Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 14 (1): 18-25, 2014
ISSN 1812-5654 / DOI: 10.3923/jas.2014.18.25

A Statistical Evaluation of Image Quality Analyzer for Assessment of Histogram Equalization-based Contrast Enhancement Methods

Soong-Der Chen
Department of Graphics and Multimedia, College of Information Technology, Universiti Tenaga Nasional, Kajang, 43000, Selangor D.E, Malaysia

Abstract: Absolute Mean Brightness Error (AMBE) and Entropy are two popular Image Quality Analyzers (IQA) used for assessment of Histogram Equalization (HE)-based contrast enhancement methods. This study highlights their shortcomings in quality assessment of contrast enhanced images. The IQAs were evaluated using human judgment scores derived from 1935 "ground truth" data. The results showed that they have poor correlation to human judgment with Pearson Correlation Coefficient (PCC) <0.4. Their performance is not even comparable to Multi-Scale Structural Similarity Index Metric (MSSIM), an IQA originally designed for assessment of image compression algorithm. The results strongly suggest that there is an urgent need to design new IQA specifically for assessment of contrast enhancement.

Key words: Histogram equalization, contrast enhancement, quality assessment, quality metric, distortions, visual perception

INTRODUCTION

Histogram Equalization (HE) is among the most commonly used contrast enhancement method for medical and radar imaging but it is seldom used in consumer electronics such as digital camera. This is due to the fact that HE tends to cause the following problems: Excessive brightness change, noise amplification and loss of details. Many variants of HE have been proposed to overcome the above mentioned problem. They can be broadly classified into two categories:

- Automatic-human intervention not required in the process of enhancement (Kim, 1997; Wongsritong *et al.*, 1998; Wang *et al.*, 1999; Chen and Ramli, 2003a; Wang and Ye, 2005; Ibrahim and Kong, 2007; Menotti *et al.*, 2007; Kim and Chung, 2008; Ooi *et al.*, 2009; Ooi and Isa, 2010a; Yun *et al.*, 2010; Sengee *et al.*, 2010)
- Adjustable-user can interactively regulate the degree of enhancement by altering parameter's value (Chen and Ramli, 2003b; Abdullah-Al-Wadud *et al.*, 2007; Wang and Ward, 2007; Chen and Suleiman, 2008; Ibrahim and Kong, 2009; Arici *et al.*, 2009; Sheet *et al.*, 2010; Ooi and Isa, 2010b)

This study focuses only on automatic methods because for consumer electronics, it is desirable to enhance image's contrast without human intervention. Therefore, adjustable methods are left beyond the scope of this study.

Although, all automatic methods are designed to overcome the problem of distortion, the extent to which they are resilient to distortions remains questionable. In fact Der and Sidhu (2009) have reported that the automatic methods proposed in references Kim (1997), Wongsritong *et al.* (1998), Wang *et al.* (1999) and Chen and Ramli (2003a) are not resilient to noise. This study aims to review the existing Image Quality Analyzers (IQA) used to assess HE-based methods.

REVIEW OF IQAs

Table 1 lists the available automatic HE-based methods together with the IQAs that have been used to evaluate them. Absolute Mean Brightness Error (AMBE) and Entropy appear to be the two most frequent used analyzers.

AMBE: AMBE is the abbreviation of Absolute Mean Brightness Error. It is the absolute difference between the mean of input and output image. It is formally defined by Eq. 1:

$$\text{AMBE} = |\,E(X) - E(Y)\,| \tag{1}$$

where, X and Y denote the input and output image, respectively and E(.) denotes the expected value, i.e., the statistical mean. Equation 1 clearly shows that AMBE is designed to detect one of the distortions-excessive brightness change. AMBE was proposed by Chen and Ramli (2003b) to evaluate the performance in preserving

Table 1: List of automatic HE-based methods and their IQA(s)

Methods	IQA
Brightness preserving Bi-HE (BBHE)	- AMBE
Multi-peak HE (Multi-peak)	- AMBE
Equal area dualistic sub-image HE (DSIHE)	- AMBE - Entropy - Background brightness
Minimum mean brightness error Bi-HE (MMBEBHE)	- AMBE
Brightness preserving histogram equalization with maximum entropy (BPHEME)	- AMBE - Entropy
Brightness preserving dynamic histogram Equalization (BPDHE)	- AMBE - Entropy
Multi-histogram equalization methods for Contrast enhancement and brightness Preserving (Multi-HE)	- AMBE
Recursively separated and weighted histogram Equalization for brightness preservation and Contrast enhancement (RSWHE)	- AMBE - Entropy - PSNR
Bi-histogram equalization with a plateau Limit For digital image enhancement (BHEPL)	- Average AMBE
Adaptive contrast enhancement methods with brightness preserving (DQHEPL and BHEPL-D)	- Average AMBE - Average entropy - Average PSNR
Fusion framework of histogram equalization and laplacian pyramid (FFHELP)	- Standard deviation of AMBE - Enhancement by entropy
Image contrast enhancement using Bi-histogram equalization with neighborhood metrics (BHENM)	- Average AMBE

original brightness. The idea of preserving brightness was originated by the author of BBHE, who suggested that the fundamental reason HE could produce undesirable distortions is because, it does not take the mean brightness of an image into account. Consequently, lower AMBE is associated with better performance in preserving brightness and hence, better quality of output image.

However, AMBE does not take in account the problem of noise so it alone may not be comprehensive to gauge the overall quality of an image. Figure 1 and 2 shows two versions of image Plane of different AMBE. There is no noise observed in Fig. 1 despite having a much higher AMBE (21.81) as compared to Fig. 2 which has lower AMBE (1.53) but clearly shows the presence of noise (false contour in the background). Based on this observation, this study suggests that such weakness of AMBE in detecting the presence of noise could be the reason why automatic methods which are designed to be brightness-preserving but later found not resilient to noise.

Entropy: The entropy here refers to the Shannon Entropy. It measures the uncertainty of a random variable. It quantifies the expected value of the information contained in an information source. Entropy is typically measured in bits. It is formally defined by Eq. 2:

$$H(X) = \sum_{i=1}^{n} p(x_i)I(x_i) = -\sum_{i=1}^{n} p(x_i)\log_b I(x_i) \qquad (2)$$

Fig. 1: Modified image of plane, AMBE = 21.81

Fig. 2: Modified image of plane, AMBE = 1.53

Where:

X	=	Image
x_i	=	Level i
$p(x_i)$	=	Probability of level i
b	=	Units, (image pixel is coded in bit, so b = 2)
n	=	No. of levels

Theoretically, the higher the entropy, the more information is available from the information source. HE is designed to maximize the entropy of an image by remapping the gray levels using the gray levels' probability density function such that they are distributed uniformly. It is assumed that by increasing the entropy, the image could reveal more information. Consequently, an image with higher entropy is regarded to have better quality.

For global gray-level transformation, remapping gray levels using their probability density to obtain uniform distribution can only be achieved if the data is in continuous (non-discrete) form. In discrete form, the mapping using probability density function which is always monotonic can never increase the entropy. Furthermore, HE tends to combine gray levels of relatively

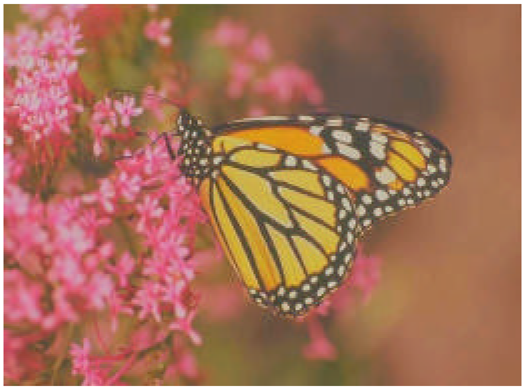

Fig. 3: Modified image of Monarch, Entropy = 6.23

Fig. 4: Modified image of Monarch, Entropy = 6.20

low probability density and results in decrease of entropy despite the fact that such action tends to increase the contrast of an image. Figure 3 and 4 shows two modified images of Monarch with slightly different entropy. Despite having lower entropy (6.20), Fig. 4 clearly shows better contrast than Fig. 3 with higher entropy (6.23).

EVALUATION OF IQAS' CORRELATION TO HUMAN JUDGMENT

Test images: A set of 9 source images of diverse image content was selected from "Lossless True Color Image Suite" provided by Rich Franzen and "LIVE Image Quality Assessment Database" provided by Laboratory of Image and Video Engineering at University of Texas, Austin. Figure 5 shows these source images. These images were preprocessed to simulate poor contrast image which shows distortions after contrast enhancement using HE-based techniques as follows:

- Original images were JPEG compressed at quality Q = 50 using MATLAB's imwrite function
- The JPEG compressed images were contrast-reduced using MATLAB's imadjust function with [low_out hight_out] = [0.2 0.8]

Each contrast-reduced image was then processed using SGHESE (Chen and Ramli, 2003a) with different parameters' value to generate one stimulus for each of the following distortion level (VQEG, 2003):

Fig. 5(a-i): The 9 source images and their spatial resolution (height×width), (a) Caps (512×768), (b) Flower (512×768), (c) Sailboats (720×480), (d) Plane (512×768), (e) Parrots (512×768), (f) Buildings (505×634), (g) Lighthouse (512×768), (h) Monarch (512×768) and (i) Carnival (488×610)

- Very annoying
- Annoying
- Slightly annoying
- Perceptible but not annoying
- Imperceptible

A total of 43 stimuli were generated. The parameters' values were chosen by human expert by interacting with graphical user interface as shown in Fig. 6. Table 2 list the chosen parameters' value. Notice that the higher the Max (Maximum Stretch) values, the more annoying the distortion is. In the case of Max = 1, the output image is identical to the output image of conventional HE.

Test procedures: The original and distorted images were displayed on two separate screens placed next to each other, both connected to a PC that support extended display. The screens used were both 17 liquid-crystal display screen, with a native resolution of 1920×1080 pixels and a screen refresh rate of 60 Hz. The experiment was conducted in a standard office environment and the viewing distance was approximately 60 cm. The experiment was designed based on double-stimulus method. Subjects rated the level of distortion for each stimulus by comparing it with the corresponding contrast-reduced image. The scoring scale ranged from 0 to 100, where "0-20" means "Very

Table 2: List of parameters used to generate stimuli

	Imperceptible		Perceptible but not annoying		Slightly annoying		Annoying		Very annoying	
Parameters	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
Caps	-	-	0.01	0.2000	0.01	0.7764	0.01	0.7764	0.01	1
Carnival	0.01	0.10	0.01	0.1700	0.01	0.3000	0.01	0.5000	0.05	1
Plane	0.01	0.03	0.01	0.0500	0.01	0.1600	0.01	0.3098	0.01	1
Lighthouse	-	-	0.01	0.1200	0.01	0.2300	0.01	0.6000	0.20	1
Flower	0.01	0.10	0.01	0.2025	0.01	0.3544	0.01	0.6667	0.60	1
Parrots	0.01	0.10	0.01	0.3200	0.01	0.5000	0.01	1.0000	0.88	1
Sailboats	0.01	0.12	0.01	0.1800	0.01	0.2800	0.01	0.5000	0.01	1
Monarch	0.01	0.17	0.01	0.2200	0.01	0.3200	0.01	0.6000	0.01	1
Buildings	0.01	0.10	0.01	0.3000	0.01	0.5500	0.01	0.8300	0.57	1

Fig. 6: Graphical user interface of SGHESE

Annoying", "21-40" means "Annoying", "81-100" means "Imperceptible" and so on.

The subjects participated in the experiment were Bachelor of IT students from Universiti Tenaga Nasional. The 45 students (35 male and 10 female) were inexperience with image quality assessment and distortions. They were briefed about the objective and procedures of the experiment. No training session were given as that could influence subject's opinion. After briefing, the stimuli were presented to subject in a random order and the ratings were recorded in a score sheet.

Raw data processing: Simple outlier detection and subject rejection procedures were carried out on the raw scores before the actual data analysis. Raw score for an image was considered to be an outlier if it was outside an interval of 2.33 standard deviations about the mean score for that image (VQEG, 2003). Also, all scores of a subject were rejected if more than 6 of his scores were outliers. Overall, only 1 subject was rejected and only less than 4% of the total scores were rejected.

In order to compute the MOS scores, the raw scores were first converted into Z scores after outlier removal and subject rejection. The Z scores for ith subject and jth image is as defined in Eq. 3:

$$Z_{ij} = \frac{\left(S_{ij} - \overline{S_i}\right)}{\sigma_i} \tag{3}$$

Where:

S_{ij} = The raw score for ith subject and jth image
$\overline{S_i}$ = The average of all the scores rated by subject i
σ_i = The standard deviation all the scores rated by subject i

The Z scores were then averaged across subjects to yield a Mean Opinion Score (MOS) for the jth image as defined in Eq. 4:

$$MOS_j = \frac{1}{S}\sum_{i=1}^{S} Z_{ij} \tag{4}$$

where, S is the total number of subjects after subjects rejection.

EVALUATION RESULTS

Performance metrics: According to the recommendations from VQEG (2003), the performance of an IQA can be quantitatively evaluated with respect to its ability to predict subjective quality rating in the following three aspects.

Prediction accuracy: The ability to predict the subjective quality score with low error. The metrics used were:

- Pearson Correlation Coefficient (Pearson CC)
- Root Mean Squared Error (RMSE)

Prediction monotonicity: The degree to which the model's prediction agrees with the relative magnitudes of the subjective quality rating. The metric used was Spearman Rank Order Correlation Coefficient (SROCC).

Predication consistency: The degree to which the model maintains prediction accuracy over different types of images and not to fail excessively for a subset of images. The metric used were Outlier ratio (OR-ratio of outlier to total scores). Outlier score is score outside an interval of two times the standard deviation about the MOS.

The evaluation was done using MOS after non-linear regression using a five-parameter logistic function (a logistic function with an added linear term, constrained to be monotonic) (Sheikh *et al.*, 2006) as defined in Eq. 5:

$$R(x) = b_1\left(\frac{1}{2} - \frac{1}{1+e^{[b_2(x-b_3)]}}\right) + b_4 x + b_5 \tag{5}$$

This nonlinearity was applied to the MOS or its logarithm which ever gave a better fit for all data.

Statistical significance and hypothesis testing: A hypothesis testing is conducted to determine if the differences between the analyzers' performance are statistically significant. The test aims to determine if one can make a statistically sound conclusion on the superiority or inferiority of an IQA based on a given confidence interval and number of sample.

A variance-based hypothesis test was conducted. It uses F-statistic to compare the variance of two sets of samples. The conclusion whether the two sets of samples come from same distribution or not is made based on the ratio of the variances (standard deviations) called F-ratio. The higher the F-ratio, the more unlikely the two sets of samples are from same distribution. The p-value of an F-ratio is the probability one repeats the same experiment and gets an equal or higher F-ratio while the two sets of samples are in fact from same distribution.

The hypothesis test used samples consist of residual values (e) between MOS and the predicted MOS (MOS_P-Score from IQA after non-linear regression) as defined in Eq. 6:

$$e = |MOS - MOS_P| \tag{6}$$

The Null Hypothesis is that the residual values of one analyzer come from same distribution and are statistically indistinguishable from the residual values from another IQA, within a given confidence interval.

A commonly used confidence interval is 95%. Using confidence interval 95%, p-value below 0.05 indicates that there is significant different in the performance of the two IQAs in study.

Since, variance-based hypothesis test relies on assumption that the samples are normally distributed, it is crucial to perform sample normality test.

DISCUSSION

In this experiment, Multi-Scale Structural Similarity Index Metric (MSSIM) (Wang *et al*., 2003), BLind Image Integrity Notator using DCT Statistics (BLIINDS2) (Saad *et al*., 2010), Distortion Identification-based Image Verity and Integrity Evaluation (DIIVINE) index (Moorthy and Bovik, 2011) were also evaluated besides AMBE and Entropy. MSSIM is an IQA designed to analyzer image fidelity in image compression. BLIINDS2 and DIIVINE are general purpose no-reference IQA. Table 3 show the results obtained and based on the interpretation of the correlation values (Table 4) which are widely used in many scientific journal, it is observed that:

- In terms of Pearson CC, AMBE, Entropy, BLIINDS2 and DIIVINE have poor (<0.4) correlation with HVP while MSSIM has good (>0.76) correlation with HVP
- In terms of SROCC, AMBE, BLIINDS2 and DIIVINE has poor (<0.4) correlation with HVP while Entropy and MSSIM has good (>0.76) correlation with HVP
- In general, RMSE and OR readings are consistent with the readings of Pearson CC and SROCC; the better the correlation, the lesser the error and outlier ratio

This experiment follows the recommendation from VQEG (2003) to use Kurtosis-based sample normality test. Distribution with Kurtosis between 2 and 4 is considered to be normally distributed. Table 5 shows the kurtosis of the residual values from each analyzer and they range from 2.4016-3.8093. In other words, the residual values are normally distributed and Variance-based Hypothesis Testing can be used for statistical hypothesis testing.

Table 6 lists the p-values of the hypothesis test for each pair of IQA. The results in Table 6 show that with 95% confidence interval:

- There is no significant difference (p-value>0.05) in the performance of AMBE, Entropy, BLIINDS2 and DIIVINE
- MSSIM is significantly better than AMBE, Entropy and DIIVINE

Table 3: The results of Pearson CC, RMSE, SROCC and OR for AMBE, Entropy, BLIINDS2, DIIVINE and MSSIM

Parameters	Pearson CC	RMSE	SROCC	OR
AMBE $\|E(X)-H(Y)\|$	0.1346	0.7809	0.0802	0.2791
Entropy $\{H(X)-H(Y)\}\div H(X)$	0.3291	0.7510	0.7682	0.2093
BLIINDS2 BLIINDS2(X)-BLIINDS2(Y)	0.3463	0.7475	0.2976	0.2093
DIIVINE DIIVINE(X)-DIIVINE(Y)	0.3837	0.7279	0.3013	0.2093
MSSIM	0.7174	0.5490	0.7628	0.1395

Table 4: Interpretation of correlation values

Value of correlation	Interpretation
0.00-0.40	Poor
0.41-0.75	Fair
0.76-0.85	Good
0.86-1.00	Excellent

Table 5: Kurtosis of residual values

AMBE	Entropy	MSSIM	BLIINDS2	DIIVINE
3.2364	2.725	3.8093	3.7728	2.4016

Table 6: Results of variance-based hypothesis testing, p-value

Parameters	AMBE	Entropy	MSSIM	BLIINDS2	DIIVINE
AMBE	-	0.6788	0.0085	0.4505	0.5983
Entropy	-	-	0.0258	0.7128	0.9149
MSSIM	-	-	-	0.0885	0.0306
BLIINDS2	-	-	-	-	0.7845
DIIVINE	-	-	-	-	-

- Although, the p-value for comparing MSSIM and BLIINDS2 is not low enough (0.08) but it is very much near to the cutoff point (0.05)

Despite that MSSIM was found significantly better than AMBE and Entropy, there is a problem applying it directly in contrast enhancement because it's originally designed for assessing image fidelity. A perfect rating of 1 simply means the output and input image is the same without any contrast enhancement.

CONCLUSION

In this study, the existing IQAs (AMBE and Entropy) used to evaluate the output of HE-based contrast enhancement methods have been reviewed and their shortcomings have been highlighted. An experiment has been conducted to evaluate the performance of the IQAs in term of their correlation with HVP of distortions. The experiment involved 1935 "ground truth" quality data collected from subject image quality assessment.

The experiment results showed that AMBE and Entropy have poor correlation with HVP. The results also showed that even a fidelity-based IQA, outperformed AMBE and Entropy to a statistically significant level, suggesting that there is an urgent need to develop a new IQA specifically for assessment of contrast enhancement methods.

The data used in this experiment is made publicly available to the research community for further scientific study in the field of image quality assessment.

ACKNOWLEDGMENT

This work was supported by the Malaysian Ministry of Higher Education (MOHE) under Fundamental Research Grant Scheme (FRGS). The Grant No. is FRGS/1/10/TK/UNITEN/02/37. Special thanks to all the UNITEN B.I.T second and final year students in year 2011 who have contributed to the compilation of opinion score.

REFERENCES

Abdullah-Al-Wadud, M., M.H. Kabir, M.A.A. Dewan and O. Chae, 2007. A dynamic histogram equalization for image contrast enhancement. IEEE Trans. Consumer Electron., 53: 593-600.

Arici, T., S. Dikbas and Y. Altunbasak, 2009. A histogram modification framework and its application for image contrast enhancement. IEEE Trans. Image Proc., 18: 1921-1935.

Chen, S.D. and A.R. Ramli, 2003a. Contrast enhancementusing recursive mean-separate histogram equalization for scalable brightness preservation. IEEE Trans. Consumer Electron., 49: 1301-1309.

Chen, S.D. and A.R. Ramli, 2003b. Minimum mean brightness error bihistogram equalization in contrast enhancement. IEEE Trans. Consumer Electron., 49: 1310-1319.

Chen, S.D. and A. bt. Suleiman, 2008. Scalalable global histogram equalization with selective enhancement for photo processing. Proceedings of the 4th International Conference on Information Technology and Multimedia, November 18-19, 2008, Malaysia, pp: 744-752.

Der, C.S. and M.S. Sidhu, 2009. Re-evaluation of automatic global histogram equalization-based contrast enhancement methods. J. Comput. Sci. Info. Technol., 1: 13-17.

Ibrahim, H. and N.S.P. Kong, 2007. Brightness preserving dynamic histogram equalization for image contrast enhancement. IEEE Trans. Consumer Electron., 53: 1752-1758.

Ibrahim, H. and N.S.P. Kong, 2009. Image sharpening using sub-regions histogram equalization. IEEE Trans. Consumer Electron., 55: 891-895.

Kim, Y.T., 1997. Contrast enhancement using brightness preserving bihistogram equalization. IEEE Trans. Consumer Electron., 43: 1-8.

Kim, M. and M.G. Chung, 2008. Recursively separated and weighted histogram equalization for brightness preservation and contrast enhancement. IEEE Trans. Consumer Electron., 54: 1389-1397.

Menotti, D., L. Najman, J. Facon and A.A. Araujo, 2007. Multi-histogram equalization methods for contrast enhancement and brightness preserving. IEEE Trans. Consumer Electron., 53: 1186-1194.

Moorthy, A.K. and A.C. Bovik, 2011. Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE Trans. Image Proces., 20: 3350-3364.

Ooi, C.H., N.S.P. Kong and H. Ibrahim, 2009. Bi-histogram with a plateau limit for digital image enhancement. IEEE Trans. Consumer Electron., 55: 2072-2080.

Ooi, C.H. and N.A.M. Isa, 2010a. Adaptive contrast enhancement methods with brightness preserving. IEEE Trans. Consumer Electron., 56: 2543-2551.

Ooi, C.H. and N.A.M. Isa, 2010b. Quadrants dynamic histogram equalization for contrast enhancement. IEEE Trans. Consumer Electron., 56: 2552-2559.

Saad, M.A., A.C. Bovik and C. Charrier, 2010. A DCT statistics-based blind image quality index. IEEE Signal Proces. Lett., 17: 583-586.

Sengee, N., A. Sengee and H.K. Choi, 2010. Image contrast enhancement using Bi-histogram equalization with neighbourhood metrics. IEEE Trans. Consumer Electron., 56: 2727-2734.

Sheet, D., H. Garud, A. Suveer, M. Mahadevappa and J. Chatterjee, 2010. Brightness preserving dynamic fuzzy histogram equalization. IEEE Trans. Consumer Electron., 56: 2475-2480.

Sheikh, H.R., M.F. Sabir and A.C. Bovik, 2006. A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans. Image Process., 15: 3440-3451.

VQEG, 2003. VQEG final report of FR-TV phase II validation test. Video Quality Experts Group.

Wang, Y., Q. Chen and B. Zhang, 1999. Image enhancement based on equal area dualistic sub-image histogram equalization method. IEEE Trans. Consumer Electronics, 45: 68-75.

Wang, Z., E.P. Simoncelli and A.C. Bovik, 2003. Multi-scale structural similarity for image quality assessment. Proceedings of the 37th Asilomar Conference Signals, Systems and Computers, Volume 2, November 9-12, 2003, California, USA., pp: 1398-1402.

Wang, C. and Z. Ye, 2005. Brightness preserving histogram equalization with maximum entropy: A variational perspective. IEEE Trans. Consumer Electron., 51: 1326-1334.

Wang, Q. and R.K. Ward, 2007. Fast image/video contrast enhancement based on weighted thresholded histogram equalization. IEEE Trans. Consumer Electron., 53: 757-764.

Wongsritong, K., K. Kittayaruasiriwat, F. Cheevasuvit, K. Dejhan and A. Somboonkaew, 1998. Contrast enhancement using multi-peak histogram equalization with brightness preserving. Proceedings of the Asia-Pacific Conference on Circuits and Systems, November 24-27, 1998, Chiangmai, pp: 455-458.

Yun, S.H., J.H. Kim and S. Kim, 2010. Image enhancement using a fusion framework of histogram equalization and Laplacian pyramid. IEEE Trans. Consumer Electron., 56: 2763-2771.