A Hybrid Image Quality Measure for Automatic Image Quality Assessment

Atif Bin Mansoor[1], Maaz Haider[1], Ajmal S. Mian[2], and Shoab A. Khan[1]

[1] National University of Sciences and Technology, Pakistan
[2] Computer Science and Software Engineering,
The University of Western Australia, Australia
atif-cae@nust.edu.pk, smaazhaider@yahoo.com, ajmal@csse.uwa.edu.au,
kshoab@yahoo.com

Abstract. Automatic image quality assessment has many diverse applications. Existing quality measures are not accurate representatives of the human perception. We present a hybrid image quality (HIQ) measure, which is a combination of four existing measures using an 'n' degree polynomial to accurately model the human image perception. First we undertook time consuming human experiments to subjectively evaluate a given set of training images, and resultantly formed a Human Perception Curve (HPC). Next we define a HIQ measure that closely follows the HPC using curve fitting techniques. The HIQ measure is then validated on a separate set of images by similar human subjective experiments and is compared to the HPC.The coefficients and degree of the polynomial are estimated using regression on training data obtained from human subjects. Validation of the resultant HIQ was performed on a separate validation data. Our results show that HIQ gives an RMS error of 5.1 compared to the best RMS error of 5.8 by a second degree polynomial of an individual measure HVS (Human Visual System) absolute norm (H_1) amongst the four considered metrics. Our data contains subjective quality assessment (by 100 individuals) of 174 images with various degrees of fast fading distortion. Each image was evaluated by 50 different human subjects using double stimulus quality scale, resulting in an overall 8,700 judgements.

1 Introduction

The aim of image quality assessment is to provide a quantitative metric that can automatically and reliably predict how an image will be perceived by humans. However, human visual system is a complex entity, and despite all advancements in the opthalmology, the phenomenon of image perception by humans is not clearly understood. Understanding the human visual perception is a challenging task, encompassing the complex areas of biology, psychology, vision etc. Likewise, developing an automatic quantitative measure that accurately correlates with the human perception of images is a challenging assignment [1]. An effective quantitative image quality measure finds its use in different image processing applications including image quality control systems, benchmarking and optimizing of image processing systems and algorithms [1]. Moreover, it

A.-B. Salberg, J.Y. Hardeberg, and R. Jenssen (Eds.): SCIA 2009, LNCS 5575, pp. 91–98, 2009.

can facilitate in evaluating the performance of imaging sensors, compression algorithms, image restoration and denoising algorithms etc. In the absence of a well defined mathematical model, researchers have attempted to find a quantitative metric based upon various heuristics to model the human image perception [2], [3]. These heuristics are based upon frequency contents, statistics, structure and Human Visual System. Miyahara et al [4] proposed a Picture Quality Scale (PQS), as a combination of three essential distortion factors; namely the amount, location and structure of error. Mean squared error (MSE) or its identical measure, peak signal to noise ratio (PSNR) has often been used as a quality metric. In [5], Guo and Meng have tried to evaluate the effectiveness of MSE as a quality measure. As per their findings, MSE alone cannot be a reliable quality index. Wang and Bovik [6] proposed a new universal image quality index Q, by modeling any image distortion as the combination of loss of correlation, luminance distortion and contrast distortion. The experimental results have been compared with MSE, demonstrating superiority of Q index over MSE. Wang et al [7] proposed a quality assessment named Structural Similarity Index based upon degradation of structural information. The approach was further improved by them to incorporate the multi scale structural information [8]. Shnayderman et al [9] explored the feasibility of Singular Value Decomposition (SVD) for quality measurement. They compared their results with PSNR, Universal Quality Index [6] and Structural Similarity Index [7] to demonstrate the effectiveness of the proposed measure. Sheikh et al. [10] gave a survey and statistical evaluation of full reference image quality measures. They included PSNR (Peak Signal to Noise Ratio), JND Metrix [11], DCTune [12], PQS [4], NQM [13], fuzzy S7 [14], BSDM (Block Spectral Distance Meausurement) [15], MSSIM (Multiscale Structural Similarity Index Measure) [8], IFC (Information Fidelity Criteria) [16], VIF (Visual Information Fidelity) [17] in the study and concluded that VIF performs the best among these parameters. Chandler and Hemami proposed a two staged wavelet based visual signal to noise ratio based on near-threshold and supra-threshold properties of human vision [18].

2 Hybrid Image Quality Measure

2.1 Choice of Individual Quality Measures

Researchers have devised various image quality measures following different approaches, and showed their effectiveness in respective domains. These measures prove effective in certain conditions and show restricted performance otherwise. In our approach, instead of proposing a new quality metric, we suggest an apt combinational metric benefiting from the strength of individual measures. Therefore, the choice of constituent measures has a direct bearing on the performance of the proposed hybrid metric. Avcibas et al. [15] performed a statistical evaluation of 26 image quality measures. They categorized these quality measures into six distinct groups based on the used type of information. More importantly, they clustered these 26 measures using a Self-Organizing Map (SOM) of distortion measures. Based on the clustering results, Analysis of variance (ANOVA) and subjective mean opinion score they concluded that five of the quality measures are most discriminating. These measures are edge stability measure (E_2), spectral phase magnitude error (S_2), block spectral phase magnitude error (S_5), HVS (Human Visual System) absolute norm (H_1) and HVS L2 norm (H_2). We chose four (H_1, H_2, S_2, S_5) of these five prominent quality measures due to their mutual non redundancy. E_2 was dropped due to its close proximity to H_2 in the SOM.

2.2 Experiment Setup

A total of 174 color images, obtained from LIVE image quality assessment database [19] representing diverse contents, were used in our experiments. These images have been degraded by using varying levels of fast fading distortion by inducing bit errors during transmission of compressed JPEG 2000 bitstream over a simulated wireless channel. The different levels of distortion resulted in a wide variation in the quality of these images. We carried out our own perceptual tests on these images. The tests were administered as per the guidelines specified in the ITU-Recommendations for subjective assessment of quality for television pictures [20]. We used three identical workstations with 17-inch CRT displays of approximately the same age. The resolution of displays were identical, 1024 x 768. External light effects were minimized, and all tests were carried out under the same indoor illumination. All subjects viewed the display from a distance of 2 to 2.5 screen heights. We employed Double stimulus quality scale method, keeping in view its more precise image quality assessments. A matlab based graphical user interface was designed to show the assessors a pair of pictures i.e. original and degraded. The images were rated using a five point quality scale; excellent, good, fair, poor and bad. The corresponding rating was scaled on a 1-100 score.

2.3 Human Subjects

The human subjects were screened and then trained according to the ITU-Recommendations [20]. The subjects of the experiment were male and female undergraduate students with no experience in image quality assessment. All participants were tested for vision impairments e.g., colour blindness. The aim of the test was communicated to each assessor. Before each session, a demonstration was given using the developed GUI with images different from the actual test images.

2.4 Training and Validation Data

Each of the 174 test images was evaluated by 50 different human subjects, resulting in 8,700 judgements. This data was divided into training and validation sets. The training set comprised 60 images, whereas the remaining 114 images were used for validation of the proposed HIQ.

A mean opinion score was formulated from the Human Perception Values (HPVs) adjudged by the human subjects for various distortion levels. As expected, it was observed that different humans subjectively evaluated the same image differently. To cater this effect, we further normalized the distortion levels and plotted the average MOS against these levels. It means that average mean opinion score of different human subjects against all the images with a certain level of degradation was plotted. As the images of a wide variety with different levels of degradation are used, therefore in this manner, we achieved an image independent Human Perception Curve (HPC).

Similarly, average values were calculated for H_1, H_2, S_2 and S_5 for the normalized distortion levels using code from [19]. All these quality measures were regressed upon HPC by using a polynomial of 'n' degree. The general form of the HIQ is given by Eqn. 1.

$$\mathrm{HIQ} = a_0 + \sum_{i=1}^{n}(a_i H_1^i) + \sum_{j=1}^{n}(b_j H_2^j) + \sum_{k=1}^{n}(c_k S_2^k) + \sum_{l=1}^{n}(d_l S_5^l) \tag{1}$$

We tested different combinations of these measures taking one, two, three and four measures at a time. All these combinations were tested up to fourth degree polynomial.

Table 1. RMS errors for various combination of Quality Measures. First block gives RMS error for individual measures, second, third and fourth blocks for combination of two, three and four measures respectively.

	Polynomial of degree 1		Polynomial of degree 2		Polynomial of degree 3		Polynomial of degree 4	
Comb. of Measures	Training RMS error	Validation RMS error	Training RMS error	Validation RMS error	Training RMS error	Validation RMS error	Training RMS error	Validation RMS error
S2	12.9	9.2	9.2	6.6	9.7	6.2	10.5	6.1
S5	13.2	10.2	6.9	7.3	7.2	6.9	7.7	7.1
H1	10.1	6.8	8.4	5.8	8.8	6.0	9.5	6.2
H2	14.8	10.8	15.4	10.0	14.4	20.4	10.5	75.7
S2–S5	11.7	9.0	5.6	8.1	4.9	8.5	4.8	8.8
S2–H1	7.2	5.8	4.2	6.3	4.0	6.2	3.9	6.6
S2–H2	9.4	7.5	6.6	7.2	6.5	7.5	6.8	6.4
S5–H1	7.2	6.2	2.9	6.4	2.9	6.4	2.4	6.3
S5–H2	9.4	8.3	4.2	8.0	4.1	8.9	4.0	9.1
H1–H2	4.4	5.4	3.1	6.5	2.8	9.9	2.2	23.1
S2–S5–H1	7.2	5.8	2.2	6.7	0.2	12	0.3	16.9
S2–S5–H2	9.4	8.0	2.9	9.3	1.0	15.8	0.4	21.5
S2–H1–H2	4.0	5.1	1.5	5.6	1.3	7.6	1.9	5.5
S5–H1–H2	4.2	5.1	1.9	5.4	1.1	6.0	0.0	22.9
S2–S5–H1–H2	3.7	5.5	1.3	7.2	0.0	14.1	0.3	16.9

3 Results

We performed a comparison of the mean square error for individual and various combinations of the quality measures for fast fading degradation. Table 1 shows the RMS errors obtained after regression on the training data and then verified on the validation data. The minimum RMS errors (approx equal to zero) on the training data were achieved using a third degree polynomial combination of all the four measures and a fourth degree polynomial combination of S_5, H_1, H_2. However, using the same combinations resulted in unexpected RMS errors of 14.1 and 22.9 respectively during validation indicating cases of overfitting on the training data. The most optimal results are given by a linear combination of H_1, H_2, S_2 which provide RMS errors of 4.0 and 5.1 on the training and validation data respectively. Therefore, we concluded that a linear combination of these measures gives the best estimate of human perception. Resultantly, regressing the values of these quality measures against HPC of the training data, the coefficients a_0, a_1, b_1, c_1 as given in Eqn. 1 were found. Thus, the HIQ measure achieved is given by:

$$\text{HIQ} = 85.33 - 529.51H_1 - 2164.50H_2 - 0.0137S_2 \tag{2}$$

Fig. 1 shows the HPV curve and the regressed HIQ measure plot for the training data. The HPV curve was calculated by averaging the HPVs of all images

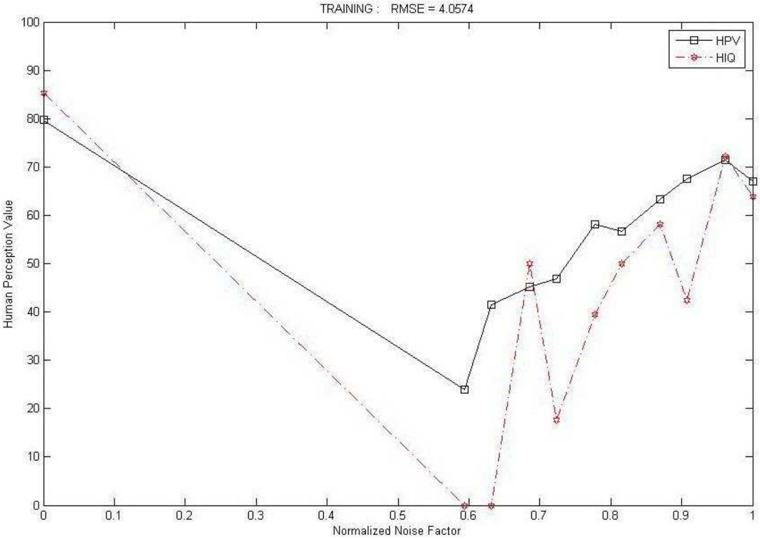

Fig. 1. Training Data of 60 images with different levels of noise degradation. Any one value e.g. 0.2 corresponds to a number of images but all suffering with 0.2% of fast fading distortion, and the corresponding value of HPV is mean opinion score of all human judgements for these 0.2% degraded images (50 human judgements for one image). HIQ curve is obtained by averaging the HIQ measures obtained from proposed mathematical model, Eqn. 2, for all images having the same level of fast fading distortion. The data is made available at http://www.csse.uwa.edu.au/ ∼ ajmal/.

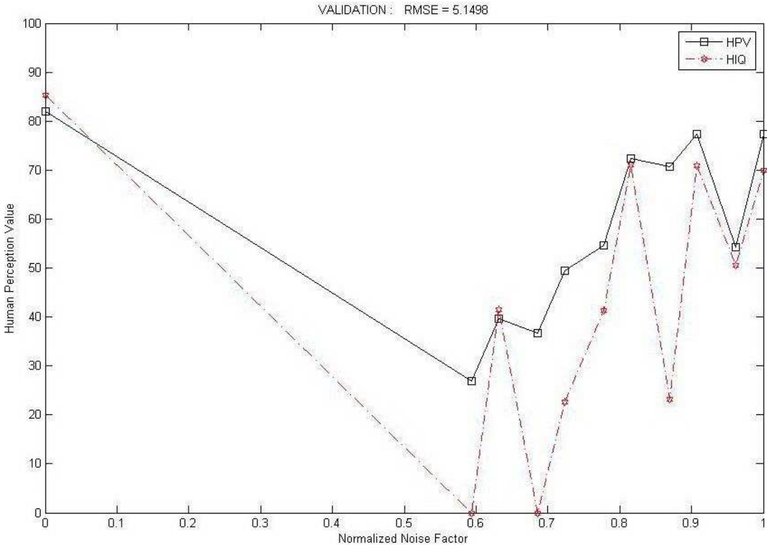

Fig. 2. Validation Data of 114 images with different levels of noise degradation. Any one value e.g. 0.8 corresponds to a number of images but all suffering with 0.8% of fast fading distortion, and the corresponding value of HPV is mean opinion score of all human judgements for these 0.8% degraded images (50 human judgements for one image). HIQ curve is obtained by averaging the HIQ measures obtained from proposed mathematical model, Eqn. 2, for all images having the same level of fast fading distortion. The data is made available at http://www.csse.uwa.edu.au/ ∼ ajmal/.

having the same level of fast fading distortion. Similarly, the HIQ curve is calculated by averaging the HIQ measures obtained from Eqn. 2 for all images having the same level of fast fading distortion. Thus Fig. 1 depicts the image independent variation in HPV and the corresponding changes in HIQ for different normalized levels of fast fading. Fig. 2 shows similar curves obtained on the validation set of images. Note that the HIQ curves, in both the cases (i.e. Fig. 1 and 2), closely follow the same pattern of the HPV curves which is an indication that the HIQ measure accurately correlates with the human perception of image quality. The following inferences can be made from our results given in Table 1. (1) H_1, H_2, S_2 and S_5 individually perform satisfactorily which demonstrates their acceptance as image quality measures. (2) The effectiveness of these measures improve by modeling them as polynomials of higher degrees. (3) Increasing the combination of these quality measures e.g., using all four measures does not necessarily increase their effectiveness, as this may suffer from overfitting on training data. (4) An important finding is validation of the fact that HIQ measure closely follows the human perception curve, as evident from Fig. 2 where HIQ curve has similar trend as of HPV, though both are calculated independently. (5) Finally, a linear combination of H_1, H_2, S_2 gives the best estimate of the human perception of an image quality.

4 Conclusion

We presented a hybrid image quality measure, HIQ, consisting of a first order polynomial combination of three different quality metrics. We demonstrated its effectiveness by evaluating it over a separate validation data consisting of a separate set of 114 different images. HIQ proved to closely follow the human perception curve and gave an error improvement over the individual measures. In the future, we plan to investigate the HIQ for other degradation models like white noise, JPEG compression, gaussian blur etc.

References

1. Wang, Z., Bovik, A.C., Lu, L.: Why is Image Quality Assessment so difficult. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, pp. 3313–3316 (2002)
2. Eskicioglu, A.M.: Quality measurement for monochrome compressed images in the past 25 years. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, pp. 1907–1910 (2000)
3. Eskicioglu, A.M., Fisher, P.S.: Image Quality Measures and their Performance. IEEE Transaction on Communications 43, 2959–2965 (1995)
4. Miyahara, M., Kotani, K., Algazi, V.R.: Objective Picture Quality Scale (PQS) for image coding. IEEE Transaction on Communications 9, 1215–1225 (1998)
5. Guo, L., Meng, Y.: What is Wrong and Right with MSE. In: Eighth IASTED International Conference on Signal and Image Processing, pp. 212–215 (2006)
6. Wang, Z., Bovik, A.C.: A universal image quality index. IEEE Signal Processing Letters 9, 81–84 (2002)
7. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error measurement to structural similarity. IEEE Transaction on Image Processing 13 (January 2004)
8. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multi-scale structural similarity for image quality assessment. In: 37th IEEE Asilomar Conference on Signals, Systems, and Computers (2003)
9. Shnayderman, A., Gusev, A., Eskicioglu, A.M.: An SVD-Based Gray-Scale Image Quality Measure for Local and Global Assessment. IEEE Transaction on Image Processing 15 (February 2006)
10. Sheikh, H.R., Sabir, M.F., Bovik, A.C.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transaction on Image Processing 15, 3440–3451 (2006)
11. Sarnoff Corporation, JNDmetrix Technology, http://www.sarnoff.com
12. Watson, A.B.: DC Tune: A technique for visual optimization of DCT quantization matrices for individual images, Society for Information Display Digest of Technical Papers, vol. XXIV, pp. 946–949 (1993)
13. Damera-Venkata, N., Kite, T.D., Geisler, W.S., Evans, B.L., Bovik, A.C.: Image Quality Assessment based on a Degradation Model. IEEE Transaction on Image Processing 9, 636–650 (2000)
14. Weken, D.V., Nachtegael, M., Kerre, E.E.: Using similarity measures and homogeneity for the comparison of images. Image and Vision Computing 22, 695–702 (2004)

15. Avcibas, I., Sankur, B., Sayood, K.: Statistical Evaluation of Image Quality Measures. Journal of Electronic Imaging 11, 206–223 (2002)
16. Sheikh, H.R., Bovik, A.C., de Veciana, G.: An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Transaction on Image Processing 14, 2117–2128 (2005)
17. Sheikh, H.R., Bovik, A.C.: Image information and Visual Quality. IEEE Transaction on Image Processing 15, 430–444 (2006)
18. Chandler, D.M., Hemami, S.S.: VSNR: A Wavelet base Visual Signla-to-Noise Ratio for Natural Images. IEEE Transaction on Image Processing 16, 2284–2298 (2007)
19. Sheikh, H.R., Wang, Z., Cormack, L., Bovik, A.C.: LIVE image quality assessment database, http://live.ece.utexas.edu/research/quality
20. ITU-R Rec. BT. 500-11, Methodology for the Subjective Assessment of the Quality for Television Pictures