Analysis and Comparison of Image Quality Metrics

Cristaudo Giuseppe
University of Bologna, Dipartimento di Informatica – Scienza e Ingegneria

August 31, 2022

Abstract

In this essay, the primary objective image quality measures are reviewed. Many image processing applications, including picture restoration, compression, acquisition, enhancement, and others, heavily rely on these metrics. Modern methodologies aim to establish the optimum balance between performance, complexity, and results that are closest to the Human Visual System. This work collects and groups reported quality measures according to their tactics and techniques and proposes a novel method for RGB images.

I. Introduction

Digital imaging systems have had exponential growth over the past forty years and have taken over as the predominant method of image processing. Digital images are susceptible to a wide range of distortions during acquisition, processing, compression, storage, transmission, and reproduction, any of which could lower visual quality. Therefore, Image Quality Assessments (IQA) are effective at evaluating the quality of images in a variety of image processing applications [1], including benchmarking for image and video processing systems and algorithms (e.g. compression), keeping track of image quality in quality control systems (e.g. streaming services), and enhancing algorithm and parameter settings in image and video processing systems.
The structure of this essay is as follows. The purposes of image quality evaluations are covered in Section II. Section III discusses objective measures, including mathematical methods and methods based on the human visual system. Section IV describes the image dataset that was used to evaluate and compare various IQA. The paper concludes in section V where IQA results are reviewed.

II. Image Quality Assessments

Image quality assessments are metrics that gauge an image's quality in relation to the source image. Historically, image quality is described in terms of the visibility of the distortions in an image (Sarnoff model [2] and Janssen model [3]), such as color shifts, blurriness, Gaussian noise and blockiness. Nowadays, Image Quality Metrics are grouped into:

- Subjective method which involves human beings to evaluate image quality.
- Objective method which computes the image quality automatically.

In multimedia applications, the naked human eye will always serve as the final arbiter of quality, hence the subjective technique is unquestionably the most precise and trustworthy way to evaluate the quality of a picture. However, this approach is too time- and money-consuming for practical or commercial usage. As a result, the objective IQA research's mission is to predict an image's quality as similar as possible to the subjective evaluation.

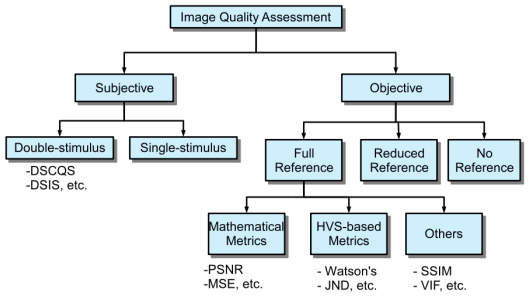

Figure 1: *How Image Quality Assessments are classified*

III. Objective Metrics

As can be seen in Fig.1, according to the availability of the source image, objective IQA are divided into three categories: full-reference (FR), reduced-reference (RR), and no-reference (NR). In this paper, we are focusing on the FR metric, which computes the quality of the deformed images by comparing them to the original image, which is presumed to be of perfect quality.

i. Mathematical Metrics

Mathematical metrics compute the quality of an image as the similarity between the reference image and the distorted image. Because of their low complexity and clear physical meaning [4], two commonly used measures are *Mean-Squared Error* (MSE) and *Peak Signal-to-Noise Ratio* (PSNR).

a Mean-Squared Error (MSE)

The MSE between a reference image X and corresponding distorted image Y is given by:

$$\text{MSE(X, Y)} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

Where N represents the number of pixels in the original image. One problem with Mean-Squared Error is the strong dependence of image intensity scaling. A mean-squared error of 100.0 for an 8-bit image looks bad, but a MSE of 100.0 for a 10-bit image is barely noticeable. Peak Signal-to-Noise Ratio (PSNR) avoids this problem by scaling the MSE according to the image range.

b Peak Signal-to-Noise Ratio (PSNR)

The PSNR between a reference image X and corresponding distorted image Y is given by:

$$\text{PSNR(X, Y)} = 10log_{10}\frac{255^2}{MSE(X,Y)}$$

where 255 is the maximum gray level of a 8-bit image. PSNR is measured in decibels (dB).

In [7], further mathematical metrics like average difference, maximum difference, absolute error, Peak MSE, Laplacian MSE, etc are discussed.
Because the Human Vision System (HVS) characteristics are not taken into account in their models, these approaches do not correlate well with perceived quality measurement [5] [6] . Their main advantages are simplicity and mathematical application.

ii. Human-Visual-System-based Metrics

These metrics aim to be as similar as possible to human perception. Typically HVS metrics are based on:

- *Contrast Sensitivity Function*: human perception is more sensitive to lower spatial frequency than higher one.
- *Luminance Contrast Sensitivity*: human eyes are sensitive to luminance contrast rather than the absolute luminance value [8].
- *Contrast Masking*: refers to the reduction in the visibility of one image component by the presence of another.

Based on the presumption that the human visual system is highly adapted to gather structural information from the seeing field, Wang [9] developed the first HVS-based measures. The *Structural Similarity* (SSIM) is a new metric proposed by Wang to assess the quality of distorted images, and various tests demonstrate increased consistency with HVS when compared to PSNR and MSE. Despite the fact that Wang's method is largely acknowledged, there are some drawbacks, including the complexity and nonlinearity of HVS and the suprathreshold problem [9].

a Structural Similarity (SSIM)

Structural similarity metric attempts to measure image quality by capturing the similarity between luminance, contrast and structure of the reference and distorted image.

Luminance comparison function l(X,Y) between a reference image X and corresponding distorted image Y is given by:

$$l(X,Y) = \frac{2\mu_x\mu_y+C1}{\mu_x^2+\mu_y^2+C1}$$

Where μ_x and μ_y are the mean values of X and Y respectively and C1 is a small positive constant used to avoid dividing by zero.
Similarly, contrast comparison function c(X, Y) is defined as:

$$c(X,Y) = \frac{2\sigma_x\sigma_y+C2}{\sigma_x^2+\sigma_y^2+C2}$$

Where σ_x and σ_y are the standard deviation value of X and Y respectively and C2 is again a stabilization constant. Finally, structure comparison function s(X, Y) is defined as:

$$s(X,Y) = \frac{2\sigma_{xy}+C3}{\sigma_x\sigma_y+C3}$$

Where σ_{xy} represents the correlation between X and Y and C3 is a constant that provides stability.
By combining the three components, the SSIM index is given by:

$$SSIM(X,Y) = [l(X,Y)]^{\alpha} * [c(X,Y)]^{\beta} * [s(X,Y)]^{\gamma}$$

Where α, β and γ are variables used to adjust the relative importance of each components and for simplicity in this paper we assume $\alpha = \beta = \gamma = 1$ and $C3 = \frac{C2}{2}$. Consequently SSIM index can be written as:

$$SSIM(X,Y) = \frac{(2\mu_x\mu_y+C1)(2\sigma_{xy}+C2)}{(\mu_x^2+\mu_y^2+C1)(\sigma_x^2+\sigma_y^2+C2)}$$

Symmetric Gaussian or Mean weighting functions are used to estimate local SSIM statics. The mean SSIM index pools the spatial SSIM values to evaluate overall image quality [11].

$$\text{SSIM(X, Y)} = \frac{1}{N}\sum_{i=1}^{N} SSIM(x_i - y_i)$$

Where x_i and y_i are image patches covered by the i^{th} window and N is the number of local windows over the image.

b Multi-Scale-SSIM (MS-SSIM)

Multi-Scale SSIM extends standard SSIM by integrating image details at different resolution.

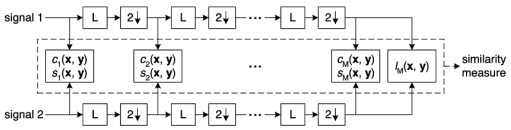

Figure 2: *MS-SSIM computation. L: low-pass filtering; 2 ↓: downsampling by 2.*

As shown in Fig.2 the original and distorted images are iteratively low-pass filtered and downsampled by a factor of two. For this multi scale operation, the reference image is taken as scale 1, meanwhile M is the highest scale. Likewise the computation of SSIM, MS-SSIM is given by the combination of contrast and structure comparison on intermediate levels, meanwhile the luminance comparison is combined only for scale M. Furthermore, the MS-SSIM between a reference image X and corresponding distorted image Y is given by:

$$MS-SSIM(X,Y) = [l_M(X,Y)]^{\alpha_M} \prod_{i=1}^{M} [c_i(X,Y)]^{\beta_i} [s_i(X,Y)]^{\gamma_i}$$

Where exponents α_M , β_i and γ_i are non zero values used to adjust the relative importance of different components. For simplicity, in this paper we assume $\alpha_M = \beta_i = \gamma_i = 1$.

c Edge-based structural similarity (E-SSIM)

Many researches pointed out that the human visual system is very sensitive to the edge and contour information of a given image. In fact, the edge and contour informations may be the most important information of an image's structure for a human to 'capture' the scene [10]. These studies resulted in an updated and improved version of SSIM: *edge-based structural similarity* (E-SSIM). E-SSIM compares the edge information between the distorted image block and the original one, and replaces the structure comparison (in the standard SSIM computation) with an edge-based structure comparison. There are several methods for obtaining edge information, including the simple edge detection algorithm, local gradients, and so on. Because of its simplicity and efficiency, the Sobel

operator has been used to obtain edge information in this paper. The Yang edge direction histogram [11] is then used to compare the edge information in the original and distorted images.
To do so, we compute the amplitude and orientation of each pixel edge in both images.

$$amplitude = g_x^2 + g_y^2$$

$$orientation = \arctan\left(\frac{g_y}{g_x}\right)\frac{180^\circ}{\pi}$$

Where g_x and g_y are gradients respectively along the x and y axis.

We quantize the continual direction $(0^\circ - 180^\circ)$ in a given number of discrete direction (e.g 8 discrete direction) and finally we sum up all the pixels' edge amplitudes with the same direction in the block.

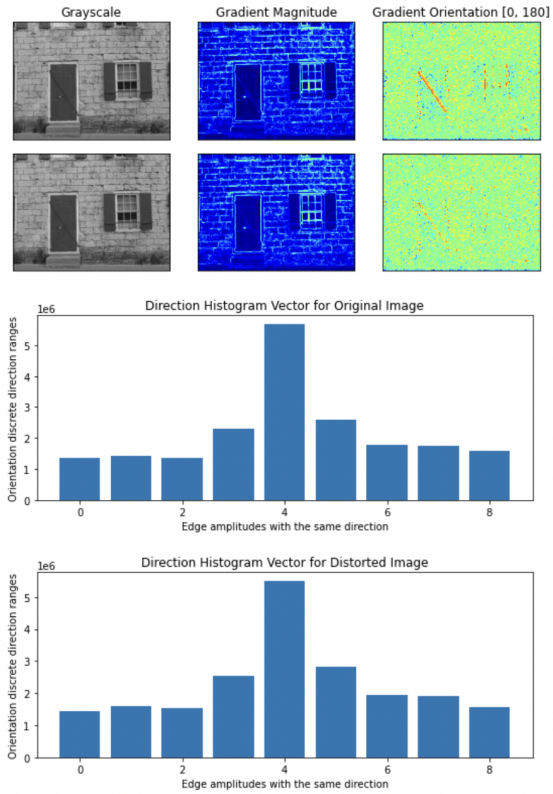

Figure 3: *The produced output by our E-SSIM implementation.*

Given the original and distorted image edge direction vector histogram, respectively D_x and D_y , the edge comparison $e(X, Y)$ is given by:

$$e(X, Y) = \frac{\sigma_{xy}+C3}{\sigma_x\sigma_y+C3}$$

Where σ_x and σ_Y are the standard deviation of vector D_x and D_y respectively, σ_{xy} is the covariance vector of D_x and D_y, and $C3$ is a small constant to avoid the denominator being zero. Finally, E-SSIM index between a reference image X and corresponding distorted image Y is given by:

$$SSIM(X, Y) = [l(X, Y)]^\alpha * [c(X, Y)]^\beta * [e(X, Y)]^\gamma$$

Where are parameters used to adjust the relative importance of each components and for simplicity in this paper we assume $\alpha = \beta = \gamma = 1$.

d Color-based-edge-based structural similarity (C-ESSIM)

All previous structural similarity methods were created for grayscale images. However, the chrominance information encoded in an RGB image is widely recognized to have a major impact on how well the human visual system understands a sample under examination. This approach can be developed by using a straightforward extension to already developed SSIM frameworks.
The RGB color images are converted into another color space where the luminance and chrominance can be separated. In this paper, YIQ color space [12], in which Y represents luminance information, I represents in-phase information, and Q represents chrominance information , is used. RGB space can be projected into YIQ space as follows:

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} \approx \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.5959 & -0.2746 & -0.3213 \\ 0.2115 & -0.5227 & 0.3112 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

Let $I(X)$, $I(Y)$ and $I(Y)$, $Q(Y)$ be the I and Q chromatic channels of the image X and Y, respectively. The similarity between chromatic features is defined as:

$$S_i(X,Y) = \frac{2I(X)I(Y)+C1}{I(X)^2+I(Y)^2+C1}$$

$$S_q(X,Y) = \frac{2Q(X)Q(Y)+C2}{Q(X)^2+Q(Y)^2+C2}$$

Where $C1$ and $C2$ are small positive constants to avoid the denominator being zero.
The final metrics are calculated by combining the similarity of chromatic features with one of the previous structure similarity frameworks. The E-SSIM method has been extended in this paper because results show that it performs better with highly blurred images than other described metrics.

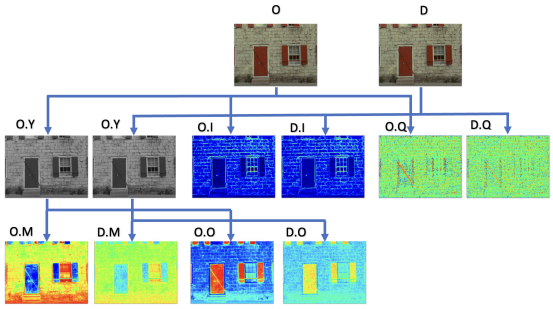

Figure 4: *Index computation of C-ESSIM. O: original image, D: distorted image, .Y, .I, .Q YIQ components, .O orientation component, .M magnitude component*

IV. Dataset

There are six main publicly available image databases in the IQA community, including TID2008 [13], CSIQ [14], LIVE [15], IVC [16], MICT [17] and A57 [18].

Database	Source Images	Distorted Images	Distortion Types	Image Type	Observers
TID2008	25	1700	17	color	838
CSIQ	30	866	6	color	35
LIVE	29	779	5	color	161
IVC	10	185	4	color	15
MICT	14	168	2	color	16
A57	3	54	6	gray	unknown

Figure 5: *Characteristics of these six databases.*

In this paper, TID2008 will be used to evaluate and compare the previously described IQA. All images in the database are of size 512x384 pixels and different kind of distortions at different levels of intensity have been applied as follow:

№	Type of distortion (four levels for each distortion)	Correspondence to practical situation	Accounted HVS peculiarities
1	Additive Gaussian noise	Image acquisition	Adaptivity, robustness
2	Additive noise in color components is more intensive than additive noise in the luminance component	Image acquisition	Color sensitivity
3	Spatially correlated noise	Digital photography	Spatial frequency sensitivity
4	Masked noise	Image compression, watermarking	Local contrast sensitivity
5	High frequency noise	Image compression, watermarking	Spatial frequency sensitivity
6	Impulse noise	Image acquisition	Robustness
7	Quantization noise	Image registration, gamma correction	Color, local contrast, spatial frequency
8	Gaussian blur	Image registration	Spatial frequency sensitivity
9	Image denoising	Image denoising	Spatial frequency, local contrast
10	JPEG compression	JPEG compression	Color, spatial frequency sensitivity
11	JPEG2000 compression	JPEG2000 compression	Spatial frequency sensitivity
12	JPEG transmission errors	Data transmission	Eccentricity
13	JPEG2000 transmission errors	Data transmission	Eccentricity
14	Non eccentricity pattern noise	Image compression, watermarking	Eccentricity
15	Local block-wise distortions of different intensity	Inpainting, image acquisition	Evenness of distortions
16	Mean shift (intensity shift)	Image acquisition	Light level sensitivity
17	Contrast change	Image acquisition, gamma correction	Light level, local contrast sensitivity

Figure 6: *Type of distortions used in TID2008 database.*

For each distorted image, the database also provides a *mean opinion score* (MOS) value. The MOS is a quality metric that represents the overall quality of a stimulus or a system and has been used to compare competing IQA metrics.

V. Results

Previously reported Image Quality Assessments are now being evaluated in order to compare their overall effectiveness. Three simulations were run on the entire TID2008 dataset, on highly blurred images and on JPEG2000 compressed (one of the most common types of images compression) images. Figures 7, 8 and 9 depict the scatter plots of the Mean Opinion Score and different metrics. All performances have been computed using default parameters for all approaches and a windows size of eleven pixels when it was requested.

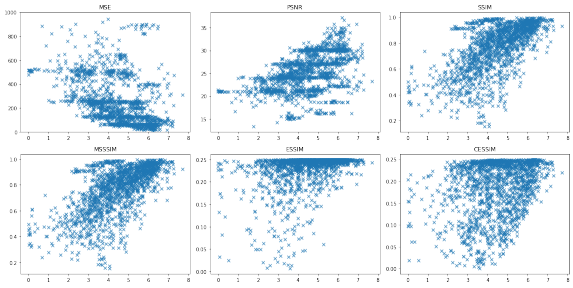

Figure 7: *Scatter plots of MOS versus model prediction for TID2008 dataset.*

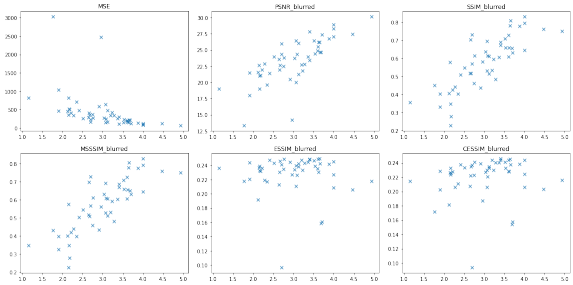

Figure 8: *Scatter plots of MOS versus model prediction for Gaussian blur distorted images.*

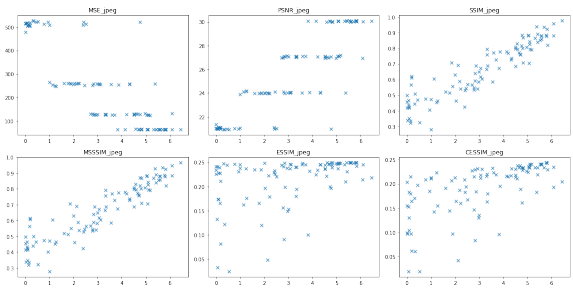

Figure 9: *Scatter plots of MOS versus model prediction for JPEG2000 compressed images.*

Tables 1, 2 and 3 show quantitative measures of performance for IQA, and dive metrics [20] are used to evaluate objectives models for the three scenarios. The prediction accuracy is provided by the correlation coefficients (CC), and the higher the CC value, the better the accuracy. The mean absolute error (MAE), root mean square error (RMS) and outlier ratio (OR), which is a percentage of the number of predictions outside the range of ±2 times of the standard deviations, are measures of prediction consistency, with a smaller value indicating better performance. Finally, the Spearman rank-order correlation coefficient (ROCC) is used to assess prediction monotonicity. The performance measures were then ranked to make the results easier to understand.

	CC	MAE	RMS	OR	ROCC
MSE	1	6	6	3	1
PSNR	4	5	5	6	4
SSIM	5	**1**	**1**	**1**	4
MS-SSIM	**6**	2	2	2	**6**
ESSIM	3	3	3	5	3
CESSIM	2	4	4	4	2

Table 1: *Performance comparison of image quality assessment models on TID2008 dataset images.*

	CC	MAE	RMS	OR	ROCC
MSE	1	6	6	3	1
PSNR	4	5	5	6	4
SSIM	5	**1**	**1**	**1**	5
MS-SSIM	**6**	2	3	2	**6**
ESSIM	3	3	3	4	3
CESSIM	2	4	4	5	2

Table 2: *Performance comparison of image quality assessment models on Gaussian blur distorted images.*

	CC	MAE	RMS	OR	ROCC
MSE	1	6	6	3	1
PSNR	4	5	5	6	4
SSIM	5	**1**	**1**	2	5
MS-SSIM	**6**	3	2	**1**	**6**
ESSIM	3	3	3	5	3
CESSIM	2	4	4	4	2

Table 3: *Performance comparison of image quality assessment models on JPEG2000 compressed images.*

VI. Conclusion

Image quality evaluation is critical in a variety of image processing applications. According to the experimental results, MSE and PSNR are very simple, easy to implement, and have low computational complexities. However, these methods do not produce satisfactory results. MSE and PSNR are acceptable image similarity measures only when the images differ by simply increasing a specific type of distortion. When used to measure across distortion types, however, they fail to capture image quality. SSIM is a popular method for measuring image quality. It works well and accurately measures across distortion types better than MSE and PSNR, but fails in the case of a highly blurred image. MS-SSIM is also not very useful for heavily blurred images, but it has a very high prediction rate and is robust when comparing images of different scales. Finally, among the HVS-model-based models, E-SSIM and its extension C-ESSIM produced the worst results. Although I believe it is due to poor implementation, studies show that E-SSIM improves performance by paying more attention to the edges and details in images, which represent the higher layer image structure information. Better outcomes might be obtained with more fine-tuning of α, β and γ. Then, the mathematical and HVS-model major goal quality techniques were compared. Results reveal that while HVS-models generally offer greater correlation with subjective measurements, they can sometimes produce subpar results if test pictures suffer from slight degradation (JPEG degradation). Mathematical methods continue to produce accurate and quick findings for the deterioration of Gaussian noise.

VII. Future Work

The field of image quality assessments is expanding daily, and new artificial intelligence-based methods have made it possible to evaluate an image's quality independently of its source. These techniques extrapolate features from images to determine the general quality of a distorted sample using high-performance convolutional neural networks and huge datasets.

References

[1] Z.Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Ap- plications*, B. Furht and O. Marqure, Eds. Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin, Austin, TX 78712: CRC Press, Sept 2003, ch. 41, pp. 1041–1078.

[2] "Video clarity website", www.videoclarity.com/WhitePapers.html.

[3] R. Janssen, *Computational Image Quality*. SPIE Press, 2001.

[4] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Addison-Wesley, New York, 1992.

[5] B. Girod, "What's wrong with mean-squared error," in *Digital images and human vision*, A. Watson, Ed. Cambridge, MA, USA: MIT Press, 1993, ch. 15, pp. 207–220.

[6] Z. Wang and A. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[7] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Transactions on communications*, vol. 43, no. 12, pp. 2959–1965, 1995.

[8] A. B. Watson, "Dct quantization metrices visually optimized for individ- ual images," in *Human Vision, Visual Processing, and Digital Display IV*, ser. Proc. SPIE, B. E. Rogowitz, Ed., vol. 1913, 1993, pp. 202–216.

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment:

from error visibility to structural similarity," *Image Processing IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[10] Guan-Hao Chen, Chun-Ling Yang, Lai-Man Po and Sheng-Li Xie, "Edge-Based Structural Similarity for Image Quality Assessment," *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006, pp. II-II, doi: 10.1109/ICASSP.2006.1660497.

[11] Y. Yang, N. P. Galatsanos, and A. K. Katsaggelos, "Regularized reconstruction to reduce blocking artifacts of block discrete cosine transform compressed images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 3, no. 6, pp. 421– 432, Dec. 1993.

[12] C. Yang and S.H. Kwok, "Efficient gamut clipping for color image processing using LHS and YIQ", *Optical Engineering*, vol. 42, no. 3, pp.701-711, Mar. 2003.

[13] 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*, vol. 10, pp. 30-45, 2009.

[14] E.C. Larson and D.M. Chandler, "Categorical Image Quality (CSIQ) Database", http://vision.okstate.edu/csiq.

[15] H.R. Sheikh, K. Seshadrinathan, A.K. Moorthy, Z. Wang, A.C. Bovik, and L.K. Cormack, "Image and video quality assessment research at LIVE", http://live.ece.utexas.edu/research/quality.

[16] A. Ninassi, P. Le Callet, and F. Autrusseau, "Subjective quality assessment-IVC database", http://www2.irccyn.ec-nantes.fr/ivcdb.

[17] Y. Horita, K. Shibata, Y. Kawayoke, and Z.M. Parves Sazzad, "MICT Image Quality Evaluation Database", http://mict.eng.u-toyama.ac.jp/mict/index2.html.

[18] D.M. Chandler and S.S. Hemami, "A57 database", http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html.

[19] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment", http://www.vqeg.org, 2000.

[20] Wang Kong-qiao,Shen Lan-sun,Xing Xin."A Quality Assessment Method of Image Based on Visual Interests". *Journal of Image and Graphics*. pp.300-303,2000