Springer International Publishing.

[4] S. Winkler, "Analysis of Public Image and Video Databases for Quality Assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625, oct 2012.

[5] A. Beghdadi and A. Le Negrate, "Contrast enhancement technique based on local detection of edges," *Computer Vision, Graphics, and Image Processing*, vol. 46, no. 2, pp. 162–174, may 1989.

[6] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics gems IV*, pp. 474–485. Academic Press Professional, Inc., San Diego, CA, USA, 1994.

[7] J. Mukherjee and S. Mitra, "Enhancement of Color Images by Scaling the DCT Coefficients," *IEEE Transactions on Image Processing*, vol. 17, no. 10, pp. 1783–1794, oct 2008.

[8] S. Mukhopadhyay and B. Chanda, "A Multiscale Morphological Approach to Local Contrast Enhancement," *Signal Processing*, vol. 80, pp. 685–696, dec 2010.

[9] S. Chen and A. Beghdadi, "Natural Enhancement of Color Image," *EURASIP Journal on Image and Video Processing*, vol. 2010, pp. 1–19, aug 2010.

[10] A. Le Négrate, A. Beghdadi, and H. Dupoisot, "An image enhancement technique and its evaluation through bimodality analysis," *CVGIP: Graphical Models and Image Processing*, vol. 54, no. 1, pp. 13–22, jan 1992.

[11] K. Panetta, C. Gao, and S. Agaian, "No reference color image contrast and quality measures," *IEEE Transactions on Consumer Electronics*, vol. 59, no. 3, pp. 643–651, aug 2013.

[12] A. Saleem, A. Beghdadi, and B. Boashash, "Image fusion-based contrast enhancement," *EURASIP Journal on Image and Video Processing*, , no. 1, pp. 10, 2012.

[13] V. L. Jaya and R. Gopikakumari, "IEM : A New Image Enhancement Metric for Contrast and Sharpness Measurements," *International Journal of Computer Applications*, vol. 79, no. 9, pp. 1–9, oct 2013.

[14] S.-D. Chen and A. R. Ramli, "Minimum mean brightness error Bi-histogram equalization in contrast enhancement," *IEEE Transactions on Consumer Electronics*, vol. 49, no. 4, pp. 1310–1319, nov 2003.

[15] A. Beghdadi, M. A. Qureshi, and M. Deriche, "A critical look to some contrast enhancement evaluation measures," in *2015 Colour and Visual Computing Symposium (CVCS)*, Gjovik, Norway, aug 2015, pp. 1–6, IEEE.

[16] K. Matkovic, L. Neumann, A. Neumann, T. Psik, and W. Purgatholer, "Global Contrast Factor - a New Approach to Image Contrast," *Computational Aesthetics in Graphics, Visualization and Imaging*, pp. 159–168, 2005.

[17] S. Siegel, *Nonparametric Statistics for The Behavioral Sciences*, McGraw-Hill, 2nd edition, 1956.