Signal Processing: Image Communication 58 (2017) 212–227

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Towards the design of a consistent image contrast enhancement evaluation measure

Muhammad Ali Qureshi [a,b,*], Azeddine Beghdadi [c], Mohamed Deriche [a]

[a] King Fahd University of Petroleum and Minerals (KFUPM), Dhahran 31261, Saudi Arabia
[b] The Islamia University of Bahawalpur, 63100, Pakistan
[c] L2TI, Institut Galilée, Université Paris 13, Sorbonne Paris Cité, France

ARTICLE INFO

Keywords:
Image enhancement
Image quality assessment
Subjective experiments
Image contrast
Contrast enhancement evaluation measures
Contrast changed database

ABSTRACT

Contrast Enhancement Evaluation (CEE) is a very challenging problem. In this work, we provide a detailed performance analysis of CEE measures. The study was conducted on our newly developed database dedicated to psychophysical Contrast Enhancement (CE) quality evaluation. The database contains 30 original images and 180 enhanced images obtained using six different CE methods as a representative set of the most common approaches used in the literature. The CE methods were subjectively evaluated and ranked by 23 observers using a Pairwise Comparison (PC) protocol. The correlation analysis between the subjective preferences and objective evaluations of the enhanced images show that most of the existing CEE metrics are not well consistent with the human judgment of quality. We also present in this paper a thorough discussion on the available CEE metrics, their strengths, their weaknesses, and their inter-correlation. In addition to the individual metrics, we show that by fusing different metrics together, a significant increase in correlation performance can be achieved. This study reveals that there is a clear need to develop more robust CEE measures which are perceptually motivated and correlated well with the quality of enhancement a given image is subjected to. The new database introduced in this paper is expected to contribute substantially promoting such research area, which is of primary importance to diverse multimedia applications.

1. Introduction

Image Quality Assessment (IQA) has attracted a lot of interest during the last three decades, and a plethora of efficient and advanced IQA measures have been proposed [1–3]. However, some simple and practical image quality measures, such as Mean Square Error (MSE) based measures, are still in use in some multimedia applications such as bit-rate optimization for video coding [4]. This is mainly due to their mathematical tractability and the absence of a well-established standard for measuring image quality. The notion of visual information fidelity or image quality is highly related to the way humans perceive distortions that may affect the quality of the observed image [5,6]. Therefore, the IQA dilemma, in its traditional sense, has been long considered as a distortion estimation problem [7]. This, of course, is an important problem as it is desirable to have ready to use techniques to evaluate the quality of images subject to distortions or artifacts that may result from processing, lossy compression, or transmission. On the other hand, very few studies have been done on the performance evaluation of image enhancement methods (better quality images rather than distorted images). Indeed, performing a quantitative evaluation of image quality enhancement methods is a very challenging task. This is due to the absence of any objective measures able to account for some high-level vision tasks and their interaction with low-level image analysis when assessing the perceptual quality of image enhancement [2]. This is also due to the difficulty in determining the most appropriate visual features to be used in the design of an overall image enhancement quality measure. Therefore, subjective evaluation is still the most reliable approach to assess the quality of enhanced images.

Enhancing image contrast is of major interest in many applications ranging from medical imaging [8], remote sensing [9], underwater imaging [10], image forensics [11], defogging [12], etc. A plethora of Contrast Enhancement (CE) methods has been proposed in the literature, and it becomes rather difficult to provide a comprehensive and complete survey of published work in this area. Moreover, there is no study to test the reliability of these measures themselves. Given the importance of CE in different applications, there is a need to

* Correspondence to: Department of Telecommunication Engineering, The Islamia University of Bahawalpur, 63100, Pakistan.
E-mail addresses: ali.qureshi@iub.edu.pk (M.A. Qureshi), azeddine.beghdadi@univ-paris13.fr (A. Beghdadi), mderiche@kfupm.edu.sa (M. Deriche).

http://dx.doi.org/10.1016/j.image.2017.08.004
Received 16 August 2016; Received in revised form 3 August 2017; Accepted 4 August 2017
Available online 14 August 2017

investigate the performance of these measures in terms of robustness and consistency with human judgment.

One of the first studies on Contrast Enhancement Evaluation (CEE) has been proposed in [13]. However, it was only restricted to images containing two classes of pixels (i.e., one object on a uniform background or many similar objects on a uniform background). The CE evaluation was based on the bimodality analysis of the gray-level distribution. Thereafter, some simple and interesting CEE measures have been proposed in [14–17]. These measures are not inspired by the classical approaches of IQA. The proposed measures are based on the computation of a global index derived from some local measures related to contrast. These are inspired originally by Michelson and Weber–Fechner contrast measures. These measures are based on min–max operations that make them more noise sensitive. The authors proposed some improvements to overcome these limitations by using entropy of local contrast, or by introducing logarithmic arithmetic operations inspired by the non-linear Human Visual System (HVS) response. We should note, that in the study conducted in [14–17], no complete subjective experiments were performed, and the performance analysis was only based on the perceptual judgment of output images. Moreover, the tests were conducted on a limited set of images (very often grayscale images), and the measures were not evaluated on any dedicated database but only on few images from the TID2013 database that has been built for traditional IQA purpose [18]. Furthermore, the statistical analysis of these measures and comparison with some representative CE methods were also missing.

In contrast, Vu et al. [19], proposed another study based on a database containing processed images obtained by changing color, saturation, brightness, sharpness, and their combinations. The subjective evaluation was performed to assess the quality of processed images. The use of classical IQA approaches in a reverse order was proposed, i.e., the given image (enhanced image) is considered as the reference and the original image as the distorted one. It has also been reported that the Visual Information Fidelity (VIF) [20] measure offers better performance as compared to many of the classical IQA measures. The authors in [19] improved the results by proposing a more efficient measure combining contrast, sharpness, and color in an empirical manner.

Following the approach of Vu et al. [19], another study of contrast change evaluation was discussed in [21] using a database consisting of 15 original and 633 enhanced images. The global contrast of images is modified using non-linear mapping functions. The conventional IQA measures designed for degradations assessment were then used to assess the quality of the processed images from the database. For this purpose, a reduced reference metric was derived combining the entropy of phase congruency image and other higher-order statistics of local features computed from the histogram of the observed image.

Another recent study, by Fang et al. [22] on quality assessment of contrast distorted images was carried using the natural scene statistics model. The contrast problem is considered only in terms of distortion. However, in our case, we enhance the contrast and, for CEE, we try to account for the undesirable side effects that may result from CE process.

Besides these works, predicting visual quality of enhanced/modified images for different applications has also been investigated in some interesting studies [23–29]. Ledda et al. [23] proposed a database for only subjective evaluation of six tone mapping methods. The Pairwise Comparison (PC) was performed in a subjective experiment to rank these methods in accordance with the perceived quality. But the authors did not perform CEE performance analysis. Virtanen et al. [24] provided another database related to tone-mapping applications. It contains images degraded with different types of distortions and images with variation of contrast due to gamut mapping. The main objective of the database was to validate the performance of existing IQA metrics designed mainly for degraded images. Another similar database was also proposed in [28] to evaluate gamut mapping, blurring, and other distortions. Similarly, sharpening algorithms are also used to enhance the perceived quality of a given image. In [29], quality evaluation of sharpened images is investigated. A new subjective experiment framework, specifically adopted for the quality assessment of sharpened images was introduced.

In addition to the above, Chen et al. [27] developed a database for CEE of images in bad visibility (i.e., haze, underwater, and low light environment). The images were enhanced through different dehazing methods and the performance of various enhancement algorithms was discussed. In this work, the original and pair of enhanced images were shown on the same screen to allow the observer to compare the enhanced images with respect to the original image.

Another less studied application, namely image retargeting quality assessment, has been addressed in [25,26]. Here, subjective and objective quality evaluation of retargeted images was performed using dedicated databases. In [25], the authors provided a database containing images by different retargeting methods. The subjective quality of the retargeted images was measured in terms of rank in a pairwise subjective experiment, and the performance of different retargeting evaluation measures were assessed in terms of correlation analysis. Similarly, Ma et al. [26] also carried out the same study except, instead of ranking, they provided the rating scores on a different proposed database. Although image retargeting, Gammut Mapping, Dehazing, and Tone Mapping have no direct relation with CE, the main purpose of discussing these works in this paper is to provide information about different subjective experiments performed with the same goal, i.e., performance evaluation of objective quality measures designed for diverse applications. To summarize the related works carried to date, we provide, in Table 1, our own perspective on the main contributions made in this field of research.

It is worth noting, that our methodology differs from previous works in many aspects; (1) The objectives are not the same. We aim here to analyze the performance of CEE measures in contrast to the work in [19,21,24] in which the performance analysis of classical IQA measures (i.e., IQA for distortions) was discussed, (2), The database is not the same compared with classical IQA databases like TID2013 [18], CSIQ [30] containing contrast images and few existing contrast databases [19,21]. These databases contain simulated changes of global contrast using a simple pixel value mapping function so as to produce a decrease in contrast. The authors consider these transformations as contrast distortion. Whereas, in our framework, we deal with the artifacts and distortion that may happen when applying CE operations. The distortion that might appear in the enhanced images after processing with CE methods are, for example, color saturation, color loss, blocking and ringing amplification in the case of compressed images, noise amplification, and halo effects and some others. The common databases did not contain any of these after effects due to CE. Moreover, in our case, we use different representative CE methods. (3), In contrast of all the databases, we do not want to estimate distortion in terms of decrease in quality like in classical IQA, rather our goal is to assess and quantify, subjectively and objectively, the increase in quality. (4), The application is entirely different compared to the CEE of tone mapped and retargeted images [23,25,26].

To the best of our knowledge, there are only two dedicated databases related to contrast manipulation [19,21], where the processed images are obtained using simple artificial pixel-based transformations. Whereas, in our proposed database, some realistic CE artifacts are considered and provided with subjective ranking of different CE methods, which can be used to validate the performance of new CEE measures. The proposed database will help in preliminary validation of new image CEE measures without performing dedicated subjective experiments. The main contributions of this work are:

- To provide a comprehensive performance analysis of the state-of-the-art CEE measures in terms of correlation with the subjective evaluation provided in the developed database as well as on other existing contrast manipulated databases.

Table 1
Summary of subjective experiments used in different image contrast databases.

	[23] (2005)	[25] (2010)	[26] (2012)	[19] (2012)	[31] (2013)	[27] (2014)	[28] (2014)	[24] (2015)	Ours (2016)
Original images	23	37	57	26	15	30	23	8	30
Processed images	–	–	171	78	400	300	–	480	180
Image resolution	–	–	–	512 × 512	720 × 576	–	800 × 800	1600 × 1200	512 × 512
Subjective method[a]	PC	PC	ACR	–	SS	PC	–	ACR-DR	PC
Score type	Rank	Rank	MOS	DMOS	MOS	Rank	9 scales catg rating	–	Rank
Screen resolution	–	–	1920 × 1280	1920 × 1200	–	–	1920 × 1080	LCD 24″	1920 × 1200
Observers	48	210	30	9	22	–	17	188	23
Viewing distance[b]	–	Web	–	–	$3H_I$	–	50 cm, 100 cm	80 cm	$2H_S$
Processing methods	6	8	10	–	-	–	–	–	6
Evaluation measures	NIL	6	6	–	-	–	-	–	12
Application[c]	TM	RT	RT	CE	CE	DH	GM	GM	CE

[a] SS—Single Stimulus, ACR (Absolute Category Rating), DR (Dynamic Reference).
[b] H_I (Image height), H_S (Screen height), Web (through web based interface).
[c] TM (Tone Mapping), RT (Retargeting), CE (Contrast Enhancement), DH (Dehazing), GM (Gamut Mapping).

- To evaluate six representative CE methods on a set of images representing different kinds of visual content. Here, our objective is to analyze the performance of CEE measures rather than CE methods. We focused only on some representative CE methods.
- To provide a comprehensive statistical analysis of the data collected from subjective experiments on a new and unique CE dedicated database.
- To propose a multi-metric fusion to improve the correlation performance with the subjective ranking.

The rest of the paper is structured as follows: Section 2 provides details of objective evaluation of CE methods, followed by subjective evaluation, which is discussed in Section 3. The particulars of the experiments and the newly developed database are provided in Section 4. The statistical analysis used to process the raw subjective data is discussed in Section 5. The experimental results are discussed in Section 6. Finally, the paper is concluded in Section 7.

2. Objective evaluation of contrast enhancement methods

Contrast enhancement is one of the most widely used low-level processing tasks. It generally consists of modifying the observed image so as to increase the visibility of details and structural information without amplifying noise or any other undesirable effects.

2.1. Contrast enhancement methods

Contrast enhancement methods can be grouped into two broad categories: direct approaches and indirect approaches [32]. The first category refers to the methods where the process is directly applied to the contrast itself; the contrast is supposed to be defined; whereas in the indirect approaches, the process implicitly affects the contrast by transforming some local or global features of the image; this is the case, for example, in histogram-based methods. Furthermore, CE can be performed in the spatial domain or the spatial-frequency domain. It is worth noting that CE is highly application-oriented and the way humans perceive the quality of images depends strongly on the semantic content of the image and the application of interest. As an example, human beings are more sensitive to the effects of CE on portrait images than on natural texture pictures or images of materials or cells observed through a microscope. This is due to high-level tasks related to the learning process and human interest.

In the literature, we can find numerous methods designed for different applications. In our work, we selected six CE methods as a representative set of the most common approaches used in the literature. These methods are: Adaptive Edge Based Contrast Enhancement (AEBCE) [32], Contrast Limited Adaptive Histogram Equalization (CLAHE) [33], Discrete Cosine Transform based (DCT) [34], Global Histogram Equalization (GHE) [35], Top Hat Transformation based (TOPHAT) [36], and Multi-scale Retinex (MRETINEX) [37]. The above were selected to cover the different classes of CE including histogram-based, edge-based, transform-based, morphological-based, and HVS-inspired methods [32–37]. We have used the codes for some methods accessible from the original papers author's websites. For GHE and CLAHE, we have used the MATLAB built-in functions `histeq` and `adapthisteq` respectively. Since, the main goal of the study is the performance comparison of CEE measures instead of CE methods, therefore, for our experiments, we have used the CE algorithms with their default parameters without tuning the algorithms for performance optimization.

2.2. Contrast enhancement evaluation measures

The improvement in image quality after CE can be evaluated using a multitude of objective measures. Although, we can see a lot of research efforts towards the development of CE algorithms, the objective CEE measures are limited and specific to different applications. The CE evaluation is different from conventional IQA. The reason is that in conventional IQA, the image which is similar to the original is considered as of good quality, and the similarity decreases with the increase in degradation. It is worth noticing that when using classical IQA, like SSIM, which is FR metric for quality assessment of degraded images, its value is close to one, when there is no distortion in an image and its value is less than one in the case of a degraded image. However, in the case of CE, we start from an input image and try to improve its quality. This processing is expected to produce more visible structures and the obtained images are rather different from the original one. If we use the SSIM for the contrast enhanced image, it will give value less than one, which does not correspond to an image of good quality. It has been observed that only the VIF measure [20], which is based on classical FR-IQA approach, yields interesting results. Indeed, the VIF produces a value less than one for the degraded images and greater than one for the case of enhancement.

Vu et al. [19], proposed that, to assess the quality of enhanced images, one can use the given image (enhanced image) as the reference and the original image as the distorted one and apply conventional IQAs. Whereas, in the case of NR-IQA, the CEE measures are derived from the given image. Some measures like sharpness, blurriness, Singular Value Decomposition based measures, details visibility map, after CE could be used to derive NR-CE quality measures. Recently, Fang et al. [22] used natural scene statistics to quantify the quality of contrast-enhanced images by looking the enhancement process as degradation process and apply the conventional IQA for the contrast distorted images in classical IQA databases. But in general, it is not applicable to CE applications.

In this section, we provide a brief overview of the measures used in our study. For the sake of completeness, we also provide the mathematical expressions for the measures as well. Based on the availability of the original image, we can group these measures into two broad classes,

Table 2
Notations used for CEE measures.

Notation	Description
I_r	Original image
I_e	Enhanced image
H	Image height (rows)
W	Image width (columns)
b	Block size
i, j	Pixel indices
L	Number of gray levels
I_{ij}	Image pixel value at index (i, j)
$\bar{I}$	Mean pixel value in an image
c	Constant ($c = 0.0001$) to avoid division by zero
$E(.)$	Statistical expectation
B_1, B_2	Number of blocks along rows and columns
I^c	Color channel in an RGB image (red, green, blue)
$h(.)$	Image histogram
$p(.)$	Probability mass function
I_{ij}^{max}	Maximum pixel intensity within the block (i, j)
I_{ij}^{min}	Minimum pixelintensity within the block (i, j)
I_{ij}^{cen}	Center pixel intensity within the block (i, j)
n	Pixel neighborhood index
$\bar{I}_b$	Average pixel value within the block centered at index (i, j)

i.e., Full Reference (FR) and No Reference (NR) measures (see Table 3). Moreover, based on the methodology used, we have also categorized these measures into Statistics-based, Gradient/Energy-based, and HVS-inspired CE evaluations. These measures are usually derived from grayscale images. For color images, the luminance component is used for contrast assessment. In this work, we adopt some state-of-the-art measures and our aim is to investigate how well these measures are consistent with the human judgment of quality. These measures are computed using only the luminance component of images. The different parameters used for the CEE measures are also mentioned in Table 3. To be consistent with the use of variables in the mathematical expressions of CEE measures as well as in the rest of the paper, we list the description of each variable in Table 2. In the following, we start with a brief description of each category of CEE measures.

2.2.1. Statistics based CEE measures

Absolute Mean Brightness Error (AMBE): It is used to evaluate how much original brightness is preserved in the enhanced image [38]. It is calculated as the deviation of the mean intensity of the enhanced image from that of the original image. For CE, it is desirable that the original brightness of an image is to be preserved. The lower value of AMBE means that the enhanced image has good brightness preservation. Here, brightness preservation does not mean that the image natural look (quality) also preserves. Either a very low value or the highest value of AMBE also indicates poor performance in case of CE.

Root Mean Square Contrast (RMSC): It is a pixel-based NR metric [39]. It gives high values for the images containing a major bright portion (e.g., sky, sea, etc.). High values of RMSC correspond to image with better contrast. However, it is not considered as an effective measure of CE since its value also increases with the appearance of some undesirable artifacts and noise amplification.

Reduced-reference Image Quality Metric for Contrast change (RIQMC): It is a Reduced Reference (RR) metric used to quantify image contrast and naturalness [21]. It combines the entropy of phase congruency image, and four statistical features computed from image histogram (i.e., mean, variance, skewness, and kurtosis). The first order statistical feature, F_1, is computed by penalizing very large and very small mean values using the Gaussian kernel and is defined as follows:

$$F_1 = \exp\left[-\left(\frac{E(I_e) - \mu}{\beta}\right)^2\right] \tag{1}$$

where μ and β determine the mean and shape of the Gaussian kernel.

The context-free contrast feature, F_2, defined as a function of variance computed from the image histogram is expressed as follows:

$$F_2 = E[h(I_e)^2] - E[h(I_e)]^2 \tag{2}$$

where $h(I_e)$ represents the histogram of an enhanced image.

Table 3
The expressions of the CEE measures used in our experiments.

CEE measures expressions	Type	Parameters used
Statistics-based		
AMBE [38] $= \lvert E(I_r) - E(I_e) \rvert$	FR	–
VIF [20] $= \frac{\sum_{k \in subbands} I_M(\vec{C}^{N,k} \mid \vec{F}^{N,k})}{\sum_{k \in subbands} I_M(\vec{C}^{N,k} \mid \vec{E}^{N,k})}$	FR	–
RMSC [39] $= \sqrt{\frac{1}{MN-1} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(I_{ij} - \bar{I}\right)^2}$	NR	–
DE [40] $= -\sum_{x=0}^{255} p(x) \log_2 p(x)$	NR	–
Gradient/Energy-based		
EC [46] $= \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \lvert \Delta I(i,j) \rvert$	NR	–
RSE [47] $= \log\left(\frac{1}{\omega_{max}} \sum_{\omega} \lvert E_{R_r}(\omega) - E_{R_e}(\omega) \rvert\right)$	NR	$\omega = 1$
IEM [45] $= \frac{\sum_{i=1}^{B_1} \sum_{j=1}^{B_2} \sum_{n=1}^{8} \lvert I_{ij}^{e,c} - I_{ij}^{e,n} \rvert}{\sum_{i=1}^{B_1} \sum_{j=1}^{B_2} \sum_{n=1}^{8} \lvert I_{ij}^{r,c} - I_{ij}^{r,n} \rvert}$	FR	$b = 3$ (non-overlapping blocks 3×3)
HVS-Inspired		
EME [14] $= \frac{1}{B_1 \times B_2} \sum_{i=1}^{B_1} \sum_{j=1}^{B_2} 20 \ln\left(\frac{I_{ij}^{max}}{I_{ij}^{min} + c}\right)$	NR	$b = 8$ (non-overlapping blocks 8×8)
EMEE [15] $= \frac{1}{B_1 \times B_2} \sum_{i=1}^{B_1} \sum_{j=1}^{B_2} \alpha \left(\frac{I_{ij}^{max}}{I_{ij}^{min} + c}\right)^{\alpha} \ln\left(\frac{I_{ij}^{max}}{I_{ij}^{min} + c}\right)$	NR	$b = 8$ (non-overlapping blocks 8×8), $\alpha = 1$
AME [16] $= \frac{-1}{B_1 \times B_2} \sum_{i=1}^{B_1} \sum_{j=1}^{B_2} 20 \ln\left(\frac{I_{ij}^{max} - I_{ij}^{min}}{I_{ij}^{max} + I_{ij}^{min}}\right)$	NR	$b = 8$ (non-overlapping blocks 8×8)
AMEE [16] $= -\frac{1}{B_1 \times B_2} \sum_{i=1}^{B_1} \sum_{j=1}^{B_2} \alpha \left(\frac{I_{ij}^{max} - I_{ij}^{min}}{I_{ij}^{max} + I_{ij}^{min}}\right)^{\alpha} \ln\left(\frac{I_{ij}^{max} - I_{ij}^{min}}{I_{ij}^{max} + I_{ij}^{min}}\right)$	NR	$b = 8$ (non-overlapping blocks 8×8), $\alpha = 1$
SDME [17] $= \frac{-1}{B_1 \times B_2} \sum_{i=1}^{B_1} \sum_{j=1}^{B_2} 20 \ln \left\lvert \frac{I_{ij}^{max} - 2I_{ij}^{cen} + I_{ij}^{min}}{I_{ij}^{max} + 2I_{ij}^{cen} + I_{ij}^{min}} \right\rvert$	NR	$b = 5$ (non-overlapping blocks 5×5)
RME [15] $= \frac{1}{B_1 \times B_2} \sum_{i=1}^{B_1} \sum_{j=1}^{B_2} \left\lvert \frac{\log \lvert I_{ij}^{cen} - \bar{I}_b \rvert}{\log \lvert I_{ij}^{cen} + \bar{I}_b \rvert} \right\rvert$	NR	$b = 5$ (non-overlapping blocks 5×5)

–E_{R_r} and E_{R_e} represents radial spectral energy of reference and enhanced image respectively, ω is the radial frequency.
– $I_M(\vec{C}^{N,k} \mid \vec{F}^{N,k})$ and $I_M(\vec{C}^{N,k} \mid \vec{E}^{N,k})$, represents the mutual information that can be extracted from a particular wavelet subband k in the original (F) and test (E) images respectively, C represents the wavelet coefficients.

Similarly, the higher-order statistical features, i.e., skewness, F_3 and kurtosis, F_4, are computed as follows:

$$F_3 = \frac{E[I_e - E(I_e)]^3}{\sigma^3(I_e)} \tag{3}$$

$$F_4 = \frac{E[I_e - E(I_e)]^4}{\sigma^4(I_e)} - 3 \tag{4}$$

where σ is the standard deviation of the gray-levels in the image.

The similarity feature, F_5, defined as the difference between the entropy of phase congruency of enhanced image and original image:

$$F_5 = H_{PC}(I_e) - H_{PC}(I_r) \tag{5}$$

where $H_{PC}(.)$ represents the entropy of phase congruency image.

Finally, the RIQMC is computed as a linear combination of the five features using the following expression.

$$\mathrm{RIQMC} = \sum_{i=1}^{5} w_i F_i \tag{6}$$

where the weights, w_i, for $i = 0, 1, \ldots, 4$, represent the contribution of each feature in the final contrast metric.

The RIQMC fusion depends on the weights, but they are not provided in the text. Through experiments, we have observed that with increasing the contrast (improvement in quality), the RIQMC value decreases.

Visual Information Fidelity (VIF): It is a FR quality metric used to quantify the loss of original image information due to processing or transmission of the given image. The original and test images are decomposed into different subbands and the mutual information to be perceived by HVS from these subbands is calculated for both images. The measure is expressed as the fraction of original image information that can be perceived by HVS from the test image. VIF measures can be used as a quality metric for both degraded and enhanced images [20]. Its values are equal to, less than, and greater than one for the original, degraded, and enhanced images respectively.

Discrete Entropy (DE): It measures the amount of information or randomness of gray-levels in an image [40]. It is well known that the increase in contrast highlights the subtle details in an image and results in an increase in entropy value. It is a global measure based on the overall histogram of an image and fails to consider the local details and spatial correlations among the pixels. The higher values of DE correspond to image with more details visibility amplification and is considered as image with good quality.

Mutual Information based Contrast Measure (MICM): It is a NR metric used to quantify the global image contrast and to detect and control the side effects of CE in few neighborhood-based methods [41,42]. It is based on mutual information derived from the joint probability mass function of a gray level co-occurrence matrix and is given by:

$$\mathrm{MICM} = \sum_{i=0}^{L}\sum_{j=1}^{L} p_{ij}\log_2\left(\frac{p_{ij}}{p_x(i)p_y(i)}\right) \tag{7}$$

where p_{ij} is the joint probability mass function of the luminance channel, whereas p_x and p_y represent the marginal probabilities calculated along the rows and columns of co-occurrence matrix respectively. It is better than 1st order entropy, we take into consideration the spatial correlation among the pixels using the gray level co-occurrence matrix. It is simple to compute, however, it does not provide information about image unnaturalness.

Lightness Order Error (LOE): The enhancement methods should be designed to increase the contrast and preserve structural information as well as to result in natural looking images. However, naturalness, in general, is a very subjective quantity. It is often considered by the observers when evaluating overall image quality. Naturalness is typically defined as the degree of correspondence between the visual representation of the image and the knowledge of reality (colors of familiar objects) as stored in memory [43]. Wang et al. [44] proposed an objective metric to measure the naturalness preservation in the enhanced imaged based on estimating the lightness order error between the original image and enhanced image. The lightness of an RGB image, I^c, is obtained by taking the maximum of the three color components.

It measures the naturalness preservation in the enhanced image based on estimating the lightness order error between the original image and enhanced image [44]. The lightness of an RGB image, I^c, is obtained by taking the maximum of the three color components.

$$L_{ij} = \max_{c\in\{r,g,b\}} I_{ij}^c \tag{8}$$

where r, g, and b represent the red, green, and blue color components in an RGB image.

The relative order difference of the lightness between the original image and its enhanced version is calculated as follows:

$$RD_{ij} = \sum_{x=1}^{M}\sum_{y=1}^{N}(U(L_{ij}^r, L_{xy}^r) \oplus U(L_{ij}^e, L_{xy}^e)) \tag{9}$$

$$U(a,b) = \begin{cases} 1 & \text{for } a \geq b \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $U(a, b)$ is a unit step function and $\oplus$ is exclusive-or operator.

The final LOE measure is calculated as:

$$LOE = \frac{1}{(M \times N)}\sum_{i=1}^{M}\sum_{j=1}^{N} RD_{ij}. \tag{11}$$

In the original implementation, the downsampled version of both images was used to reduce the computational complexity with the downsampling ratio of $r = 50/\min(H, W)$.

Since the relative order of lightness represents the light source directions and the brightness variations, the naturalness of an enhanced image is related to the relative order of lightness in different local areas. Small values of LOE indicate that the naturalness is well preserved in an enhanced image in comparison with the original image. The authors claimed that metric well provides naturalness information. However, naturalness is a complex property and is difficult to define.

2.2.2. Gradient/energy-based CEE measures

The following measures are based on either local signal activity or energy as measured through the gradient operator or the spectral energy distribution. Indeed, any increase/decrease of the contrast inevitably affects the pixel intensity gradient and the spectral energy distribution in the spatial-frequency domain.

Image Enhancement Metric (IEM): It is a FR metric proposed by Jaya et al. [45] and is calculated by subdividing an image into non-overlapping blocks. The ratio of the sum of absolute values of differences of the center pixel from its eight neighbors in all blocks of the enhanced image and the corresponding blocks in the original image represents the IEM value. The absolute intensity differences between a pixel and its neighbors corresponding to the reference and enhanced images are used to account for the change in contrast and sharpness. Typical values for the image blocks are 3×3 or 5×5. For identical images, IEM is equal to one. The values of IEM greater than one means image contrast and sharpness are increased.

Edge Content (EC): It is a blind objective measure based on the local gradient of the image intensity [46]. In its expression, $\Delta I\,(.)$ represents the gradient magnitude of the pixel value computed from the Sobel edge operator. Higher values of EC correspond to images with more contrast. The overall EC value for a complete image is calculated by averaging the local EC values for each block.

Radial Spectral Energy (RSE): It is based on radial spectral energy analysis developed for blind image sharpness assessment [47]. It is based on the idea that the effect of adding a certain amount of blur to a given image depends on the original quality of this image. In other words, a contrasted image is more sensitive to blur effect than a less contrasted image. The enhancement measure is computed as the variation of the radial spectrum due to contrast enhancement. The radial energy on the original image and its contrasted version are calculated as follows:

$$E_{R_r}(\omega) = \frac{1}{K}\sum_k |\mathcal{I}_r(\omega, \theta_k)| \quad (12)$$

$$E_{R_e}(\omega) = \frac{1}{K}\sum_k |\mathcal{I}_e(\omega, \theta_k)| \quad (13)$$

where $\mathcal{I}(.)$ is the Fourier transform of the image signal $I(.)$ at a particular radial frequency ω and in θ_k direction, $\theta_k = \frac{k\pi}{K}$ and $\omega = \sqrt{(u^2+v^2)}$ where u, v are spatial frequencies, and K is total number of directions.

Then the blur index is computed as:

$$\mathrm{RSE} = \log\Big(\frac{1}{\omega_{max}}\sum_\omega |E_{R_r}(\omega) - E_{R_e}(\omega)|\Big) \quad (14)$$

where ω_{max} is the maximum radial frequency within the image and can be calculated as $\omega_{max} = \sqrt{u_{max}^2 + v_{max}^2}$, where u_{max} and v_{max} are the maximum values of spatial frequencies u and v. The $\log(.)$ is used in the expression to make the measure non-linear in accordance with HVS response. An increase of RSE corresponds to increase in contrast.

2.2.3. HVS-inspired CEE measures

Some simple CEE measures have been proposed in [14–17,48]. These approaches are not inspired by the traditional IQA measures as suggested by [19]. The proposed measures are based on the computation of a global index derived from some local measures related to contrast and gradient. These CEE measures are mainly inspired by the Michelson and Weber–Fechner contrast measures which are not really adapted to natural scenes. These measures have been evaluated on a limited set of images processed by some CE methods. However, these studies do not provide a comprehensive analysis of the validity of these measures on various images and different CE methods. The main CEE measures of this class are now briefly discussed.

Measure of Enhancement (EME): The EME was proposed by Agaian et al. [14] and is a NR metric based on a contrast measure using the pixel value dynamic range (min–max values) within a block. The image is first divided into non-overlapping blocks of the same size (say 8×8). The EME value is computed based on the minimum and maximum pixel values in each block, respectively. The overall measure is computed by averaging the local EME values for image blocks. Since log of ratios of maximum and minimum intensities within each block can be written as difference, EME may represent signal dynamic range of the image. EME increases with the increase in image contrast.

Measure of Enhancement by Entropy (EMEE): The EMEE measures the entropy in the local contrast as defined in [15]. It also increases with the increase in image contrast. The use of entropy is motivated by the fact that any small variation in the contrast would convey additional amount of information on the spatial content of the image. This consequently would affect the entropy value.

Absolute Measure of Enhancement (AME): Similarly to EME, the AME [16], is also a block-based logarithmic Michelson contrast based measure. The AME decreases with the increase in image contrast.

Absolute Measure of Enhancement by Entropy (AMEE): The AMEE measures the entropy in the local Michelson contrast of an image as defined in [16]. It increases with the increase in contrast. The reason for using the entropy is also the same as for EMEE measure.

Second Derivative like MEasurement (SDME): This measure is based on the fact that the local contrast is highly related to the local variations of the signal [17]. This could be captured by any derivative operator, here, a pseudo-second order derivative operator is used. It is also a block-based measure with default block size either 3×3 or 5×5. The authors claimed that this measure is less noise sensitive than the other similar measures based on only min–max operations. It decreases with an increase in image contrast.

Root Mean Enhancement (RME): It incorporates both RMS contrast and properties of HVS [15]. It measures the relative RMS contrast in the log domain. It is calculated by subdividing an image into non-overlapping blocks (say 3×3 or 5×5). For low contrast images, RME value is small, whereas it is large for high contrast images.

It is worth noting that the goal of CEE measures should not be limited to quantify improvement in quality but also to provide a quantitative measure that could be used to control some unpredictable after-effects due to CE. The side effects due to CE are color mismatch, color bleeding, saturation, overshooting, halo effects, blocking/ringing artifacts amplification, and other undesirable effects. The existing CEE measures either increase or decrease with the increase in contrast and none of the measures could predict the side effects due to CE. It has been demonstrated in [41] that AMBE, EC, AME, EME measures do not provide information about the optimal contrast value. In [41], a mutual information based measure computed from gray level co-occurrence matrix was proposed. A sharp decay in the proposed measure curve was observed after reaching a certain saturation point (this point could be considered as the optimal contrast).

3. Subjective evaluation of contrast enhancement methods

Subjective methods involve human judgment of perceived quality and they are considered as more reliable methods for quality assessment applications. The detailed explanations of testing material, testing environment, testing methods, and statistical analysis of subjective data are discussed in [49,50].

The subjective experiments are direct methods, and different subjects rate the quality of a given image. These methods need careful design considerations and are performed in a well-controlled environment and involve at least 15 observers. They are time-consuming and cannot be used in real-time applications. However, they can only be used in benchmarking of different objective image quality evaluation measures. The subjective experiments can be grouped into rating-based and ranking-based methods [51]. In rating-based methods, participants assign a score to each stimulus presented to them either on an interval scale or categorical scale. The rating-based methods are further classified into three types: the Single-Stimulus (SS), Double-Stimulus (DS), and Multi-Stimulus methods. In SS methods, the observers only rate on quality of a test image/video. Whereas in DS approaches, the observers are asked to rate the change in quality between two images/videos based on the perceived quality while compared to the original image/video. Each methodology has some strengths and weaknesses [52]. The DS approach is not too much affected by the context. The ratings are less influenced by the severity and ordering of the degradations within the test session. The SS method is more sensitive to the context. However, it is faster compared with the DS method and provides more representative quality scores [53].

The SS subjective methods are further categorized into Absolute Category Rating (ACR) [50], Absolute Category Rating with Hidden Reference (ACR-HR), and Single Stimulus Continuous Quality Rating (SSCQE) [49] methods. In the ACR subjective method, the test stimuli are rated individually on a 5-level categorical rating scale (excellent, good, fair, poor, and bad) with values 5, 4, 3, 2, and 1 respectively for MOS calculations. The ACR methods are strongly affected by the observers' opinion of the content. The observer may give a poor rating to the disliked content. The ACR-HR is a variant of ACR method, in which the original image/video is shown with the test stimulus. The observers are unaware of the original image/video and they are asked to rate the quality of both sequences. The final score for the test stimulus is calculated by subtracting the ACR ratings of the test stimulus from the reference stimulus. The ACR-HR method is less content sensitive compared to the ACR method. The SSCQE method is only used for monitoring applications, e.g., continuous quality evaluation of videos using a slider. The quality ratings are recorded at regular intervals (usually every half second) and a quality curve is generated for a specific time period.

In addition, DS subjective methods are classified as Double-Stimulus Continuous Quality Scale (DSCQS) [49] and Double Stimulus Impairment Scale (DSIS) [49] or DCR (Degradation Category Rating) [50] methods. In DSCQS methods, both original and test stimuli are shown in random order. The observers are unaware of the original image/video and they are asked to rate the quality of both on a continuous scale labeled with the ACR categories.

The DSIS and DCR, both refer to the same method where both reference and test stimuli are shown one after the other and the observer knows about the reference. The observers are asked to rate the difference in quality on a discrete 5-point impairment scale (Imperceptible (5), perceptible but not annoying (4), slightly annoying (3), annoying (2), and very annoying (1)). The DCR method is less influenced by the observer's rating of the content.

Subjective Assessment of Multimedia Video quality (SAMVIQ) [54,55], is a non-interactive multi-stimulus subjective assessment method used for video or audio-visual quality evaluation. In SAMVIQ test, observers are allowed freely to view stimuli multiple times using a fine resolution continuous quality rating scale (0–100). In SAMVIQ, there is a possibility to change the vote before proceeding. Between the category and interval rating based methods, the main problem in interval (continuous) rating scales is that people may have their own perceptual judgment scales in their mind and it is tough to rate if the number of points on the scale is large.

The ranking-based methods can be grouped into rank order-based methods and Pairwise Comparison (PC)-based methods [50]. In rank-order-based methods, the observers are asked to rank different stimuli according to perceived quality displayed at once. This protocol seems to be time-efficient; however, it is sometimes difficult to differentiate among the stimuli, especially when the number of stimuli are more than three or four having subtle differences. Whereas, in PC-based methods, the stimuli are presented to the observers in pairs, and the observers choose whether the stimulus A is better than stimulus B and vice versa, or both stimuli are alike. In this way, each stimulus is compared with the other. The PC-based methods are simple as only one stimulus is compared with the other, and they are effective when there are subtle differences between the stimuli compared to the rank-order or rating-based methods. More research on the PC-based subjective methods associated with objective quality metrics has been carried in [56–59].

The pairwise ranking raw data can be statistically analyzed in terms of coefficients of transitivity and consistency to sort out the bad participants as well as pathological stimuli [60]. Moreover, the pairwise ranking data can also be easily converted to the rating scores. The PC ranking data can be further extended for more stimuli and images. However, since each stimulus is compared in pairs with the others, the number of comparisons increases with the stimulus. For M stimuli (or methods in our case), the maximum number of pairwise comparisons is $\binom{M}{2} = \frac{M(M-1)}{2}$. For a large number of comparisons, some procedures exist to reduce the number of comparisons [61]. An overview of different subjective methodologies used in IQA applications is shown in Fig. 1. Note that in both rating and ranking based subjective experiments, the aggregated scores from all observers are considered as the overall ratings or ranking scores for each image.

The subjective methodologies discussed above have been used in various applications [19,23–29,31,62]. Considering the advantages of PC-based methods, we opted to use the non-forced choice pairwise ranking protocol in our subjective experiments. In the following section, we provide a detailed description of our new database containing images obtained using six CE methods as well as a brief discussion on the testing methodologies and testing environment.

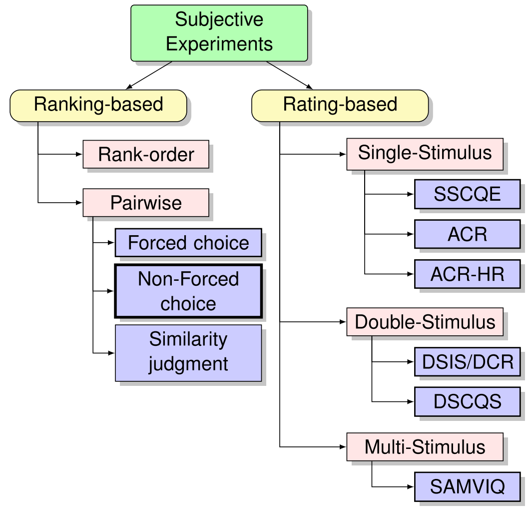


Fig. 1. Classes of subjective methodologies used for quality assessment.

4. Experimental framework and database description

For performance analysis of different CEE measures, we performed an intensive psychophysical experiment. The correlation between the subjective ranking and the objective scores should indicate how much a given CE evaluation, is consistent with the human judgment of quality. We have followed the main relevant ITU guidelines designed for the subjective experiments [49,50]. Here, we provide a brief discussion of the new database, the testing environment, and the testing procedure used in our experiments.

4.1. Database creation

We constructed a new database named as Contrast Enhancement Evaluation Database (CEED2016), containing 30 original color images and 180 enhanced images with a size of 512 $\times$ 512 pixels. The database is built with our own captured images and some common pictures used by the image processing community. The reason for using the old images is that these images are widely used in the research related to the contrast enhancement and contrast enhancement evaluation. For consistency, we opted to include some of these images in our proposed database. The images in the proposed database are shown in Fig. 3.

It is well-understood that the human perception of image quality is highly dependent upon the scene content under observation. For this reason, we selected images with different textures, color distributions, and contrast variations. We have used three quantitative measures for the selection of images. These measures are Colorfulness (CF) [63], Spatial Information (SI) [63], and Global Contrast Factor (GCF) [64]. A brief description of each measure is given below:

Colorfulness (CF): It is a perceptual indicator of the variety and intensity of colors in the image [63]. The Red (R), Green (G), and Blue (B) color components are converted into opponent color space as follows:

$$r_g = R - G \tag{15}$$

$$y_b = \frac{(R+G)}{2} - B. \tag{16}$$

The CF is then given by:

$$\mathrm{CF} = \sqrt{\sigma_{r_g}^2 + \sigma_{y_b}^2} + 0.3\sqrt{\mu_{r_g}^2 + \mu_{y_b}^2} \tag{17}$$

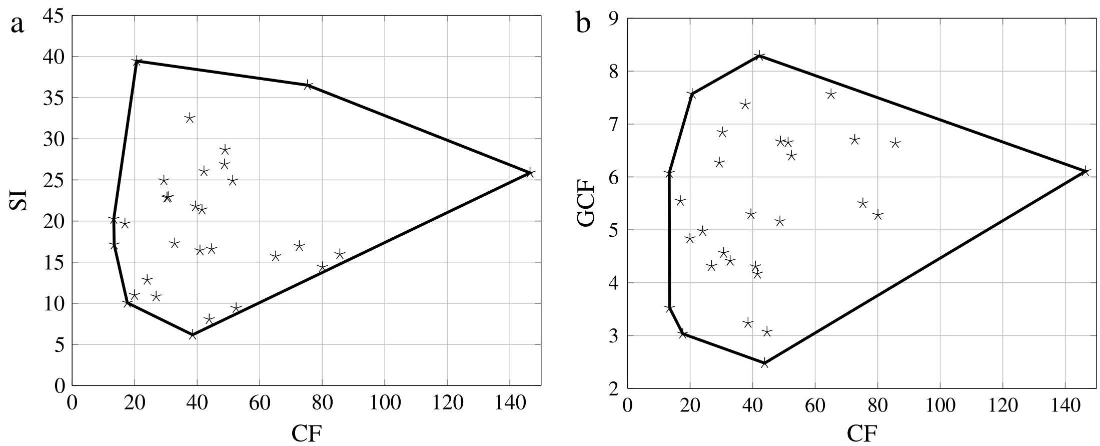


Fig. 2. Scatter plots for all images in the new database between (a) Spatial information versus colorfulness, (b) Global contrast factor versus colorfulness.

where σ_i and μ_i for $i \in [r_g, y_b]$ represent the standard deviations and the mean of the pixel values in the color space.

Spatial information (SI): It is an indicator of edge energy and is calculated as the root mean square of the edge magnitude over the entire image [62]:

$$SI = \sqrt{\frac{V}{1080}} \sqrt{\sum_{k=1}^{N} \left(\frac{\Delta_k^2}{N} \right)} \tag{18}$$

where Δ_k represents the gradient magnitude computed from the Sobel operator at the kth pixel, N is the total number of image pixels, and V is the vertical resolution of the image.

Global Contrast Factor (GCF): It is a global measure of the overall image contrast as perceived by the HVS. This contrast measure accounts for the multi-scale characteristics of the HVS. It is based on a multi-resolution decomposition scheme and a weighting process. The global contrast is then expressed as the weighted average of the local contrast computed at different resolution levels. The contrast weighting function is derived from a psychophysical experiment [64]. It is calculated as follows:

$$GCF = \sum_{i=k}^{N} w_k c_k \tag{19}$$

where w_k and c_k represents weights and average local contrast of the image for a given resolution and N is the number of resolution levels.

$$c_k = \frac{1}{W^k \times H^k} \sum_{i=1}^{W^k \times H^k} \frac{1}{l_{c_i^k}}, \text{ for } k = 1, \ldots, N \tag{20}$$

where $l_{c_i^k}$ represents local contrast for ith pixel at the kth resolution, W^k and H^k represents image width and height at kth resolution. The local contrast is computed as the average of the differences of pixel values with its four nearest neighbors.

Since the database contains images enhanced by different CE methods, and to see the effect of improvement in quality in a clearer way, we have used this measure to select images with varying contrast from low to high.

Using the measures above, we provided a scatter plot for the images in our database (Fig. 2). Here, '$\star$' symbol is used, to represent images. From the plots, it could be observed that the database contains images with diverse spatial information, colorfulness, and global contrast features.

The contrast-enhanced images were generated from the six selected CE methods mentioned in Section 2. Among the original images, we have also included six compressed images (three for JPEG and three for JPEG2000) with moderate compression so as to observe the effect of contrast enhancement that may increase the visibility of these artifacts. In this way, we can also observe the capabilities of different CEE metrics in quantifying these particular CE after effects i.e., blur, ringing amplification. In Fig. 4, we show some enhanced images with visible artifacts due to CE.

Table 4
Display setup used in the subjective experiment.

Parameter	Description
Type	LCD
Model	EIZO Color Edge CG242W
Screen	24.1 inch
Resolution	1920 × 1200 pixels
Calibration device	Eye-One Match 3
Color space	sRGB
Color temperature	6500 K
White point luminance	119 cd/m^2
Display frame rate	60 Hz
Contrast	80
Room environment	Dark
Gamma	2.2
Background color	Gray (128, 128, 128)

4.2. *Testing environment*

The subjective experiments were performed at Universit*é* Paris 13 at Laboratoire de Traitement et Transport de l'Information (L2TI). The images were displayed on a calibrated LCD monitor in a dark room environment to avoid any problem with the illumination adaptation of background. The details of display parameters are shown in Table 4.

Twenty-three observers, 10 experts and 13 non-experts, from different age groups and background participated in the experiments. Among these, there were fourteen male and nine female observers. The age distributions of the participants is also shown in Fig. 5.

All the observers had either normal vision or corrected to normal vision and they were undergone a pre-screening procedure for color vision and visual acuity. The observers were forced to perform the experiments at a fixed distance of twice the screen height which is equivalent in our case to be 4.5 times the image height of the image. The non-expert 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. They were allowed to give the same rank for both images in case of the equivalent degree of quality.

The database contains 30 original images. Each original image is enhanced by six CE methods, resulting in 180 enhanced images in addition to the original images. For each original image, the six enhanced

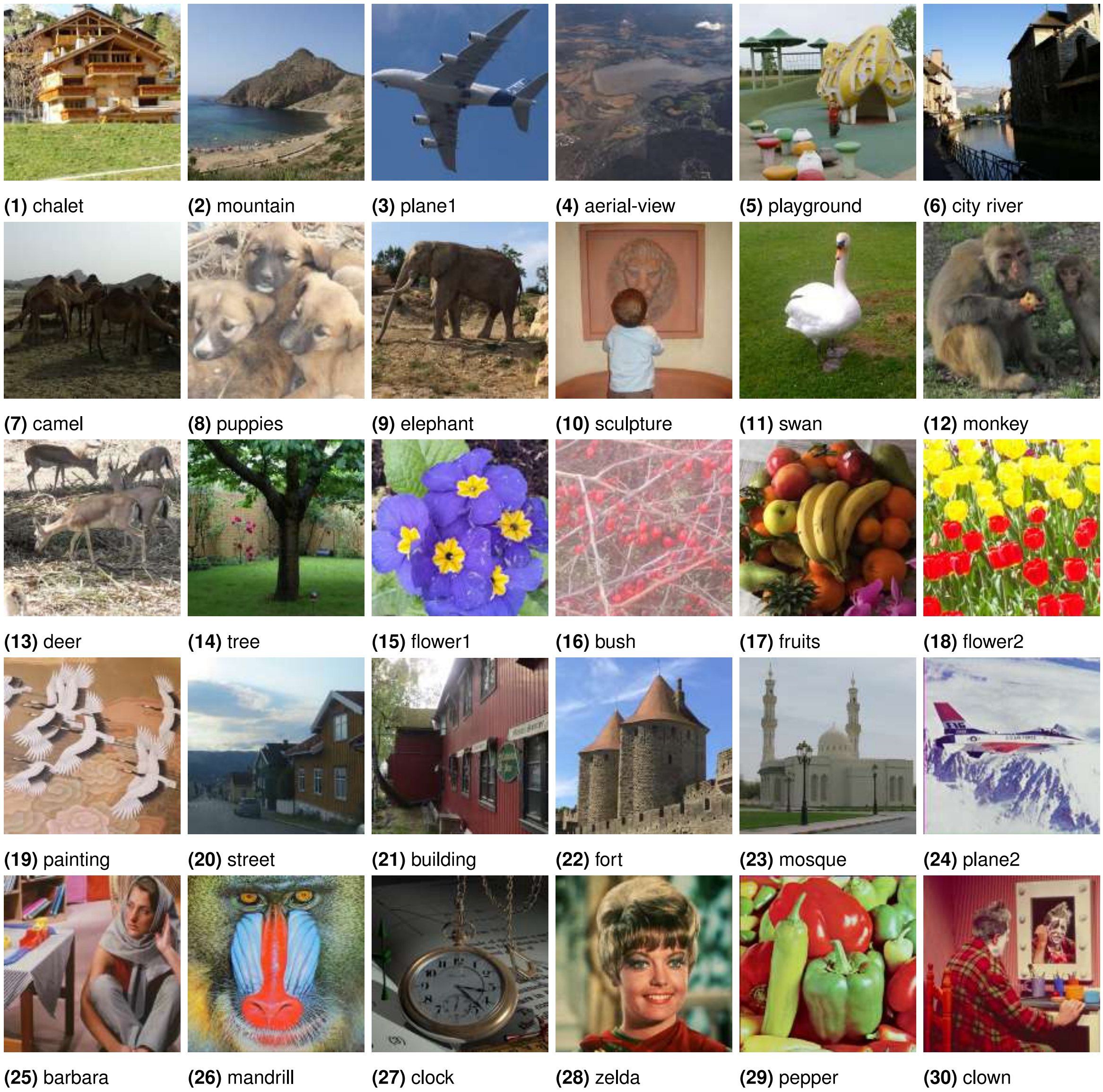


Fig. 3. Images in the database (Images 1–23 are self captured while images 24–30 are standard test images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

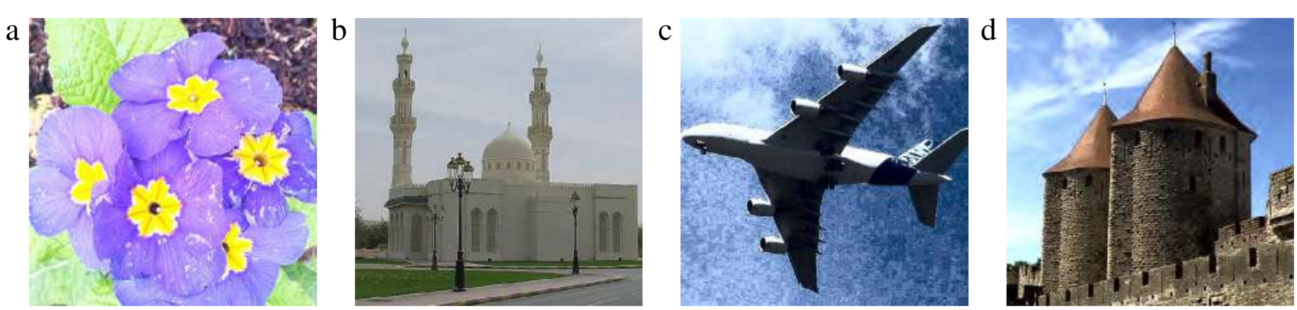


Fig. 4. Illustration of some artifacts due to CE (a) color shift, (b) halo effects, (c) blocking, and (d) ringing. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

images were shown in pairs to the observers. The number of possible combinations to display for each original image are $\binom{6}{2} = \frac{(6)(5)}{2} = 15$. We allowed the observers to take their time for the subjective experiments and they were not forced to finish early. However, they were informed that the whole subjective tests take in average 50–60 min.

4.3. Testing procedure

To obtain the ranking scores, we adopted a balanced pairwise preference based ranking protocol (Condorcet method). The interface for the subjective experiments was developed in Matlab, where, for each

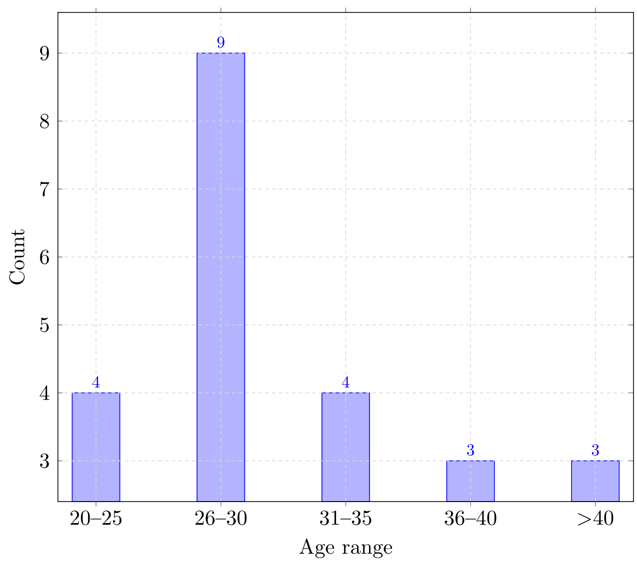


Fig. 5. Age distributions of observers participated in the subjective experiment.

original image, we randomly displayed all possible pair combinations of enhanced images to the observers. We also showed the original image in the center of the screen (a pair of enhanced images are to its left and right), to facilitate the analysis of after effects of CE. The observers had the choice to rank equally similar stimuli. A screenshot of the graphical interface is shown in Fig. 6. A Variability analysis of the results was carried to see the influence of the side of the image viewing, and the changes in the results were negligible. In the PC ranking protocol, each enhanced image is compared with the others in pairs and ranking results are stored in a preference matrix. An aggregated preference matrix for the 23rd image in the database is shown in Table 5. From Table 5, it can be observed that the CLAHE method is highly preferred, whereas GHE is least preferred by all the observers. The preference data was collected for all the images in our database for statistical and correlation analysis.

Table 5
A sample preference matrix for 23rd image (i.e., mosque) aggregated over preferences of 23 observers. In our experiment, M_1 = AEBCE, M_2 = CLAHE, M_3 = DCT, M_4 = GHE, M_5 = TOPHAT, M_6 = MRETINEX.

–	M_1	M_2	M_3	M_4	M_5	M_6	p_i
M_1	–	2.5	10	23	22	3	60.5
M_2	20.5	–	17.5	23	23	15.5	99.5
M_3	13	5.5	–	23	23	10	74.5
M_4	0	0	0	–	1	0	1
M_5	1	0	0	22	–	0	23
M_6	20	7.5	13	23	23	–	86.5

5. Statistical analysis

The data gathered from the subjective experiment was processed to verify its reliability and validity. The reliability relates to the consistency and it is further related to the closeness of agreement in the preference ranking among different observers (also called inter-rater reliability). Whereas, validity relates to the accuracy of the data. However, it does not mean that the data with high reliability is also accurate. For the preference based pairwise rank data, the reliability was measured using Kendall's Coefficient of Concordance (K_W) [60]. Other measures are Kendall's Tau (τ) and Spearman's Rank Order Correlation Coefficient (ρ).

5.1. Coefficient of agreement (u)

The coefficient of agreement or inter-rater reliability is a measure of understanding among a group of observers in their judgments. It is measured on a continuous scale in the range $[0 - 1]$. Kendall and Babington et al. [65] proposed coefficient of agreement, u, among the observers and defined it as:

$$u = \frac{2\sum_{i,j=1,i\neq j}^{M} \binom{p_{ij}}{2}}{\binom{S}{2}\binom{M}{2}} - 1 \tag{21}$$

where M is the number of CE methods, S is the number of observers, and p_{ij} represents the number of times image enhanced by method M_i is preferred over the image enhanced by method M_j. Its value is equal to one, when all the observers (or raters) agree on their preferences.

To test for the significance of coefficient of agreement (u), we have performed a chi-squared test (χ^2). The χ^2 values are calculated as

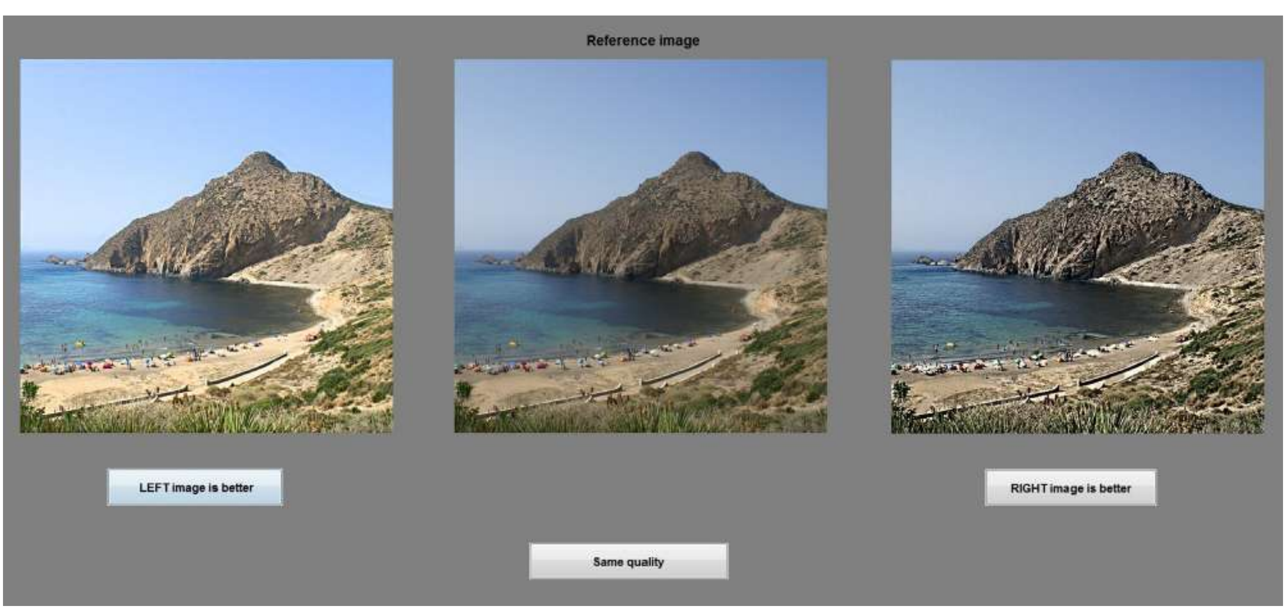


Fig. 6. The display environment where the original and enhanced images are displayed at the same time.

Table 6
Summary of statistical coefficients for the subjective experiment data.

Image	u	χ^2	p_{value}	$\bar{\zeta}_{observer}$	K_W	χ^2	p_{value}	Image	u	χ^2	p_{value}	$\bar{\zeta}_{observer}$	K_W	χ^2	p_{value}
1	0.67	235.83	<0.001	0.94	0.82	94.44	<0.001	16	0.24	92.70	<0.001	0.77	0.41	47.13	<0.001
2	0.38	141.30	<0.001	0.84	0.55	63.47	<0.001	17	0.29	109.43	<0.001	0.77	0.49	56.65	<0.001
3	0.55	196.91	<0.001	0.86	0.71	81.48	<0.001	18	0.39	145.30	<0.001	0.81	0.58	67.02	<0.001
4	0.47	170.70	<0.001	0.88	0.64	73.06	<0.001	19	0.36	133.57	<0.001	0.86	0.53	60.91	<0.001
5	0.57	202.91	<0.001	0.91	0.77	88.63	<0.001	20	0.35	131.04	<0.001	0.83	0.52	60.07	<0.001
6	0.12	54.00	<0.001	0.76	0.24	27.69	<0.001	21	0.41	150.96	<0.001	0.86	0.59	68.07	<0.001
7	0.45	164.74	<0.001	0.87	0.67	77.35	<0.001	22	0.37	135.57	<0.001	0.83	0.57	65.12	<0.001
8	0.20	82.35	<0.001	0.72	0.38	43.39	<0.001	23	0.48	173.52	<0.001	0.86	0.71	81.18	<0.001
9	0.46	167.13	<0.001	0.86	0.67	76.68	<0.001	24	0.50	181.09	<0.001	0.85	0.72	83.32	<0.001
10	0.58	205.35	<0.001	0.87	0.79	91.18	<0.001	25	0.37	135.87	<0.001	0.78	0.58	66.47	<0.001
11	0.52	188.09	<0.001	0.84	0.76	87.10	<0.001	26	0.26	101.57	<0.001	0.74	0.42	48.76	<0.001
12	0.19	76.78	<0.05	0.66	0.37	42.55	<0.001	27	0.37	135.52	<0.001	0.75	0.56	64.65	<0.001
13	0.47	169.74	<0.001	0.85	0.66	76.12	<0.001	28	0.39	145.17	<0.001	0.78	0.64	73.67	<0.001
14	0.35	131.26	<0.001	0.76	0.56	64.06	<0.001	29	0.40	147.91	<0.001	0.81	0.60	68.73	<0.001
15	0.44	159.30	<0.001	0.87	0.62	71.71	<0.001	30	0.31	117.83	<0.001	0.80	0.50	57.90	<0.001

Table 7
Consistency coefficients for 23 observers.

Observer	$\bar{\zeta}_{image}$	Observer	$\bar{\zeta}_{image}$	Observer	$\bar{\zeta}_{image}$
1	0.7583	9	0.9208	17	0.8406
2	0.7990	10	0.8354	18	0.8271
3	0.9042	11	0.7896	19	0.8094
4	0.6885	12	0.7708	20	0.8688
5	0.8688	13	0.8906	21	0.7917
6	0.7323	14	0.8354	22	0.7167
7	0.9208	15	0.8604	23	0.8344
8	0.8187	16	0.7406	–	–

follows:

$$\chi^2 = \frac{M(M-1)(1+u(S-1))}{2}. \tag{22}$$

The degree of freedom for this χ^2 statistic is selected as $\frac{M(M-1)}{2}$. The minimum value of u is $\frac{-1}{(S-1)}$ and $\frac{-1}{S}$ for even and odd number of observers respectively. For our experiment, with 23 observers, the minimum value of the consistency coefficient (u_{min}) is $\frac{-1}{23} = -0.0435$. The null hypothesis H_0 is rejected when the observed χ^2 is greater than its critical value.

5.2. Coefficient of consistency or transitivity (ζ)

The pairwise rank data is further assessed for inconsistency. It relates to the transitivity property in a paired comparison. It is determined from the number of intransitivity or circular triads in a set of ranking. It is also called intra-rater agreement and is calculated for each observer and image. The number of circular triads in a set of pairwise comparison can be calculated using the relation [65]:

$$\zeta = 1 - \frac{C}{C_{max}} \tag{23}$$

where C represents the number of circular triads and C_{max} is the maximum possible circular triads in a pairwise comparison. C is calculated using the following relation:

$$C = \frac{M}{24}(M^2-1) - \frac{1}{2}M \tag{24}$$

where $M = \sum(p_i - (M-1)/2)^2$, p_i is the number of times stimulus i was preferred over other stimuli. The maximum value of C is given by:

$$C_{max} = \begin{cases} \frac{(M^3-4M)}{24} & M \text{ is even} \\ \frac{(M^3-M)}{24} & M \text{ is odd.} \end{cases} \tag{25}$$

Note that, $\zeta = 1$ represents a perfect consistency in the pairwise comparisons.

The consistency coefficient for each observer is calculated by averaging consistency coefficients across all the images used in the experiment. Whereas, the consistency coefficient for each image is computed by averaging consistency coefficients for all observers participated in the experiments (see Tables 6 and 7).

5.3. Kendall's coefficient of concordance

Kendall's Coefficient of Concordance (K_W) is also used to measure the degree of agreement in the rankings among different observers. It is calculated as follows:

$$K_W = \frac{(12 \times S)}{(S^2(M^3-M) - S \times T)} \tag{26}$$

where S and M are the number of observers and the number of methods respectively. T is the correction factor, when there are ties in the rank. T is zero when there is no tie within the rank. The correction factor T is calculated as follows:

$$T = \sum_{k=1}^{K} t_k^3 - t_k \tag{27}$$

where K is the total number of tie groups, and t_k is the total number of ties in a particular group.

To determine the significance of K_W, we calculated the χ^2 value given by:

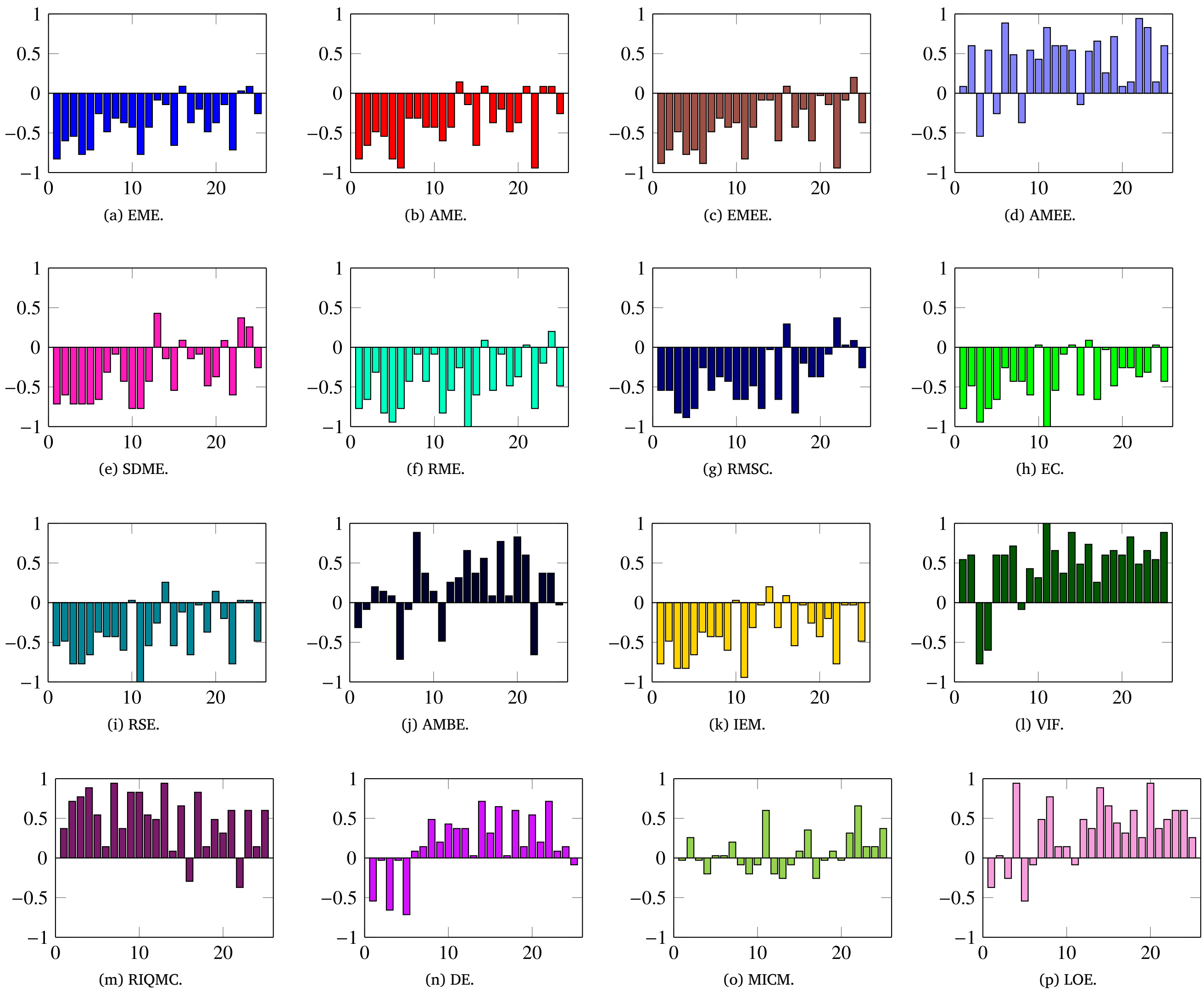
$$\chi^2 = S(M-1)K_W. \tag{28}$$

Then, the probability of getting the results by chance (p-value) is also calculated using the χ^2 distribution. The p-values for the experimental data indicates that the consistency coefficients are significant. However, we removed some images in the comparisons where these coefficients values are low.

For our preference based pairwise rank data collected from the subjective experiment, the values of these coefficients are presented in Table 6. From the significant tests, we have noticed that inter-observers' and intra-observers' consistency coefficients for the images in our new database are high except for the images numbered 6, 8, 12, 16, and 26. We then discarded these images and their related data and did not use these in further experiments.

6. Results and discussions

From the subjective experiments, we have derived the preference scores, i.e., the number of times an image enhanced by a particular CE method is preferred over other enhanced images. We have used the Spearman Rank Order Correlation Coefficient (SROCC) and the Kendall Rank Order Correlation Coefficient (KROCC) to observe the consistency of the CEE measures with the human visual perception [60,66]. The SROCC and KROCC are widely used non-parametric measures to determine the monotonicity between the ranks of two variables and their values ranges from −1 to +1. The values are close to +1 in the case of



(a) EME. (b) AME. (c) EMEE. (d) AMEE.

(e) SDME. (f) RME. (g) RMSC. (h) EC.

(i) RSE. (j) AMBE. (k) IEM. (l) VIF.

(m) RIQMC. (n) DE. (o) MICM. (p) LOE.

Fig. 7. SROCC plots for 25 images in the database. The x and y-axes represent image index and correlation values respectively.

strong correlation between the ranks of two variables and close to −1 in the case of strong disagreement between the two variables. The SROCC and KROCC give zero values when there is no correlation between the ranks. In our study, the aim is to observe how well a CEE measure is consistent in capturing the ranking for the six enhanced versions of each original image in the database. Therefore, before performing the correlations, we must consider, how the change in magnitude of metric values affects the image quality. For some metrics, high values correspond to good quality, whereas for other metrics, the opposite is true (see Section 2.2). For preference ranking, the highest score is highly ranked. Whereas, the metrics with high/low values corresponding to good quality are also highly ranked. Using the concept of strength of correlation alone without signs will not convey the desired purpose [25]. It is worth noting that for AMBE, AME, LOE, RIQMC, and SDME, the small values correspond to good image quality, whereas for other metrics in comparison, good quality corresponds to large values. The correlation is then calculated between the two rankings.

We calculate the median and mean correlations for each of CEE measures under study. For each image in the database, we have the ranking scores for its six enhanced versions as well as the quantitative scores. If I_i represents an original image and $I_{i,j}$, its enhanced version processed by method M_j, for $i = 1, 2, \dots, n_I$ and $j = 1, 2, \dots, n_J$, for $(n_I = 25, n_J = 6)$. Here n_I and n_J represent the number of original images and the number of CE methods respectively. We compute the SROCC for each image using the following relation:

$$\rho_{i,k} = 1 - \frac{6d_{i,j}^2}{n_J(n_J^2 - 1)}, \text{ for } i = 1, 2, \dots, n_I \tag{29}$$

where $d_{i,j}$ represents the difference in the ranks of subjective preferences and objective scores of kth CEE measure for the ith image. The correlation for each image for all the CEE measures are shown in Fig. 7. Finally, the median and mean of the SROCC's for the 25 images are computed as a single performance measure of each CEE measure and reported in Table 8. Similarly, two different types of KROCC across the images, i.e., τ_{median} and τ_{mean} are also calculated and shown in Table 8. The median statistic is more robust to outliers than the mean statistic. Therefore median correlation is used for further analysis in the experiments. The mean values are only mentioned here for completeness. Moreover, to get an idea of deviations of correlation values for a given CEE metric, the tolerance information of the individual correlation per test image in terms of minimum, maximum, and standard deviations are also shown in Table 8.

The performance of different CEE metrics is also compared with other existing databases with contrast manipulated images and results are shown in Table 9.

From the correlation results on different databases, we observed that the performance of various CEE metrics might differ on some databases. The metrics which perform better for some databases do not work

Table 8
Correlation analysis of CEE measures on the proposed CEED2016 database.

Measure	SROCC (ρ)				KROCC (τ)			
	median	mean	[min,max]	std.dev.	median	mean	[min,max]	std.dev.
EME [14]	−0.3714	−0.3896	[−0.8286, +0.0883]	+0.2721	−0.3333	−0.2905	[−0.7333, +0.0716]	+0.2336
AME [16]	−0.4286	−0.3896	[−0.9429, +0.1429]	+0.3258	−0.2000	−0.2958	[−0.8667, +0.2000]	+0.3021
EC [46]	−0.4286	−0.4079	[−1.000, +0.0883]	+0.3078	−0.3333	−0.3118	[−1.000, +0.0716]	+0.2743
AMBE [38]	0.2000	0.1892	[−0.7143, +0.8857]	+0.4276	0.2000	0.1747	[−0.6000, +0.7333]	+0.3391
SDME [17]	−0.4286	−0.3325	[−0.7714, +0.4286]	+0.3696	−0.3333	−0.2425	[−0.6000, +0.4667]	+0.3026
IEM [45]	−0.4286	−0.3782	[−0.9429, +0.2000]	+0.3214	−0.3333	−0.2958	[−0.8667, +0.0716]	+0.2684
VIF [20]	0.6000	0.4797	[−0.7714, +1.000]	+0.4163	0.4667	0.3884	[−0.6000, +1.000]	+0.3519
AMEE [16]	0.5429	0.3892	[−0.5429, +0.9429]	+0.4014	0.3333	0.3241	[−0.4667, +0.8667]	+0.3361
EMEE [15]	−0.4286	−0.4239	[−0.9429, +0.2000]	+0.3238	−0.3333	−0.3225	[−0.8667, +0.0716]	+0.2776
RME [15]	−0.4857	−0.4468	[−1.000, +0.2000]	+0.3349	−0.3333	−0.3598	[−1.000, +0.0716]	0.2959
RSE [47]	−0.4286	−0.3819	[−1.000, +0.2571]	+0.3274	−0.3333	−0.2802	[−1.000, +0.2000]	0.2827
RMSC [39]	−0.4286	−0.3905	[−0.8857, +0.3714]	+0.3513	−0.3333	−0.3167	[−0.7333, +0.2148]	+0.2822
MICM [41]	0.0286	0.0713	[−0.2571, +0.6571]	+0.2472	0.0667	0.0726	[−0.2000, +0.4667]	+0.1747
LOE [44]	0.3714	0.3377	[−0.5429, +0.9429]	+0.3971	0.3333	0.2810	[−0.3333, +0.8667]	+0.3153
DE [40]	0.1429	0.1676	[−0.7143, +0.7143]	+0.3867	0.0667	0.1161	[−0.6000, +0.6000]	+0.3257
RIQMC [21]	0.5429	0.4865	[−0.3714, +0.9429]	+0.3565	0.4667	0.4021	[−0.2148, +0.8667]	+0.2972

Table 9
Rank correlation values of different CEE metrics calculated on existing contrast-changed databases.

Metrics	CCID2014 [31]		DRIQ [19]		TID2013 [18]		CSIQ [30]	
	SROCC	KROCC	SROCC	KROCC	SROCC	KROCC	SROCC	KROCC
EME [14]	0.3543	0.2487	0.3221	0.2234	0.4986	0.3714	0.8360	0.6546
AME [16]	−0.0414	−0.0106	0.5367	0.3660	0.4530	0.3430	0.7376	0.5376
EC [46]	0.8286	0.6328	0.7457	0.5504	0.8060	0.6252	0.9532	0.8150
AMBE [38]	0.5965	0.4359	0.0129	0.0010	0.1715	0.1550	0.4963	0.3559
SDME [17]	0.0096	0.0335	0.5860	0.4059	0.4263	0.3221	0.8460	0.6549
IEM [45]	0.8352	0.6406	0.7427	0.5465	0.8065	0.6218	0.9519	0.8138
VIF [20]	0.8349	0.6419	−0.6679	−0.4619	0.7716	0.5795	0.9345	0.7769
AMEE [16]	0.6285	0.4489	−0.4353	−0.2927	0.0734	0.0395	0.4403	0.3040
EMEE [15]	0.1021	0.0807	0.3802	0.2734	0.4908	0.3581	0.8208	0.6129
RME [15]	−0.1625	0.0879	0.5274	0.3706	0.4660	0.3557	0.3290	0.2459
RSE [47]	0.7391	0.5396	0.7154	0.5098	0.7620	0.5678	0.9507	0.8078
RMSC [39]	0.7348	0.5336	0.3449	0.2268	0.7855	0.5922	0.9516	0.8126
MICM [41]	0.6917	0.5016	−0.6305	−0.4252	0.5733	0.4430	0.8005	0.6222
LOE [44]	0.6770	0.4844	−0.7716	−0.5564	0.1440	0.1157	0.7130	0.5316
DE [40]	0.8592	0.6690	0.2830	0.1975	0.7356	0.5598	0.9501	0.8120
RIQMC [40]	−0.8464	−0.6507	−0.3336	−0.2328	−0.8044	−0.6178	−0.9580	−0.8279

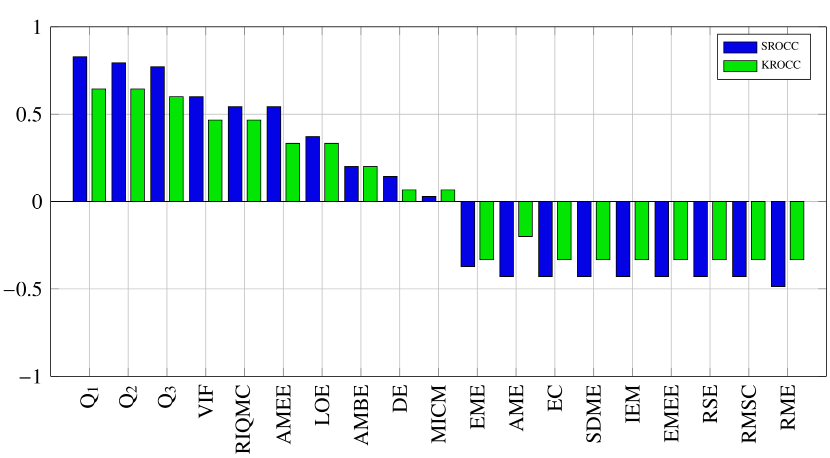


Fig. 8. Subjective vs objective comparisons of CE evaluation in terms of median correlations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

well on others. The reasons are that the metrics are not adapted to different CE distortions, and the databases do not contain enhanced images affected by various CE distortions.

The comparison of the median correlations between subjective and objective data for the CEE measures is also shown in Fig. 8. From Fig. 8, we observe that only VIF, RIQMC, AMEE, LOE, AMBE, DE, and MICM metrics have positive correlations with the subjective ranking. The negative correlation of other CEE measures show the inconsistencies of these measures with the human perception of quality judgment. This is due to the inconsistencies that exist between the metric values and subjective preferences for the TOPHAT and GHE methods, where these two approaches were ranked worst based on the observers' preferences. We can also draw a conclusion from the correlation analysis, that most of the CEE measures are not well suited for the CE evaluation of GHE and TOPHAT based CE methods. From these results, it is also clear that using simple local features, such as contrast or gradient, in the design of CE measure is not sufficient. It is important to include the color aspects in the design of CEE measures for color images. Through this study,

Table 10
Statistical significance analysis for CEE metrics.

	EME	AME	EC	AMBE	SDME	IEM	VIF	AMEE	EMEE	RME	RSE	RMSC	MICM	LOE	DE	RIQMC
EME	0	1	1	−1	1	1	−1	−1	1	1	1	1	−1	−1	−1	−1
AME	−1	0	1	−1	1	1	−1	−1	1	1	1	1	−1	−1	−1	−1
EC	−1	−1	0	−1	1	1	−1	−1	1	1	1	1	−1	−1	−1	−1
ANBE	1	1	1	0	−1	−1	−1	1	−1	−1	−1	−1	1	1	1	−1
SDME	−1	−1	−1	1	0	1	−1	−1	1	1	1	1	−1	−1	−1	−1
IEM	−1	−1	−1	1	−1	0	−1	−1	1	1	1	1	−1	−1	−1	−1
VIF	1	1	1	1	1	1	0	1	−1	−1	−1	−1	−1	1	−1	1
AMEE	1	1	1	−1	1	1	−1	0	−1	−1	−1	−1	−1	1	1	1
EMEE	−1	−1	−1	1	−1	−1	1	1	0	1	1	1	−1	−1	−1	−1
RME	−1	−1	−1	1	−1	−1	1	1	−1	0	1	1	−1	−1	−1	−1
RSE	−1	−1	−1	1	−1	−1	1	1	−1	−1	0	1	−1	−1	−1	−1
RMSC	−1	−1	−1	1	−1	−1	1	1	−1	−1	−1	0	−1	−1	−1	−1
MICM	1	1	1	−1	1	1	1	1	1	1	1	1	0	−1	1	−1
LOE	1	1	1	−1	1	1	−1	−1	1	1	1	1	1	0	−1	−1
DE	1	1	1	−1	1	1	1	−1	1	1	1	1	−1	1	0	−1
RIQMC	1	1	1	1	1	1	−1	−1	1	1	1	1	1	1	1	0

Table 11
Computational time for different CEE metrics.

Metric	Time (s)	Metric	Time (s)	Metric	Time (s)	Metric	Time (s)
EME [14]	0.143	AME [16]	0.121	EC [46]	0.417	AMBE [38]	0.031
SDME [17]	0.655	IEM [45]	2.593	VIF [20]	3.419	AMEE [16]	0.295
EMEE [15]	0.269	RME [15]	3.13	RSE [47]	1.929	RMSC [39]	0.705
MICM [41]	1.661	LOE [44]	0.248	DE [40]	0.652	RIQMC [40]	1.271

it is shown that CE evaluation is still a very challenging problem. We believe that the introduction of some learning based approaches would offer better solutions to this very challenging problem.

6.1. Statistical significance analysis

The statistical significance performance of the CEE metrics has been evaluated. A two sample T-test with 95% confidence level has been conducted. The results are shown in Table 10. A value of 1 indicates that the row metric is statistically superior to the column metric, whereas a value of −1 indicates that the row metric is statistically inferior to the column metric. A value of "0" indicates that the two metrics are statistically indistinguishable. From the obtained result, we observe that RIQMC metric is statistically superior compared to the other CEE metrics.

6.2. Computational time analysis

In real-time applications, the low-complexity CEE metric is needed. Therefore, in addition to comparing the correlation performance, we also compute the computational time of each CEE metric. We list in Table 11, the time taken (in seconds) to compute the metric score for a single image of resolution 512×512. The experiments are performed on notebook Intel Core i5-24500M CPU@2.5 GHz and 4G RAM. The software platform is MATLAB R2013b under Windows 8.1. The results show that the VIF which is superior in terms of correlation requires the largest computational time. Whereas the RIQMC, AMEE, and LOE have moderate computational complexity. The results in Table 11 only provide an idea about the comparative complexity of different CEE metrics. In real-time applications, the algorithms can be substantially optimized.

6.3. Multi-metric fusion based CEE measure

It is also evident that a single metric cannot perform very well. This is due to the reason, that no metric is sensitive to different types of artifacts introduced due to CE process. We have seen through the experiments that only seven CEE metrics show a positive correlation with the subjective preferences. We believe that a good contrast image should have prominent textural details (high entropy), no color loss (low LOE), provides more information (large VIF, and MICM), high local contrast (high AMEE) and some ordered-statistical features like mean, variance, skewness, kurtosis (RIQMC).

Table 12
Median correlation results for combining different CEE metrics.

Fused metrics	SROCC	KROCC	Weights
(Q_1) VIF + RIQMC + AMEE + LOE	+0.8286	+0.6445	[0.1,0.1,0.1,0.7]
(Q_2) VIF + RIQMC + LOE	+0.7945	+0.6445	[0.1,0.1,0.8]
(Q_3) VIF + LOE	+0.7714	+0.6000	[0.1,0.9]

Therefore, we propose to combine some best metrics to benefit from their strengths in quantifying the image contrast. We use a simple weighting based fusion and tune the weights to avoid the limitations of each metric and increase the correlation performance. We use the top four metrics with positive correlations, (i.e., VIF, RIQMC, LOE, and AMBE), and fuse their possible combinations using different weights. We show in Table 12, only the top three combinations with high correlations. Compared with the single metrics, the multi-metric fusion results in a substantial increase in correlation performance at the cost of an increase in complexity. From the fusion weights, we observe that LOE metric which captures the naturalness property of an image is considered more important by giving more weights in the fusion process. These observations provide us hints in designing new metrics to consider different quality parameters (e.g., naturalness, lightness, saturation, color-shift, visibility of edges, etc.) in contrast enhancement applications.

7. Conclusion

In this paper, a comprehensive psychophysical-based performance comparison of different state-of-the-art CEE measures is presented. The analysis was carried over a new database, that we introduced, which consists of enhanced images using different CE methods most commonly found in practice. Extensive subjective experiments were performed using a balanced pairwise preference-based ranking protocol to rank the CE methods by perceived quality. The correlation between subjective preferences and objective measures showed that most of the existing CEE measures are not well adapted with human perception of enhancement quality. Our analysis revealed that only seven measures, namely VIF, RIQMC, AMEE, LOE, AMBE, DE, and MICM exhibit positive correlations with perceptual quality of contrast enhancement. This is due to the fact that a single metric may be unable to capture various CE artifacts. For each available metric, we provided a detailed analysis of its strengths, weaknesses, and its inter-correlation with other metrics. We also showed that multi-metric fusion results in substantial improvement in correlation performance. To the best knowledge of the authors, the analysis provided here is most extensive and most comprehensive analysis of CEE metrics, to date.

Moreover, most existing research on IQA uses handcrafted image features for objective quality assessment of enhanced images. However, the selection of the most representative and relevant features is by itself a challenging task. Recently, deep learning approaches have been shown to provide excellent alternative models and remarkable results in diverse image processing applications. In deep learning, rather than focusing on feature extraction, the features are automatically learned from the training data. Recently, Li et al. [67] proposed an NR-IQA metric using convolutional neural network and Prewitt magnitude of segmented image patches showing very good correlation performance for different types of image distortions. We expect to see more work on using deep learning strategies in introducing new CEE measures for automatic perceptual feature learning to evaluate the perceptual quality of both distorted and enhanced images.

Furthermore, the paper provides an insight to consider various CE distortions in designing a new CEE metric. The newly developed database is expected to provide a platform for developing new CEE measures and benchmarking the results without the need for dedicated subjective experiments. The developed database along with the subjective experimental data is available for use free of charge [68].

Acknowledgments

The authors would like to thank the editor and the anonymous reviewers for their valuable comments. The work presented here has been developed in collaboration with L2TI Research Lab, Univ. Paris 13. The research was supported in part by the project GTEC 1401-1402 under the joint Center of Energy and Geoprocessing (CeGP) at King Fahd University of Petroleum & Minerals (KFUPM) and Georgia Tech.

References

[1] D.M. Chandler, Seven challenges in image quality assessment: Past, present, and future research, ISRN Sig. Process. 2013 (2013) 1–53. http://dx.doi.org/10.1155/2013/905685.

[2] A. Beghdadi, M.C. Larabi, A. Bouzerdoum, K.M. Iftekharuddin, A survey of perceptual image processing methods, Signal Process. Image Commun. 28 (8) (2013) 811–831. http://dx.doi.org/10.1016/j.image.2013.06.003.

[3] M. Qureshi, M. Deriche, A. Beghdadi, Quantifying blur in colour images using higher order singular values, Electron. Lett. 52 (21) (2016) 1755–1757. http://dx.doi.org/10.1049/el.2016.1792.

[4] J.-R. Ohm, G.J. Sullivan, H. Schwarz, T.K. Tan, T. Wiegand, Comparison of the coding efficiency of video coding standards - including high efficiency video coding (HEVC), IEEE Trans. Circuits Syst. Video Technol. 22 (12) (2012) 1669–1684. http://dx.doi.org/10.1109/TCSVT.2012.2221192.

[5] P.G. Engeldrum, A theory of image quality: The image quality circle, J. Imaging Sci. Technol. 48 (5) (2004) 447–457.

[6] P.G. Engeldrum, A short image quality model taxonomy, J. Imaging Sci. Technol. 48 (2) (2004) 160–165.

[7] S. Winkler, F. Dufaux, D. Barba, V. Baroncini, Special issue on image and video quality assessment, Signal Process. Image Commun. 25 (7) (2010) 467–468. http://dx.doi.org/10.1016/j.image.2010.07.001.

[8] J. Lu, D.M. Healy Jr, J.B. Weaver, Contrast enhancement of medical images using multiscale edge representation, Opt. Eng. 33 (7) (1994) 2151–2161. http://dx.doi.org/10.1117/12.172254.

[9] E. Lee, S. Kim, W. Kang, D. Seo, J. Paik, Contrast enhancement using dominant brightness level analysis and adaptive intensity transformation for remote sensing images, IEEE Geosci. Remote Sens. Lett. 10 (1) (2013) 62–66. http://dx.doi.org/10.1109/LGRS.2012.2192412.

[10] R. Schettini, S. Corchs, Underwater image processing: State of the art of restoration and image enhancement methods, EURASIP J. Adv. Signal Process. 2010 (2010) 1–15. http://dx.doi.org/10.1155/2010/746052.

[11] M.A. Qureshi, M. Deriche, A bibliography of pixel-based blind image forgery detection techniques, Signal Process. Image Commun. 39 (2015) 46–74. http://dx.doi.org/10.1016/j.image.2015.08.008.

[12] L.K. Choi, J. You, A.C. Bovik, Referenceless prediction of perceptual fog density and perceptual image defogging, IEEE Trans. Image Process. 24 (11) (2015) 3888–3901. http://dx.doi.org/10.1109/TIP.2015.2456502.

[13] A. Le Négrate, A. Beghdadi, H. Dupoisot, An image enhancement technique and its evaluation through bimodality analysis, CVGIP Graph. Model. Image Process. 54 (1) (1992) 13–22. http://dx.doi.org/10.1016/1049-9652(92)90030-2.

[14] S.S. Agaian, K. Panetta, A. Grigoryan, Transform-based image enhancement algorithms with performance measure, IEEE Trans. Image Process. 10 (3) (2001) 367–382. http://dx.doi.org/10.1109/83.908502.

[15] K. Panetta, C. Gao, S.S. Agaian, No reference color image contrast and quality measures, IEEE Trans. Consum. Electron. 59 (3) (2013) 643–651. http://dx.doi.org/10.1109/TCE.2013.6626251.

[16] S.S. Agaian, B. Silver, K. Panetta, Transform coefficient histogram-based image enhancement algorithms using contrast entropy, IEEE Trans. Image Process. 16 (3) (2007) 741–758. http://dx.doi.org/10.1109/TIP.2006.888338.

[17] K. Panetta, Yicong Zhou, S.S. Agaian, Hongwei Jia, Nonlinear unsharp masking for mammogram enhancement, IEEE Trans. Inf. Technol. Biomed. 15 (6) (2011) 918–928. http://dx.doi.org/10.1109/TITB.2011.2164259.

[18] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.-C. Jay Kuo, Image database TID2013: Peculiarities, results and perspectives, Signal Process. Image Commun. 30 (2015) 57–77. http://dx.doi.org/10.1016/j.image.2014.10.009.

[19] C.T. Vu, T.D. Phan, P.S. Banga, 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, IEEE, Santa Fe, NM, USA, 2012, pp. 181–184. http://dx.doi.org/10.1109/SSIAI.2012.6202483.

[20] H. Sheikh, A. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444. http://dx.doi.org/10.1109/TIP.2005.859378.

[21] K. Gu, G. Zhai, W. Lin, M. Liu, The analysis of image contrast: From quality assessment to automatic enhancement, IEEE Trans. Cybern. 46 (1) (2016) 284–297. http://dx.doi.org/10.1109/TCYB.2015.2401732.

[22] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, G. Zhai, No-reference quality assessment of contrast-distorted images based on natural scene statistics, IEEE Signal Process. Lett. 22 (7) (2014) 838–842. http://dx.doi.org/10.1109/LSP.2014.2372333.

[23] P. Ledda, A. Chalmers, T. Troscianko, H. Seetzen, Evaluation of tone mapping operators using a High Dynamic Range display, ACM Trans. Graph. 24 (3) (2005) 640. http://dx.doi.org/10.1145/1073204.1073242.

[24] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, J. Hakkinen, CID2013: A database for evaluating no-reference image quality assessment algorithms, IEEE Trans. Image Process. 24 (1) (2015) 390–402. http://dx.doi.org/10.1109/TIP.2014.2378061.

[25] M. Rubinstein, D. Gutierrez, O. Sorkine, A. Shamir, A comparative study of image retargeting, ACM Trans. Graph. 29 (6) (2010) 1. http://dx.doi.org/10.1145/1882261.1866186.

[26] L. Ma, W. Lin, C. Deng, K.N. Ngan, Image retargeting quality assessment: A study of subjective scores and objective metrics, IEEE J. Sel. Top. Signal Process. 6 (6) (2012) 626–639. http://dx.doi.org/10.1109/JSTSP.2012.2211996.

[27] Z. Chen, T. Jiang, Y. Tian, Quality assessment for comparing image enhancement algorithms, in: IEEE Conf. Comput. Vis. Pattern Recognit., IEEE, Columbus, OH, 2014, pp. 3003–3010. http://dx.doi.org/10.1109/CVPR.2014.384.

[28] X. Liu, M. Pedersen, J.Y. Hardeberg, CID:IQ - A new image quality database, in: International Conference on Image and Signal Processing, (ICISP), in: Lecture Notes in Computer Science, vol. 8509, Springer International Publishing, Cherbourg, Normandy, France, 2014, pp. 193–202. http://dx.doi.org/10.1007/978-3-319-07998-1. URL http://www.colourlab.no/cid.

[29] L. Krasula, P. Le Callet, K. Fliegel, M. Klima, Quality assessment of sharpened images: Challenges, methodology, and objective metrics, IEEE Trans. Image Process. 26 (3) (2017) 1496–1508. http://dx.doi.org/10.1109/tip.2017.2651374.

[30] E.C. Larson, D.M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, J. Electron. Imaging 19 (1) (2010) 011006. http://dx.doi.org/10.1117/1.3267105. URL http://vision.okstate.edu/csiq/.

[31] K. Gu, G. Zhai, X. Yang, W. Zhang, M. Liu, Subjective and objective quality assessment for images with contrast change, in: 2013 IEEE International Conference on Image Processing, IEEE, Melbourne, VIC, 2013, pp. 383–387. http://dx.doi.org/10.1109/ICIP.2013.6738079.

[32] A. Beghdadi, A. Le Negrate, Contrast enhancement technique based on local detection of edges, Comput. Vis. Graph. Image Process. 46 (2) (1989) 162–174. http://dx.doi.org/10.1016/0734-189X(89)90166-7.

[33] K. Zuiderveld, Contrast limited adaptive histogram equalization, in: Graph. Gems IV, Elsevier, San Diego, CA, USA, 1994, pp. 474–485. http://dx.doi.org/10.1016/b978-0-12-336156-1.50061-6.

[34] J. Mukherjee, S. Mitra, Enhancement of color images by scaling the DCT coefficients, IEEE Trans. Image Process. 17 (10) (2008) 1783–1794. http://dx.doi.org/10.1109/TIP.2008.2002826.

[35] R. Hummel, Image enhancement by histogram transformation, Comput. Graph. Image Process. 6 (2) (1977) 184–195. http://dx.doi.org/10.1016/S0146-664X(77)80011-7.

[36] S. Mukhopadhyay, B. Chanda, A multiscale morphological approach to local contrast enhancement, Signal Process. 80 (4) (2010) 685–696. http://dx.doi.org/10.1016/s0165-1684(99)00161-9.

[37] S. Chen, A. Beghdadi, Natural enhancement of color image, EURASIP J. Image Video Process. 2010 (1) (2010) 1–19. http://dx.doi.org/10.1155/2010/175203.

[38] S.-D. Chen, A.R. Ramli, Minimum mean brightness error Bi-histogram equalization in contrast enhancement, IEEE Trans. Consum. Electron. 49 (4) (2003) 1310–1319. http://dx.doi.org/10.1109/TCE.2003.1261234.

[39] E. Peli, Contrast in complex images, J. Opt. Soc. Am. A 7 (10) (1990) 2032–2040. http://dx.doi.org/10.1364/JOSAA.7.002032.

[40] C.E. Shannon, A mathematical theory of communication, ACM SIGMOBILE Mob. Comput. Commun. Rev. 5 (1) (2001) 3. http://dx.doi.org/10.1145/584091.584093.

[41] A. Beghdadi, M.A. Qureshi, M. Deriche, A critical look to some contrast enhancement evaluation measures, in: 2015 Colour and Visual Computing Symposium, (CVCS), IEEE, Gjovik, Norway, 2015, pp. 1–6. http://dx.doi.org/10.1109/CVCS.2015.7274888.

[42] M.A. Qureshi, M. Deriche, A. Beghdadi, M. Mohandes, An information based framework for performance evaluation of image enhancement methods, in: 2015 International Conference on Image Processing Theory, Tools and Applications, (IPTA), IEEE, Orleans, France, 2015, pp. 519–523. http://dx.doi.org/10.1109/IPTA.2015.7367201.

[43] R. Halonen, S. Westman, P. Oittinen, Naturalness and interestingness of test images for visual quality evaluation, in: Image Quality and System Performance VIII, Vol. 7867, SPIE, 2011. http://dx.doi.org/10.1117/12.872390. 78670Z–78670Z–12.

[44] S. Wang, J. Zheng, H.-M. Hu, B. Li, Naturalness preserved enhancement algorithm for non-uniform illumination images, IEEE Trans. Image Process. 22 (9) (2013) 3538–3548. http://dx.doi.org/10.1109/TIP.2013.2261309.

[45] V.L. Jaya, R. Gopikakumari, IEM : A new image enhancement metric for contrast and sharpness measurements, Int. J. Comput. Appl. 79 (9) (2013) 1–9. http://dx.doi.org/10.5120/13766-1620.

[46] A. Saleem, A. Beghdadi, B. Boashash, Image fusion-based contrast enhancement, EURASIP J. Image Video Process. 2012 (10) (2012) 1–17. http://dx.doi.org/10.1186/1687-5281-2012-10.

[47] A. Chetouani, A. Beghdadi, M. Deriche, A new reference-free image quality index for blur estimation in the frequency domain, in: 2009 IEEE International Symposium on Signal Processing and Information Technology, (ISSPIT), IEEE, Ajman, 2009, pp. 155–159. http://dx.doi.org/10.1109/ISSPIT.2009.5407502.

[48] K. Panetta, C. Gao, S. Agaian, No reference color image contrast and quality measures, IEEE Trans. Consum. Electron. 59 (3) (2013) 643–651. http://dx.doi.org/10.1109/TCE.2013.6626251.

[49] ITU-R Recommendations BT.500-13, Methodology for the subjective assessment of the quality of television pictures, Tech. rep., International Telecommunications Union, Geneva, Switzerland, Jan 2012, URL http://www.itu.int/rec/R-REC-BT.500-13-201201-I/en.

[50] ITU-T Recommendations P.913, Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment, Tech. rep., International Telecommunications Union, Geneva, Switzerland, 2014.

[51] R.K. Mantiuk, A. Tomaszewska, R. Mantiuk, Comparison of four subjective methods for image quality assessment, Comput. Graph. Forum 31 (8) (2012) 2478–2491. http://dx.doi.org/10.1111/j.1467-8659.2012.03188.x.

[52] M.H. Pinson, S. Wolf, Comparing subjective video quality testing methodologies, in: SPIE Visual Communications and Image Processing, Vol. 5150, SPIE, Lugano, 2003, pp. 573–582. http://dx.doi.org/10.1117/12.509908.

[53] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, A. Raake, Study of rating scales for subjective quality assessment of high-definition video, IEEE Trans. Broadcast. 57 (1) (2011) 1–14. http://dx.doi.org/10.1109/tbc.2010.2086750.

[54] F. Kozamernik, V. Steinmann, P. Sunna, E. Wyckens, SAMVIQ - a new EBU methodology for video quality evaluations in multimedia, SMPTE Motion Imaging J. 114 (4) (2005) 152–160. http://dx.doi.org/10.5594/j11535.

[55] Technical Report, European Broadcast Union (EBU), SAMVIQ - subjective assessment methodology for video quality, Tech. rep., Tech. Rep. BPN 056, May 2003.

[56] H. Liu, A.R. Reibman, Software to stress test image quality estimators, in: 2016 Eighth International Conference on Quality of Multimedia Experience, (QoMEX), IEEE, Lisbon, Portugal, 2016, pp. 1–6. http://dx.doi.org/10.1109/qomex.2016.7498945.

[57] A.R. Reibman, A strategy to jointly test image quality estimators subjectively, in: 19th IEEE International Conference on Image Processing, (ICIP), IEEE, 2012, pp. 1501–1504. http://dx.doi.org/10.1109/icip.2012.6467156.

[58] P. Hanhart, L. Krasula, P. Le Callet, T. Ebrahimi, How to benchmark objective quality metrics from paired comparison data?in: 2016 Eighth International Conference on Quality of Multimedia Experience, (QoMEX), IEEE, Lisbon, Portugal, 2016, pp. 1–6. http://dx.doi.org/10.1109/qomex.2016.7498960.

[59] L. Krasula, K. Fliegel, P. Le Callet, M. Klíma, On the accuracy of objective image and video quality models: New methodology for performance evaluation, in: 2016 Eighth International Conference on Quality of Multimedia Experience, (QoMEX), IEEE, Lisbon, Portugal, 2016, pp. 1–6. http://dx.doi.org/10.1109/qomex.2016.7498936.

[60] S. Siegel, Nonparametric Statistics for The Behavioral Sciences, 2nd ed., McGraw-Hill, 1956, pp. 1–312.

[61] J. Li, M. Barkowsky, P. Le Callet, Boosting paired comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs, in: Stereoscopic Displays and Applications XXIV, vol. 8648, SPIE, 2013. http://dx.doi.org/10.1117/12.2002075. 86481V–86481V-12.

[62] S. Winkler, Analysis of public image and video databases for quality assessment, IEEE J. Sel. Top. Signal Process. 6 (6) (2012) 616–625. http://dx.doi.org/10.1109/JSTSP.2012.2215007.

[63] D. Hasler, S.E. Suesstrunk, Measuring colorfulness in natural images, in: B.E. Rogowitz, T.N. Pappas (Eds.), Human Vision and Electronic Imaging VIII, Vol. 5007, SPIE, 2003, pp. 87–95. http://dx.doi.org/10.1117/12.477378.

[64] K. Matkovic, L. Neumann, A. Neumann, T. Psik, W. Purgatholer, Global contrast factor - a new approach to image contrast, in: Computational Aesthetics in Graphics, Visualization and Imaging, The Eurographics Association, 2005, pp. 159–167. http://dx.doi.org/10.2312/COMPAESTH/COMPAESTH05/159-167.

[65] M.G. Kendall, B.B. Smith, On the method of paired comparisons, Biometrika 31 (3/4) (1940) 324–345. http://dx.doi.org/10.2307/2332613.

[66] ITU-R Recommendations P.1401, Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models, Tech. rep., International Telecommunications Union, Jul 2012.

[67] J. Li, L. Zou, J. Yan, D. Deng, T. Qu, G. Xie, No-reference image quality assessment using Prewitt magnitude based on convolutional neural networks, Signal Image Video Process. 10 (4) (2016) 609–616. http://dx.doi.org/10.1007/s11760-015-0784-2.

[68] M.A. Qureshi, M. Deriche, A. Beghdadi, Contrast Enhancement Evaluation Database (CEED2016), 2017. Mendeley Data, v1. http://dx.doi.org/10.17632/3hfzp6vwkm.1.