Structural Similarity Measure for Color Images

Mohammed Hassan
DCIS
University of Hyderabad
Hyderabad-500046, India

Chakravarthy Bhagvati
DCIS
University of Hyderabad
Hyderabad-500046, India

ABSTRACT
Color images reveal more meaningful information to the human observers rather than grayscale ones. Regardless of the advantages of the existing well-known objective image quality measures, one of the common and major limitations of these measures is that they evaluate the quality of grayscale images only and don't make use of color information. In this paper we propose an improved method for image quality assessment that adds a color comparison to the criteria of the well-known Multiscale Structural Similarity index (MSSIM). We evaluated the new color image quality measure through human subjective experiments. Our human subjective evaluation data contains 25 reference images and 875 test images produced by five popular color quantization algorithms. Each of the quantized images was evaluated by twenty two subjects and more than 19200 individual human quality judgments were carried out to obtain the final mean opinion scores. We also tested the proposed method on TID2008 image database to further verify our results. These results indicate that adding color comparison improves MSSIM for many distortions in TID2008 and for assessing quantized images in our database.

Keywords
Image quality assessment, Structural similarity index, Color quantization.

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 involve a costly process which requires a large number of observers and takes significant time. On the other hand the objective image quality assessment methods are computer based methods that can automatically predict the perceived image quality. Hence the objective image quality assessment methods gained more popularity although they do not necessarily correlate well with the quality as perceived by humans [1, 2]

Objective image quality assessment methods also may be classified into full reference, reduced reference, and no reference methods based on the availability of the reference image. Full reference image quality assessment requires complete information about the reference image; and partial information about the reference image is required for the reduced reference image quality assessment; while no information about the reference image is needed in no reference image quality assessment. This paper focuses on the full reference image quality assessment methods for color images where both the original and the test images are available.

Many researchers have contributed significantly in the design of objective image quality methods starting from the widely used mean square error (MSE) metric and its correlated peak signal to noise ratio (PSNR). The weighted signal to noise ratio (WSNR) [3] simulates the human visual system properties by filtering both the reference and distorted images with contrast sensitivity function and then compute the SNR. Miyahara [4] proposed a picture quality scale (PQS) based on three distortion factors; the amount, location and structure of error. Wang and Bovik [5] proposed a new universal image quality index (UQI) and its improved form the single-scale structural similarity index (SSIM) [6] by modeling the image distortion as a combination of loss of luminance, contrast, and correlation. In [7] the single-scale structural similarity index was extended to a multi-scale structural similarity index (MSSIM) that achieved better results than SSIM. Information fidelity criterion (IFC) [8] and visual information fidelity (VIF) [9] 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 [10] explored the feasibility of singular value decomposition (SVD) for quality measurement. In [11] a two staged wavelet based visual signal to noise ratio (VSNR) was proposed based on the low-level and the mid-level properties of human vision.

It is known that the number of colors that can be distinguished by human eye is much more than that of the grayscale levels and more information is contained in color image than what is contained in grayscale one, but in the traditional image quality assessment methods the assessment of color image is always performed by assessing its luminance layer or its grayscale conversion therefore color information in the image is largely ignored and precision of assessment result is influenced accordingly. One of the cases where color plays an important role in determining the quality of the color image is color quantization distortion. In color image quantization only a small set of colors (usually 256 or less [12]) is used to represent the full set of colors in the color image that may contain more than sixteen millions different colors. This paper proposes an extension to MSSIM for color images and explores whether the extension correlates well with human perception of color image quality.

Fig 1: Some of the reference images used in the study

The paper is organized as follows. In section 2, the grayscale SSIM is simply introduced. Section 3 describes the proposed improvement to the grayscale SSIM. In Section 4 the generation of the experimental human subjective evaluation data is presented. Section 5 discusses the results. Finally, section 6 draws the conclusion.

2. THE GRAYSCALE STRUCTURAL SIMILARITY INDEX

Wang [6] proposed the structural similarity index (SSIM) based on the assumption that the HVS is highly adapted to extract structural information from visual scenes. Structural similarity index consists of three local comparison functions namely luminance comparison, contrast comparison, and structure comparison between two signals *x* and *y*:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \qquad (1)$$

$$C(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \qquad (2)$$

$$S(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \qquad (3)$$

Where μ_x and μ_y are the sample means of *x* and *y*, respectively, σ_x and σ_y are the sample standard deviations of *x* and *y*, respectively, and σ_{xy} is the sample correlation coefficient between *x* and *y*. The constants C_1, C_2, and C_2 are used to stabilize the algorithm when the denominators approach to zero. These statistics are calculated within a local window.

Then the general form of the SSIM index is given by combining the three comparison functions:

$$SSIM = l(x,y)^{\alpha} \cdot C(x,y)^{\beta} \cdot S(x,y)^{\gamma} \quad (4)$$

Where α, β and γ are parameters which define the relative importance of the three components.

Usually, $\alpha = \beta = \gamma = 1$ and $C_1=C_2$ yielding

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (5)$$

In [7] Wang introduced an improved version of SSIM index that is accomplished over multiple scales of the reference and distorted images in which the contrast and the structure comparisons are calculated at each scale, and the luminance comparison is computed only at highest scale *M*=5.

$$MSSIM = [l_M(x,y)]^{\alpha_M} \prod_{i=1}^{M} [C_i(x,y)]^{\beta_i} \cdot [S_i(x,y)]^{\gamma_i} \quad (6)$$

The parameters α_i, β_i and γ_i are selected such that $\alpha_i = \beta_i = \gamma_i$ and $\sum_{i=1}^{M} \gamma_i = 1$.

3. THE COLOR IMAGE QUALITY MEASURE

Color images reveal more meaningful information to the human observers rather than grayscale ones; based on this fact we developed an improved multiscale structural similarity index that adds a color comparison to the criteria of the grayscale MSSIM.

The CIELAB color space is designed to improve the organization of colors that are not uniform in a linear color space so that Euclidean distances between different colors in the CIELAB correspond approximately to perceived color differences. A useful rule of thumb in CIELAB color space is that any two colors can be distinguished if the Euclidean distance between these two colors is greater than threshold value of 3 [13]. This threshold is known as the Just Noticeable Color Difference (JNCD) threshold. Therefore all the colors within a sphere of radius equal to the JNCD threshold are perceptually indistinguishable from each other.

The Color comparison can be obtained by the following steps:

- First, the reference and test images are preprocessed by spatial filtering to simulate the spatial blurring by the human visual system in a way that the filtering operation to the color image affects only the fine patterned colors [14].
- Then, the reference and test images are transformed to the CIELAB color space.
- After that the color comparison is computed by averaging the number of colors in the reference image that are undistinguishable from their corresponding colors in the test image based on the JNCD threshold 3.

The general form of the proposed image quality measure is given as:

$$CMSSIM = [Clr(x,y)]^{\delta} [l_M(x,y)]^{\alpha_M} \prod_{i=1}^{M} [C_i(x,y)]^{\beta_i} \cdot [S_i(x,y)]^{\gamma_i} \quad (7)$$

Where the contrast and the structure comparisons *C* and *S* are calculated at each scale, the luminance comparison *l* is computed only at highest scale, and the color comparison *Clr* is calculated at the lowest scale only. $\beta_1 = \gamma_1 = 0.04448$ $\beta_2 = \gamma_2 = 0.2856$ $\beta_3 = \gamma_3 = 0.3001$ $\beta_4 = \gamma_4 = 0.2363$ $\alpha_5 = \beta_5 = \gamma_5 = 0.1333$, and the optimal value of δ is 0.7.

4. SUBJECTIVE EXPERMENT SETUP

Our human subjective evaluation data [15] contains 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 some of the reference images used in the study. All images in our 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 popular color image quantization algorithms namely: Kmeans algorithm [16], Median Cut algorithm [17], Wu's Algorithm [18], Octree [19], and Dekker's SOM [20].

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. In our tests we followed the recommendations given by ITU [21] that define how to carry out subjective quality tests. A group of twenty two undergraduate students participated in our 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.

Before carrying out the experiments the observers were briefly explained 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.

Since fidelity of the quantized images to the reference images has to be evaluated, simultaneous double stimulus for continuous evaluation (SDSCE) method [21] was used in conducting the psychometric experiment where a set of subjects is watching simultaneously the two images (the reference and the quantized images). The observers are asked to assess the overall quantized image quality with respect to the reference image of each presentation by simply dragging a slider on a quality scale. The quality scale which is of range [0,100] was labeled and divided into five equal categories: "Bad," "Poor," "Fair," "Good," and "Excellent." The position of the slider reflects the rate given by the observer for that image and its position was reset after each presentation.

There are 875 test images and each session should not last more than 30 minutes [21], therefore the overall subjective tests were divided into six 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 to the rating process. Subjects were shown images in a random order and this order is unique for each subject.

Before starting the analysis of the data, a screening of the subjective raw values was conducted [21] to eliminate observers with unstable values. The generalized ESD [22] many-outlier procedure 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 subjective raw data was considered as outlier and one of the observers was rejected.

To calculate Mean Opinion Scores (MOS), the subjective raw data is first converted to Z-score (after outlier 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 [23]:

$$z_{ij} = \frac{\left(v_{ij} - \overline{v_i}\right)}{\sigma_i} \qquad (8)$$

Where v_{ij} is the raw values 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 values over all 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.

5. RESULTS AND DISCUSSIONS

In this section, the performance of the proposed image quality measure in terms of the ability of predicting the subjective ratings is analyzed. The proposed quality measure was applied to the set of images used in the psychometric experiment and the results were compared to the subjective MOS. For comparison, the same set of images were presented to five well-known objective image quality measures that are commonly used and their implementations are publicly available on the Internet namely: Peak Noise to signal Ratio (PNSR), Structural Similarity Index (SSIM), Multiscale Structural Similarity Index (MSSIM), Visual Information Fidelity (VIF), and Visual Signal to Noise Ratio (VSNR).

The scores given by the objective image quality measures are transferred into a predicted MOS to map the objective image quality measures' scores into the range of the subjective MOS and to remove any nonlinearity between them using nonlinear regression [21]. The function chosen for regression is a four parameters logistic function [24]:

$$MOS_p(Q) = \frac{p_1 - p_2}{1 + \exp\left(\frac{Q - p_3}{p_4}\right)} + p_2 \qquad (9)$$

Where MOS_p is the predicted MOS, *Q* is the quality rating given by the measure, and p_1, p_2, p_3, and p_4 are parameters.

In order to show the advantages of the proposed quality measure over the traditional MSSIM and the other quality measures included for comparison reason, the performance of each quality measure was quantified based on Pearson Linear Correlation Coefficient (PLCC) as an indication on correlation, the Root Mean Square Error (RMSE) to measure the prediction accuracy, and the Spearman Rank Order Correlation Coefficient (SROCC) to measure prediction monotonicity.

Tables I-III show the testing results of scores given by the six image quality measures included in this study for comparison (after nonlinear mapping) with the mean opinion scores (MOS) obtained from the psychometric experiments for the complete set of images as well as for separate subsets of the images. It is clear from the tables that the proposed method CMSSIM greatly improves the performance of traditional

Table I. PLCC between MOS and IQM's ratings after nonlinear mapping

	PSNR	***VIF***	***VSNR***	***SSIM***	***MSSIM***	***CMSSIM***
SOM	0.956	0.950	0.949	0.929	0.935	**0.959**
Median	0.965	0.938	0.929	0.940	0.934	0.962
Kmeans	0.960	0.951	0.943	0.911	0.91	**0.969**
Octree	0.970	0.967	0.955	0.935	0.944	**0.977**
Wu's algo	0.957	0.957	0.953	0.930	0.94	**0.959**
ALL Data	0.945	0.942	0.926	0.913	0.917	**0.960**

Table II. SROCC between MOS and IQM's ratings after nonlinear mapping

	PSNR	***VIF***	***VSNR***	***SSIM***	***MSSIM***	***CMSSIM***
SOM	0.950	0.945	0.944	0.921	0.934	**0.955**
Median	0.961	0.936	0.925	0.938	0.931	0.957
Kmeans	0.952	0.946	0.938	0.909	0.912	**0.959**
Octree	0.965	0.962	0.954	0.934	0.95	**0.977**
Wu's algo	0.953	0.955	0.952	0.929	0.939	**0.956**
ALL Data	0.939	0.938	0.923	0.911	0.918	**0.954**

Table III. RMSE

	PSNR	***VIF***	***VSNR***	***SSIM***	***MSSIM***	***CMSSIM***
SOM	8.540	9.127	9.168	10.831	10.334	**8.264**
Median	8.004	10.609	11.372	10.490	10.97	8.395
Kmeans	8.678	9.586	10.304	12.787	12.862	**7.642**
Octree	7.008	7.409	8.629	10.302	9.532	**6.158**
Wu's algo	8.815	8.815	9.187	11.176	10.39	**8.646**
ALL Data	9.774	10.100	11.289	12.209	11.933	**8.421**

Table IV. The Pearson's linear correlation coefficients (PLCC) and the Spearman's rank order correlation coefficient (SROCC) for the TID2008 image database

	PLCC			***SROCC***		
	SSIM	***MSSIM***	***CMSSIM***	***SSIM***	***MSSIM***	***CMSSIM***
Additive Gaussian noise	0.771	0.748	**0.919**	0.797	0.81	**0.917**
Additive noise in color components	0.797	0.778	**0.926**	0.811	0.806	**0.916**
Spatially correlated noise	0.788	0.76	**0.863**	0.827	0.819	**0.869**
High frequency noise	0.851	0.822	**0.957**	0.843	0.868	**0.939**
Impulse noise	0.717	0.625	**0.874**	0.746	0.687	**0.902**
Quantization noise	0.740	0.757	**0.853**	0.803	0.**854**	0.849
Image denoising	0.903	0.915	**0.958**	0.927	**0.957**	0.950
JPEG compression	0.902	0.931	**0.943**	0.899	0.935	**0.936**
JPEG2000 compression	0.858	0.939	**0.944**	0.887	**0.973**	0.945
JPEG transmission errors	0.816	0.824	**0.864**	0.819	**0.874**	0.864
JPEG2000 transmission errors	0.810	0.788	**0.830**	0.840	0.853	**0.859**
Non eccentricity pattern noise	0.668	0.665	**0.724**	0.695	0.733	**0.763**

MSSIM and also outperformed the other image quality measures.

We also used the popular Tampere Image Database (TID 2008) [25] to further test the performance of CMSSIM quality measure. This database is the most recent and largest database so far available that includes more images and more distortion types for verification of full reference quality metrics. The TID2008 database contains 1700 distorted images (25 reference images × 17 types of distortions × 4 levels of distortions). Mean Opinion Scores for this database have been obtained as a result of 838 subjective experiments. During these tests, observers from three countries (Finland, Italy, and Ukraine) have carried out about 256000 individual human quality judgments. Table IV shows the Pearson's linear correlation coefficients (PLCC) and the Spearman's rank order correlation coefficients (SROCC) between the MOS form the TID2008 database and the scores given by proposed CMSSIM as well as for the state-of-the-art SSIM and MSSIM indexes for some distortion types.

In terms of linear correlation, it is clear from Table IV that the CMSSIM measure greatly improves the performance of MSSIM and SSIM in all distortion types. For the Spearman Rank Order Correlation Coefficient we can see from Table IV that CMSSIM outperforms MSSIM and SSIM in most of distortion types and competitive in the rest few types. It may be noted from Table IV that CMSSIM has a competitive performance for the Spearman's rank order correlation coefficient with MSSIM for Quantization noise on TID2008 image database white it has better performance on our psychometric data this due to the number of distorted images used; where there are only one hundred quantized images with four quantization levels created by single color quantization algorithm on the TID2008 database while our psychometric experiment contains 875 quantized images with seven quantization levels produced by five color quantization algorithms. This number of quantized images is large enough to distinguish between the performances of the two quality measures.

6. CONCLUSION

In this paper, we present an improvement to the well-known Multi-Scale Structural Similarity index (MSSIM) by adding a color comparison to the criteria of the grayscale MSSIM. The new image quality measure fully uses the color information of the image for the assessment of color distortions that are difficult to be noticed using the luminance channel only or grayscale conversion of the color image. Results show that the proposed quality measure provides results that are more consistent with human perception of color image quality assessment and also greatly improves the performance of MSSIM on many distortion types.

7. REFERENCES

[1] Ouni, S., Chambah, M., Herbin, M., and Zagrouba, E., 2008. Are Existing Procedures Enough? Image and Video Quality Assessment: Review of Subjective and Objective Metrics. Electronic Imaging, Image Quality and System Performance, SPIE, San Jose, CA, USA.Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[2] Wang, Z., Bovik, A. C., and Lu, L., 2002. Why is image quality assessment so difficult?. IEEE International Conference on Acoustics, Speech, & Signal Processing, 4, 3313-3316.

[3] Mitsa, T., and Varkur, K.., 1993. Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms. IEEE International Conference on Acoustic, Speech, and Signal processing, 5, 301- 304.

[4] Miyahara, M., Kotani, K., Algazi, V.R., 1998. Objective Picture Quality Scale (PQS) for image coding. IEEE Transactions on Communications, 46(9), 1215-1226.

[5] Wang, Z., Bovik, A.C., 2002. A universal image quality index. IEEE Signal Processing Letters, 9(3),81–84.

[6] Wang, Z., Bovik, A.C., Sheikh, H.R., and Simoncelli, E.P., 2004. Image quality assessment: From error measurement to structural similarity. IEEE Transaction on Image Processing, 13 (4), 600-612.

[7] Wang, Z., Simoncelli, E.P., and Bovik, A.C., 2003. Multiscale structural similarity for image quality assessment. 37th IEEE Asilomar Conference on Signals, Systems, and Computers, 2, 1398- 1402.

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

[9] Sheikh, H. R., and Bovik, A. C., 2006. Image Information and Visual Quality. IEEE Transactions on Image Processing, 15(2), 430-444.

[10] Shnayderman, A., Gusev, A., and Eskicioglu, A.M., 2006. An SVD-Based Gray-Scale Image Quality Measure for Local and Global Assessment. IEEE Transaction on Image Processing, 15(2), 422-429.

[11] Chandler, D.M., Hemami, S. S., 2007. VSNR: A Wavelet base Visual Signal-to-Noise Ratio for Natural Images. IEEE Transaction on Image Processing, 16(9), 2284–2298.

[12] Scheunders, P., 1997. A genetic C-means clustering algorithm applied to color image quantization. Pattern Recognition, 30(6), 859-866.

[13] Mahy, M., Van Eycken, L., and Oosterlinck, A., 1994. Evaluation of uniform color spaces developed after the adoption of CIELAB and CIELUV. Color research and application, 19(2), 105–121.

[14] Zhang, X., Wandell, B. A., 1996. A spatial extension of CIELAB for digital color image reproduction. Proceedings of the SID Symposium Technical Digest, 27, 731-734.

[15] Color Quantization Database. Available: http://dcis.uohyd.ernet.in/~hassan/Color_Quantization_Database.rar.

[16] Lloyd, S. P., 1982. Least squares quantization in PCM. IEEE Transactions on Information Theory, 28 (2): 129–137.

[17] Heckbert, P., 1982. Color image quantization for frame buffer display. ACM Trans. Computer Graphics (SIGGRAPH), 16 (3), 297–307.

[18] Wu, X., 1991. Efficient statistical computations for optimal color quantization. Graphics Gems, 11, J. Arvo, Ed. New York: Academic, 126-133.

[19] Gervautz, M., and Purgathofer, W., 1988. A simple Method for Color Quantization: Octree Quantization. New Trends in Computer Graphics, Springer Verlag, Berlin. 219-231.

[20] Dekker, A. H., 1994. Kohonen neural networks for optimal colour quantization. Network Computation in Neural Systems, 5(3), 351-367.

[21] ITU-R, 2002. Methodology for the Subjective Assessment of the Quality for Television Pictures, Recommendation ITU-R BT.500-11. Geneva.

[22] Rosner, B., 1983. Percentage points for a generalized ESD many-Outlier Procedure. Technometrics, 25(2), 165–172.

[23] Van Dijk, A. M, Martens, J-B, Watson, A. B., 1995. Quality assessment of coded images using numerical category scaling. Proc. SPIE, 2451, 90–101.

[24] Sheikh, H. R., Sabir, M. F., and Bovik, A. C., 2006. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. IEEE Transactions on Image Processing, 15(11), 3441-3452.

[25] Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., and Battisti, F., 2009. TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics. Advances of Modern Radio electronics, 10, 30-45.