International Journal of Applied Information Systems (IJAIS) ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 9 - No. 1, June 2015 - www.ijais.org

Evaluation of Image Quality Assessment Metrics: Color Quantization Noise

Mohammed Hassan
Assistant Professor, DACS
Seiyun Community College
Hadramout, Yemen

Chakravarthy Bhagvati
Professor, DCIS
University of Hyderabad
Hyderabad-500046, India

ABSTRACT

Although color quantization noise is frequently met in practice, it has not been given too much attention in color image visual quality assessment. In this paper, a new image database for the evaluation of image quality metrics over color quantization noise is described. It contains 25 reference images and 875 test images produced by five popular quantization algorithms. Each of the quantized images was evaluated by 22 human subjects and more than 19200 individual human quality judgments were carried out to obtain the final mean opinion scores. A comparative analysis of several well-known image quality metrics is presented and their correlation with the human opinion scores is evaluated. This image database has been made freely available for downloading for research in image quality assessment and other applications [10].

General Terms:

Subjective image quality assessment, Objective image quality assessment

Keywords:

Image quality metrics,Color quantization, Image database

1. INTRODUCTION

Image quality assessment is an important tool in image processing systems. Image quality assessment methods can be classified into two categories: subjective and objective. The subjective image quality assessment methods are accurate in estimating the visual quality of an image because they are carried out by human subjects but are costly and inconvenient processes which require a large number of observers, take a significant time and nor can they be automated. On the other hand the objective image quality assessment methods are computer based methods that can automatically predict the perceived image quality. Therefore the objective image quality assessment methods gained more popularity although they tend to correlate rather poorly with human perception of images quality.

Originally, color quantization has been used to satisfy the display hardware constraints that allow only a limited number of colors to be displayed simultaneously. Today the original motivation of color quantization has changed due to the availability of inexpensive full color displays. However, color quantization is still an important problem in the fields of image processing and computer graphics [2]. It can be used in mobile and hand-held devices where memory is usually small [25], it can be used for low-cost display and printing devices where only a small number of colors can be displayed or printed simultaneously [27], it also has been used as a preprocessing step for many applications such as object recognition [33], image compression [41], and content-based image retrieval (CBIR) [34]. Another aspect of importance of color quantization is that the human visual system cannot perceive a large number of colors at the same time, nor is it able to distinguish close colors well [21] even though under appropriate adaptation, it cannot distinguish more than two million colors [18] while a full color image may contain up to 16 million different colors. This large number of colors makes it difficult to handle a variety of color-based tasks such as edge detection, enhancement, computing histograms,and color adjustment. These tasks are more efficiently carried out on a small set of colors.

Today, a large variety of objective image quality assessment algorithms has been proposed starting from the widely used mean square error (MSE) metric and its signal processing counterpart, the peak signal to noise ratio (PSNR). The weighted signal to noise ratio (WSNR) [19] simulates the human visual system properties by filtering both the reference and distorted images with contrast sensitivity function and then compute the signal to noise ratio. Miyahara et al. [20] proposed a picture quality scale (PQS) based on three distortion factors: the amount, location and structure of error. The perceptual color fidelity metric (S-CIELAB) [44] is a spatial extension to the CIELAB metric for measuring color reproduction errors of digital images. It simulates the spatial sensitivity of the human visual system by spatial filter processes on images. Wang and Bovik [36] proposed a new universal image quality index (UQI) and its improved form, the single-scale structural similarity index (SSIM) [37], by modeling the image distortion as the combination of loss of luminance, contrast, and correlation. The single-scale structural similarity index was extended into a multi-scale structural similarity index (MSSIM) [39] that works on scale space of an image and achieved a better result than SSIM. Information fidelity criterion (IFC) [29] and visual information fidelity (VIF) [28] both are based on information-theory in which the distorted image is modeled as a sequence of passing the reference images through distortion channels and quantify the visual quality as a mutual information between the test image and the reference image. Shnayderman et al. [32] explored the feasibility of singular value decomposition (SVD) for image quality measurement. A two staged wavelet based visual signal to noise ratio (VSNR) [4] was proposed based on the low-level and the mid-level properties of human vision. A structural information-based image quality assessment algorithm [9] uses LU factorization for representation of the structural information of an image. An image quality metric using the phase quantization code [14] was proposed and extended to amplitude/phase quantization code [13]. Wang and Li [38] incorporated the idea of information content weighted pooling and applied it to peak signal to noise ratio (PSNR) and structural similarity measure (SSIM) leading to an information content weighted PSNR (IW-PSNR) and an information content weighted SSIM (IW-SSIM). A feature similarity index (FSIMc) [43] for color image quality assessment is proposed based on the fact that human visual system understands an image mainly according to its low-level features. Two kinds of features, the phase congruency (PC) and the image gradient magnitude (GM), are used in FSIMc.

Most of the objective image quality metrics are claimed by their authors to be human perception correlated; this raises the need to judge them. Although the subjective image quality assessments are very difficult to carry out, expensive, and time consuming task but are still the only reliable way of evaluating the correla-

tion of the objective image quality assessment algorithms with the human perception. This necessitates the creation of image databases with subjective opinion scores defining the human perception of quality.
A lot of researchers have contributed significantly in the design of subjective database for the assessment of image quality metrics. The studies [7, 1, 17, 26] present image databases consist of only compression distortions. The entire database A57 from Cornell University [5] consists of three reference images distorted by compression distortion, Gaussian blur, and Gaussian white noise. IVC database [3, 22] contains 235 distorted images generated from four distortion types JPEG, JPEG2000, LAR coding, and Blurring. In LIVE database [31, 30] there are twenty nine reference images distorted with compression distortions, Gaussian blur, White noise, and fast-fading to produce 779 test images. TID2008 database [23] which is the largest so far available image database with seventeen distortions types and 1700 test images. The CSIQ image database [15] consists of 30 original images distorted using six different types of distortions at four to five different levels of distortion. The distortions used in CSIQ are: JPEG compression, JPEG-2000 compression, global contrast decrements, additive pink Gaussian noise, additive white Gaussian noise, and Gaussian blurring. Recently, VCL@FER [42] is a new image database which consists of four degradation types: JPEG, JPEG2000, white noise and Gaussian blur with 23 different images and 6 levels of each degradation.
This paper presents a new image database for color image quantization noise which contains 875 quantized images produced by five popular color quantization algorithms. The performance of recent state-of-the-art full-reference image quality metrics over the new database is analyzed and reported. The paper is organized as follows: Section 2 describes the proposed image database including the subjective quality tests. In section 3, the prediction of the image quality metrics is evaluated. Section 4 shows statistical significance analysis of the image quality metrics. The study is concluded in section 5.

2. DESCRIPTION OF THE PROPOSED IMAGE DATABASE

2.1 Choice of Input Images

This image database [10] consists of 25 reference images collected from the Internet based on the number of segments and number of distinct colors. Those images reflect a variety of image contents includes important objects, uniform regions, slowly varying color gradients, edges, and high level of details. Fig. 1 shows the reference images used in the study: images from 1 to 7 have small number of segments and colors, images from 8 to 13 have small number of segments but large number of colors, images from 14 to 18 have large number of segments but small number of colors, while images from 19 to 25 have large number of segments and colors.

2.2 Color Quantization Algorithms

All images in the database are of size 512x512 pixels for the purpose of carrying out subjective experiments. Each of the resized images has been quantized into seven levels (4, 8, 16, 32, 64, 128, and 256 colors) using five color image quantization algorithms that are popular in literature and represent different approaches (dividing approach, merging approach, clustering approach, and neural networks approach). The color quantization algorithms are as follows:

—*K-means algorithm* [16] where a set of k initial centroids is randomly selected. At each step, a scan through all pixels of the original image is performed to assign each pixel to the nearest centroid. After that, a new set of centroids is generated as the means of the pixels associated to each centroid. These steps are repeated until the algorithm converges or the number of iterations reaches a specified value.

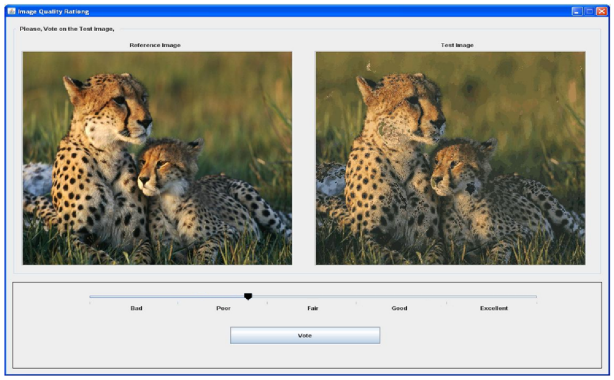

Fig. 2. Interface for the subjective study.

—*Median Cut algorithm* [11] which repeatedly bipartitions the color space into smaller and smaller rectangular boxes until the desired number of clusters is obtained with an approximately equal number of pixels at each level. The cutting plane is chosen normal to the longest axes and passes through the median point of the color distribution projected on this axis.

—*Wu's algorithm* [40] is similar to the Median cut algorithm except that the cutting plane is chosen perpendicularly to the R, G, B axes separately, and the plane that minimizes the sum of variances at both sides is chosen to cut the cube into two boxes. Next, the box with the larger variance is partitioned into two smaller boxes by the same cutting criterion.

—*Octree algorithm* [8] repeatedly divides the color space into eight smaller and smaller cubes in a way that the entire color space is treated as a hierarchy of octants and each individual color as a leaf of the octree. The octree is then reduced by replacing the leaves by their parent containing the average of that leaves until the desired number of clusters is obtained.

—*Dekkers SOM* [6] uses a one-dimensional Self-organizing Neural Network. The network contains one neuron for each desired cluster. Through the learning process each neuron acquires a weight vector which is used as possible representative. After learning is completed, pixels are mapped to the closest weight vector.

2.3 Number, Selection and training of Subjects

A group of twenty two undergraduate students participated in the psychometric experiment. The majority of the subjects were males and they were non-experts with image quality assessment. The reliability of the assessors was qualitatively evaluated by checking their behavior when reference/reference pairs where reliable subjects are expected to give evaluations very close to the maximum point in the quality scale.
Each subject was individually briefed about the goal of the experiment, what they are going to see, what they have to evaluate and how they express their opinion, the grading scale, the sequence, and timing. The subjects also have been shown some examples in how to evaluate the quality of quantized images. Those examples approximate the range of quality of the images for different quantization levels. Images in the training phase were different from those used in the actual experiment.

2.4 Display Equipments

The psychometric experiments were conducted in a lab with normal indoor illumination environment using Microsoft Windows workstations. The display monitors were all 19-inch CRT and were all approximately the same age. Although the monitors

Fig. 1. The reference images used in the study

were not calibrated, they were set to the same display settings. A java-based interface was used to show the images and to enable the observers vote the quality of the test images. The software was designed in a way that the observers assess the overall quality of quantized image with respect to reference image of each assessment trail presentation by simply dragging a slider on a quality scale. The quality scale which is of range [0,100] was unmarked numerically but labeled and divided into five equal categories: “Bad”, “Poor”, “Fair”, “Good”, and “Excellent” to be used as general guidance which range from the lowest to the highest perceptual quality grade. The position of the slider reflects the rate given by the observer for that image and its position was reset after each presentation. Fig. 2 shows the user interface of the quality assessment software.

2.5 Subjective Quality Tests

To evaluate the quality of the quantized images, a subjective quality test is used in which a number of human subjects are asked to judge the quality of the sequence images. The subjective tests are based on the recommendations given by the Interna-

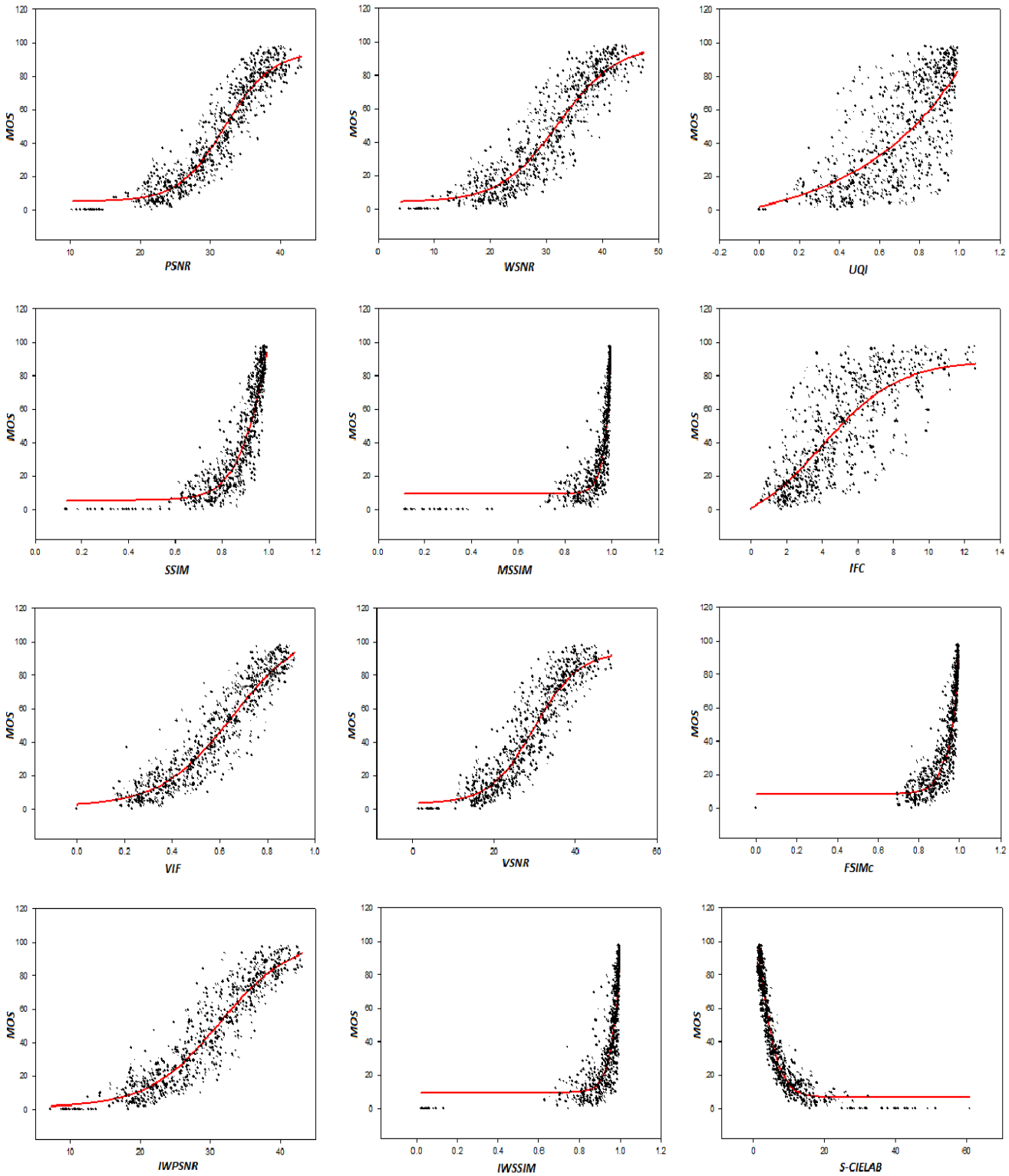

Fig. 3. Scatter plots for the subjective MOS versus the quality scores form different image quality metrics: PSNR, WSNR, UQI, SSIM, MSSIM, IFC, VIF, VSNR, FSIMc, IW-PSNR, IW-SSIM, S-CIELAB.

tional Telecommunication Union (ITU) [12] that define how to carry out subjective quality tests. The set of subjects is watching two images (reference and test) at the same time. The observers are asked to assess the quality of the test image with respect to the reference image of each assessment trail. In the proposed image database, there are 875 test images to be assessed and as recommended by ITU each session should not last more than 30 minutes. Therefore overall subjective tests were divided into five sessions (175 test images for each session). Five dummy images were added at the beginning of the first session and not considered in the calculation; their purpose is to stabilize the subjects

Table 1. Pearson's correlation coefficient of the scores given by different quality assessment methods against MOS from the subjective study after a non-linear mapping.

	SOM	*Median*	*Kmeans*	*Octree*	*Wu*	*All Data*
PSNR	0.956	0.965	0.960	0.970	0.957	0.945
WSNR	0.942	0.935	0.940	0.957	0.945	0.920
UQI	0.732	0.772	0.662	0.804	0.720	0.728
SSIM	0.929	0.940	0.911	0.935	0.930	0.913
MSSIM	0.935	0.934	0.910	0.944	0.940	0.917
IFC	0.806	0.783	0.791	0.869	0.812	0.805
VIF	0.950	0.938	0.951	0.967	0.957	0.942
VSNR	0.949	0.929	0.943	0.955	0.953	0.926
FSIMc	0.943	0.928	0.924	0.958	0.949	0.924
IWPSNR	0.959	0.935	0.957	0.969	0.963	0.934
IWSSIM	0.910	0.897	0.868	0.930	0.913	0.888
S-CIELAB	0.963	0.966	0.969	0.977	0.961	0.961

Table 2. Root Mean Square Error

	SOM	*Median*	*Kmeans*	*Octree*	*Wu*	*All Data*
PSNR	8.540	8.004	8.678	7.008	8.815	9.774
WSNR	9.801	10.908	10.603	8.383	9.974	11.722
UQI	19.871	19.511	23.205	17.243	21.147	20.574
SSIM	10.831	10.490	12.787	10.302	11.176	12.209
MSSIM	10.334	10.970	12.862	9.532	10.390	11.933
IFC	17.283	19.099	18.938	14.332	17.774	17.811
VIF	9.127	10.609	9.586	7.409	8.815	10.100
VSNR	9.168	11.372	10.304	8.629	9.187	11.289
FSIMc	9.692	11.403	11.823	8.298	9.647	11.474
IWPSNR	8.281	10.860	9.029	7.216	8.234	10.715
IWSSIM	12.076	13.596	15.375	10.672	12.429	13.771
S-CIELAB	7.885	7.920	7.679	6.158	8.400	8.312

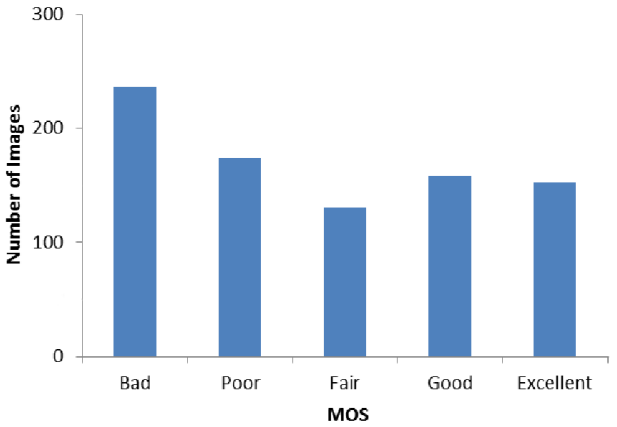

Fig. 4. Histogram of the MOS values for the five perceptual quality range.

to the rating process. Subjects were shown images in a random order and this order is unique for each subject.

2.6 Outliers Detection and Subject Rejection

Before starting analysis of the data, a screening of the subjective raw scores was conducted to eliminate observers with unstable scores [12]. The generalized ESD many-outlier procedure [24] was run twice to detect outliers within the subjective raw data. The generalized ESD many-outlier procedure selects the maximum k deviations from the mean and compares them with their corresponding critical values $\lambda_i, i = 1, .., k$ that define a cut points to decide whether an observation is an outlier. The values of $\lambda_i s$ are computed based on the percentage points from the Student's t distribution. If at any step i a maximum deviation is greater than its corresponding critical value λ_i then the extreme observations for the first i^{th} maximum deviations are all considered to be outliers even some of them are smaller than or equal to their corresponding critical values. About 2.66 % of the subjective raw data was rejected as being outliers. All quality evaluations of a subject were rejected if more than 5 % of his evaluations were outliers. Only one of the observers was rejected.

2.7 Calculation of Mean Opinion Scores

To calculate Mean Opinion Scores (MOS), the subjective raw data is first converted to Z-score (after outliers removal) to minimize the variation between individual subjective values due to not using the full range of quality scale by the different subjects during the image quality rating process [35]:

$$z_{ij} = \frac{(v_{ij} - \overline{v_i})}{\sigma_i} \quad (1)$$

where v_{ij} is the raw scores given by the i^{th} subject to j^{th} test image, $\overline{v_i}$ and σ_i are the mean and the standard deviation of raw scores over all test images evaluated by the i^{th} subject. The final MOS for each test image j is obtained by averaging all Z-scores z_{ij} given to that image by all subjects. Fig. 4 depicts a histogram of the subjective MOS scores from the proposed image database. Notice how the scores are uniformly distributed and span the entire range of perceptual qualities from low to high values.

3. IMAGE QUALITY METRICS AND HUMAN PERCEPTION

In this section, the performance of several well-known objective image quality metrics is evaluated. These quality metrics are commonly used and their implementations are publicly available

Table 3. Spearman's Rank Order Correlation Coefficient of the scores given by different quality assessment methods against MOS from the subjective study after a non-linear mapping.

	SOM	*Median*	*Kmeans*	*Octree*	*Wu*	*All Data*
PSNR	0.950	0.961	0.952	0.965	0.953	0.939
WSNR	0.934	0.932	0.933	0.952	0.941	0.916
UQI	0.742	0.774	0.679	0.802	0.728	0.735
SSIM	0.921	0.938	0.909	0.934	0.929	0.911
MSSIM	0.934	0.931	0.912	0.950	0.939	0.918
IFC	0.812	0.790	0.798	0.873	0.816	0.810
VIF	0.945	0.936	0.946	0.962	0.955	0.938
VSNR	0.944	0.925	0.938	0.954	0.952	0.923
FSIMc	0.942	0.930	0.928	0.959	0.949	0.926
IWPSNR	0.954	0.938	0.950	0.961	0.960	0.931
IWSSIM	0.911	0.897	0.872	0.935	0.913	0.890
S-CIELAB	0.954	0.958	0.960	0.972	0.957	0.956

Table 4. Outlier Ratio (the percentage of the number of predictions outside the range of two times the standard deviations of the subjective MOS)

	SOM	*Median*	*Kmeans*	*Octree*	*Wu*	*All Data*
PSNR	0.206	0.143	0.189	0.286	0.206	0.280
WSNR	0.257	0.297	0.274	0.349	0.223	0.359
UQI	0.589	0.606	0.686	0.617	0.611	0.647
SSIM	0.274	0.251	0.349	0.417	0.269	0.376
MSSIM	0.257	0.297	0.320	0.349	0.240	0.329
IFC	0.469	0.577	0.537	0.520	0.497	0.542
VIF	0.217	0.240	0.223	0.280	0.194	0.282
VSNR	0.229	0.337	0.240	0.349	0.194	0.335
FSIMc	0.200	0.326	0.337	0.331	0.229	0.328
IWPSNR	0.183	0.314	0.257	0.280	0.171	0.317
IWSSIM	0.303	0.394	0.383	0.377	0.326	0.383
S-CIELAB	0.183	0.160	0.194	0.251	0.234	0.240

on the Internet namely: peak signal to noise ratio (PSNR), the weighted signal to noise ratio (WSNR) [19], S-CIELAB [44], universal image quality index (UQI) [36], structural similarity index (SSIM) [37], multiscale structural similarity index (MSSIM) [39], information fidelity criterion (IFC) [29], visual information fidelity (VIF) [28], visual signal to noise ratio (VSNR) [4], information content weighted PSNR (IW-PSNR) and information content weighted SSIM (IW-SSIM) [38], and feature similarity (FSIMc) [43]. For grayscale metrics, the reference and test images are transformed using the Matlab function rgb2gray.

The scores given by an objective image quality metric are transferred into a predicted MOS to map the scores of the objective image quality metric into the range of the subjective MOS and to remove any nonlinearity between them using non-linear regression [12]. Fig. 3 shows the scatter plots of the scores given by the different objective image quality metrics versus the subjective MOS before non-linear regression. The function chosen for regression is a four parameters logistic function [30]:

$$MOS_p(Q) = \frac{p_1 - p_2}{1 + \exp\left(\frac{Q - p_3}{p_4}\right)} + p_2 \qquad (2)$$

where MOS_p is the predicted MOS, Q is the quality rating given by an objective image quality metric. The parameters p_1, p_2, p_3, and p_4 are chosen to minimize the mean square error between the quality scores given by the objective image quality metric and the subjective MOS.

A number of measures were used to evaluate the performance of the objective image quality metrics. These measures characterize three attributes related to the prediction of each image quality metric [12]:

(1) *Prediction Accuracy:* The ability of an objective image quality metric to predict the subjective MOS with minimum average error. Root mean square error and the Pearsons linear correlation coefficient were used to measure the prediction accuracy.

(2) *Prediction Monotonicity:* The ability of given by an objective image quality metric to give values that are monotonic in their relationship to the corresponding subjective MOS values. This attribute was measured by the Spearmans rank order correlation coefficient.

(3) *Prediction Consistency:* The ability of an objective image quality metric to provide consistently accurate predictions for all types of images and not to fail badly for a subset of images. Outlier ratio was used to measure the prediction consistency of an image quality metric.

Tables 1-4 show Pearsons correlation coefficient, root mean square error, Spearmans rank order correlation coefficient, and outlier ratio of the objective image quality metrics after logistic transformation for individual datasets as well as for the whole data. It is clear from those tables that S-CIELAB metric has the highest prediction accuracy, monotonicity, and consistency among all the metrics followed by PSNR, VIF, and IWPSNR that have comparable performance, while UQI and IFC are least accurate, monotonic, and consistent.

4. STATISTICAL SIGNIFICANCE OF IMAGE QUALITY METRICS

The statistical significance of each metric's performance relative to other metrics was evaluated by performing an F-test in the set of residuals (prediction errors). For variances σ_A^2 and σ_B^2 of two

Table 5. Normality test for the set residuals (Skewness / Kurtosis)

	SOM	*Median*	*Kmeans*	*Octree*	*Wu*	*All Data*
PSNR	-0.08 / 4.14	0.25 / 4.0	-0.13 / 4.16	0.02 / 3.6	0.48 / 4.3	0.13 / 3.6
WSNR	-0.22 / 4.22	0.29 / 3.5	0.06 / 4.2	-0.19 / 3.86	0.05 / 4.8	0.13 / 3.7
UQI	-0.21 / 2.71	-0.16 / 3.05	0.10 / 2.7	0.39 / 3.4	-0.13 / 3.02	-0.01 / 3.01
SSIM	-0.30 / 3.81	-0.47 / 4.80	-0.50 / 4.12	-0.15 / 3.67	-0.06 / 4.81	-0.11 / 4.01
MSSIM	-0.66 / 4.50	-0.64 / 4.49	-0.39 / 5.80	0.14 / 4.1	-0.96 / 6.05	-0.35 / 4.94
IFC	0.01 / 3.3	0.40 / 3.5	0.05 / 3.6	0.63 / 4.3	0.01 / 3.7	0.12 / 3.6
VIF	-0.37 / 4.02	0.16 / 4.3	-0.43 / 4.55	0.94 / 5.7	-0.22 / 4.73	-0.04 / 4.47
VSNR	-0.17 / 3.50	0.41 / 3.9	0.14 / 4.7	-0.20 / 4.50	0.17 / 4.9	0.28 / 3.9
FSIMc	-0.33 / 3.65	-0.13 / 4.10	0.07 / 5.31	0.17 / 4.78	-0.30 / 4.11	-0.09 / 4.33
IWPSNR	-0.25 / 3.88	0.12 / 3.64	-0.05 / 4.26	0.28 / 4.85	0.07 / 4.20	0.25 / 4.04
IWSSIM	-0.60 / 4.48	-0.32 / 4.18	-0.01 / 5.42	0.50 / 5.21	-0.78 / 5.44	-0.20 / 4.96
S-CIELAB	0.00 / 2.78	0.10 / 3.43	0.16 / 2.76	0.22 / 3.72	0.23 / 2.61	0.11 / 3.03

Table 6. The F-test results for all datasets

	PSNR	*WSNR*	*UQI*	*SSIM*	*MSSIM*	*IFC*	*VIF*	*VSNR*	*FSIMc*	*IWPSNR*	*IWSSIM*	*S-CIELAB*
PSNR	- - - - - -	1111-1	111111	111111	111111	111111	-1- - - -	-111-1	1111-1	-1- - -1	111111	- - -0-0
WSNR	0000-0	- - - - - -	111111	- -11- -	- -11- -	111111	- - - - - 0	- - - - - -	- - - - - -	0-0000	111111	000000
UQI	000000	000000	- - - - - -	000000	000000	0-0000	000000	000000	000000	000000	000000	000000
SSIM	000000	- -00- -	111111	- - - - - -	- - - - - -	111111	0-0000	0-0000	- - -000	0-0000	-11- -1	000000
MSSIM	000000	- -00- -	111111	- - - - - -	- - - - - -	111111	- -0000	- -0- - -	- - -0- -	0-0000	111-11	000000
IFC	000000	000000	1-1111	000000	000000	- - - - - -	000000	000000	000000	000000	000000	000000
VIF	-0- - - -	- - - - -1	111111	1-1111	- -1111	111111	- - - - - -	- - -1-1	- -1- -1	- - - - -1	111111	0000-0
VSNR	-000-0	- - - - - -	111111	1-1111	- -1- - -	111111	- - -0-0	- - - - - -	- -1- - -	- -00- -	111111	0000-0
FSIMc	0000-0	- - - - - -	111111	- - -111	- - -1- -	111111	- -0- -0	- -0- - -	- - - - - -	0-0000	111111	000000
IWPSNR	-0- - -0	1-1111	111111	1-1111	1-1111	111111	- - - - -0	- -11- -	1-1111	- - - - - -	111111	-000-0
IWSSIM	000000	000000	111111	-00- -0	000-00	111111	000000	000000	000000	000000	- - - - - -	000000
S-CIELAB	- - -1-1	111111	111111	111111	111111	111111	1111-1	1111-1	111111	-111-1	111111	- - - - - -

sets of residuals from metrics A and B respectively; the F statistic is defined as $F = \frac{\sigma_A^2}{\sigma_B^2}$. If $F > F_{critical}$ ($F < 1/F_{critical}$) then it signifies that at a given confidence level, metric A has significantly larger (smaller) residuals than metric B. The $F_{critical}$ is computed based on the number of residuals and the confidence level [4]. In this study, 95% confidence level was used. The F-test assumes that the set of residuals (prediction errors) are normally distributed. A simple normality test was used based on the rule of thumb that a set of values is normally distributed if its kurtosis and skewness values between 2 to 4, and -1 to 1 respectively [30] (the Normal distribution has a kurtosis of 3 and a skewness of zero). The results of the normality test are given in Table 5.

Table 6 lists the F-test statistics results carried out on the set of residuals of each objective image quality metric for the individual subsets as well as for the full dataset. Each entry in the Table 6 is a codeword of six symbols. The position of the symbol in the code word represents the following datasets (from left to right): Dekker SOM, Median Cut, Kmeans, Octree, Wus algorithm, and all data. Each symbol gives the result of the F-test on the dataset represented by the symbols position. "1" means that the image quality metric from the row is statistically better than the image quality metric from the column, "0 means that it is statistically worse and "-" means that it is statistically indistinguishable. Thus in terms of statistical significance, as expected S-CIELAB is statistically the best performing metric because it is color fidelity based metric followed by the PSNR. Although many studies have shown that the PSNR perform badly in assessing the quality of images. Other studies also have shown that the PSNR have the best performance in assessing the quality of images for different distortions including color quantization distortions [43, 1]. The structure similarity based metrics come in the middle since the quality of a distorted image is evaluated based on how much structure is preserved within the distorted image compared with the reference image. Although color quantization is not in the first place a structural distortion, but reducing the number of colors in an image may result in a distortion of the structure of the quantized image. It is clear also that VIF is significantly better than the structure similarity based metrics while IFC and UQI metrics have the worst performance.

5. CONCLUSION

In this paper, a new image database was presented. This image database can be used to evaluate the performance of image quality metrics. The database consists of 25 reference images, 875 test images produced by five popular color quantization algorithms. The prediction performance of recent state-of-the-art image quality metrics over the new image database was analyzed and reported. The database has been put for downloading freely to the research community to further study in the field of image quality assessment.

6. REFERENCES

[1] I. Avcibas, B. Sankur, and K. Sayood. Statistical evaluation of image quality measures. *Journal of Electronic imaging*, 11(2):206–223, 2002.

[2] J. Braquelaire and L. Brun. Comparison and optimization of methods of color image quantization. *IEEE Transactions on Image Processing*, 6(7):1048–1052, 1997.

[3] P. Le Callet and F. Autrusseau. Subjective quality assessment IRCCyN/IVC database. `http://www.irccyn.ec-nantes.fr/ivcdb/`, 2005.

[4] D. Chandler and S. Hemami. VSNR: A wavelet base visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16(9):2284–2298, 2007.

[5] D. Chandler and S. Hemami. VSNR online supplement. `http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html`, 2007.

[6] A. Dekker. Kohonen neural networks for optimal colour quantization. *Network Computation in Neural Systems*, 5(3):351–367, 1994.

[7] A.M. Eskicioglu and P.S. Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, 1995.

[8] M. Gervautz and W. Purgathofer. A simple method for color quantization: Octree quantization. In *Graphics Gems*, pages 287–293. Academic Press Professional, Inc., San Diego, CA, USA, 1990.

[9] H.S. Han, D.O. Kim, and R.H. Park. Structural information-based image quality assessment using LU factorization. *IEEE Transactions on Consumer Electronics*, 55(1):165–171, 2009.

[10] M. Hassan and C. Bhagvati. Color quantization database. `http://dcis.uohyd.ernet.in/~hassan/Color_Quantization_Database.rar`, 2012.

[11] P. Heckbert. Color image quantization for frame buffer display. In *Proceedings of SIGGRAPH*, volume 16, pages 297–307, 1982.

[12] ITU-R. Methodology for the subjective assessment of the quality for television pictures, 2002. Recommendation ITU-R BT.500-11. Geneva.

[13] D.O. Kim and R.H. Park. Image quality assessment using the amplitude/phase quantization code. *IEEE Transactions on Consumer Electronics*, 56(4):2756–2762, 2010.

[14] D.O. Kim and R.H. Park. Image quality measure using the phase quantization code. *IEEE Transactions on Consumer Electronics*, 56(2):937–945, 2010.

[15] E. Larson and D. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006–011006, 2010.

[16] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer desing. *IEEE Transactions on Communications*, 28(1):84–95, 1980.

[17] A. Mayache, T. Eude, and H. Cherifi. A comparison of image quality models and metrics based on human visual sensitivity. In *Proceedings of International Conference on Image Processing*, pages 409–413, 1998.

[18] C. McCamy. On the number of discernible colors. *Color Research and Application*, 23(5):337–337, 1998.

[19] T. Mitsa and K. Varkur. Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal processing*, pages 301–304, 1993.

[20] M. Miyahara, K. Kotani, and V. Algazi. Objective picture quality scale (PQS) for image coding. *IEEE Transactions on Communications*, 46(9):1215–1226, 1998.

[21] A. Mojsilovic, J. Kovacevic, J. Hu, R. Safranek, and S. Ganapathy. Matching and retrieval based on the vocabulary and grammar of color patterns. *IEEE Transactions on Image Processing*, 9(1):38–54, 2000.

[22] A. Ninassi, P. Le Callet, and F. Autrusseau. Pseudo no reference image quality metric using perceptual data hiding. In *Proceedings of SPIE*, volume 6057, pages 146–157, 2006.

[23] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008 - A database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10:30–45, 2009.

[24] B. Rosner. Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, 25(2):165–172, 1983.

[25] X. Rui, C. Chang, and T. Srikanthan. On the initialization and training methods for kohonen self-organizing feature maps in color image quantization. In *Proceedings of the* 1^{st} *IEEE international workshop on electronic design, test and applications*, pages 321–325, 2002.

[26] Z. Parvez Sazzad, Y. Kawayoke, and Y. Horita. MICT image quality evaluation database. `http://mict.eng.u-toyama.ac.jp/mictdb.html`, 2008.

[27] P. Scheunders. A genetic c-means clustering algorithm applied to color image quantization. *Pattern Recognition*, 30(6):859–866, 1997.

[28] H. Sheikh and A. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.

[29] H. Sheikh, A. Bovik, and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, 2005.

[30] H. Sheikh, M. Sabir, and A. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3441–3452, 2006.

[31] H. Sheikh, Z. Wang, L. Cormack, and A. Bovik. LIVE image quality assessment database release 2. `http://live.ece.utexas.edu/research/quality/subjective.htm`, 2006.

[32] A. Shnayderman, A. Gusev, and A. M. Eskicioglu. An SVD-based gray-scale image quality measure for local and global assessment. *IEEE Transactions on Image Processing*, 15(2):422–429, 2006.

[33] M. Swain and D. Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991.

[34] E. van den Broek, T. Kok, T. Schouten, and L. Vuurpijl. Human-centered content-based image retrieval. In *Proceedings of SPIE*, volume 6806, page 54, 2008.

[35] A. van Dijk, J. Martens, and A. Watson. Quality assessment of coded images using numerical category scaling. In *Proceedings of SPIE*, volume 2451, pages 90–101, 1995.

[36] Z. Wang and A. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9:81–84, 2002.

[37] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[38] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185–1198, 2011.

[39] Z. Wang, E. Simoncelli, and A. Bovik. Multi-scale structural similarity for image quality assessment. In *Proceedings of the* 37^{th} *IEEE Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1398–1402, 2003.

[40] X. Wu. Efficient statistical computations for optimal color quantization. In *Graphics Gems II*, pages 126–133. New York: Academic, James Arvo edition, 1991.

[41] C. Yang and W. Tsai. Color image compression using quantization, thresholding, and edge detection techniques all based on the moment-preserving principle. *Pattern Recognition Letters*, 19:205–215, 1998.

[42] A. Zaric, N. Tatalovic, N. Brajkovic, H. Hlevnjak, M. Loncaric, E. Dumic, and S. Grgic. VCL@FER image quality assessment database. *AUTOMATIKA*, 53(4):344–354, 2012.

[43] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.

[44] X. Zhang and B. Wandell. A spatial extension of CIELAB for digital color image reproduction. In *Proceedings of SID International Symposium Digest of Technical Papers*, volume 27, pages 731–734, 1996.