Full-Reference Visual Quality Assessment for Synthetic Images: A Subjective Study

Debarati Kundu and Brian L. Evans
Embedded Signal Processing Laboratory
The University of Texas at Austin, Austin, TX
Email: debarati@utexas.edu, bevans@ece.utexas.edu

***Abstract*—Measuring visual quality, as perceived by human observers, is becoming increasingly important in the many applications in which humans are the ultimate consumers of visual information. For assessing subjective quality of natural images, such as those taken by optical cameras, significant progress has been made for several decades. To aid in the benchmarking of objective image quality assessment (IQA) algorithms, many natural image databases have been annotated with subjective ratings of the images by human observers. Similar information, however, is not as readily available for synthetic images commonly found in video games and animated movies. In this paper, our primary contributions are (1) conducting subjective tests on our publicly available ESPL Synthetic Image Database, and (2) evaluating the performance of more than 20 full reference IQA algorithms for natural images on the synthetic image database. The ESPL Synthetic Image Database contains 500 distorted images (20 distorted images for each of the 25 original images) in 1920 $\times$ 1080 format. After collecting 26000 individual human ratings, we compute the differential mean opinion score (DMOS) for each image to evaluate IQA algorithm performance.**

I. Introduction

Recent years have seen a huge growth in the acquisition, transmission, and storage of videos. In addition to videos captured with optical cameras, video traffic also consists of synthetic scenes, such as animated movies, cartoons and video games. The burgeoning popularity of multiplayer video games (esp. on handheld platforms) is causing an exponential increase in synthetic video traffic. In all these cases, the ultimate goal is to provide the viewers with a satisfactory quality-of-experience (QoE). Methods of evaluating the visual quality plays an important role in the optimal design of displays, rendering engines and maintaining a satisfactory QoE in streaming applications under given network constraints.

The 'gold-standard' in assessing the perceptual quality of images and videos is to seek human opinion. But conducting subjective studies is time consuming and infeasible for many applications. However, the ground-truth data obtained from human observers can be used to benchmark different objective IQA algorithms which aim at automating the process of visual quality assessment.

In this paper we present the results from a subjective test conducted on synthetic images. The study included 25 high definition reference images, from which 500 images were created by the controlled addition of different levels of five commonly encountered artifacts. Every image was evaluated by 64 observers under controlled laboratory conditions in a single stimulus experiment, where the observers rated the visual quality on a continuous quality scale. The DMOS obtained augment the previously released ESPL Synthetic Image Database [1] [2] containing the unannotated pristine and distorted images.

Some of the largest and most popular natural image databases are LIVE Image Quality Database (LIVE) [3], Tampere Image Database 2013 [4], Categorical Image Quality Database [5] and EPFL JPEG XR codec [6]. The performance of several publicly available state-of-the-art full-reference(FR) IQA algorithms has been evaluated for seven natural image databases in [7].

Recently Cadík *et al.* have developed a database of computer graphics generated imagery with distortions such as noise, aliasing, change in brightness, light leakage, tone mapping artifacts, etc. and evaluated the performance of six FR-IQA algorithms [8]. The authors demonstrated that the FR-IQA algorithms were sensitive towards brightness and contrast changes, could not distinguish between plausible and implausible shading and failed to localize distortions precisely.

From our ESPL Synthetic Image Database, we consider a larger number of photo-realistic images and a broader class of distortions (transmission and compression artifacts for synthetic images) than the work by Cadík *et al.* [8] [9] in the hope of providing a better representation of the types of images and artifacts encountered in watching animated movies and playing video games. The performance of FR-IQA algorithms has also been evaluated, by using hypothesis testing and statistical significance analysis. To the best of our knowledge, we have considered the largest number of FR-IQA algorithms in any previously published survey. This provides the researchers a valuable tool by which which they can evaluate the performance of the existing and proposed objective IQAs on synthetic images.

(a) (b) (c)

Fig. 1: Sample Synthetic Images in the ESPL database [1]

II. Summary of Subjective Study

A. Source Images

For the purpose of this study, 25 synthetic images have been chosen from video games and animated movies. These high quality color 8-bit images (pixel values ranging from 0-255) from the Internet are 1920×1080 pixels in size. Some video games which were considered were multiplayer role playing games (such as World of Warcraft), first person shooter games (such as Counter Strike), motorcycle and car racing games, and games with more realistic content (such as FIFA). Some of the animated movies, from which the images were collected, are, The Lion King, the Tinkerbell series, Avatar, The Beauty and the Beast, Monster series, Ratatouille, the Cars series, etc. [1]

B. Distorted Images

For this database, three categories of processing artifacts have been considered, namely interpolation (which arises frequently in texture maps, causing jaggedness of crisp edges), blurring and additive Gaussian noise. With the advent of cloud gaming, where the rendered 2D game images are streamed from the server to the 'dumb' clients, we have chosen to study the effect of compression and transmission artifacts on computer graphics generated images (which had been previously considered only for natural scenes). For this database, JPEG compression and Rayleigh fast-fading wireless channel artifacts have been considered. For each artifact type, four different levels were considered, resulting in 20 distorted image created from a single pristine image.

1) *Interpolation*: The original images were downsampled using integer downsampling factors from 3 to 6, which are upsampled back using a nearest neighbor approach.
2) *Gaussian Blur*: The RGB color channels were filtered using a circularly symmetric 2D Gaussian kernel with standard deviation ranging from 1.25 to 3.5 pixels. The same kernel was employed for each of the color channels.
3) *Gaussian Noise*: Zero mean white Gaussian noise was added to the RGB components of the images (same noise variance were used for all the color channels). The noise standard deviation ranged from 0.071 to 0.316 pixels, using the *imnoise* MATLAB function.
4) *JPEG compression*: The *imwrite* functionality of MATLAB was used to compress the reference images using JPEG format. The bits-per-pixel (bpp) ranged from 0.0445 to 0.1843.
5) *Simulated Fast Fading Channel*: The original images were compressed into JPEG2000 bitstreams (with wireless error resilience features enabled and 64 x 64 code blocks) which were transmitted over a simulated Ralyleigh-faded channel. The signal-to-noise ratio was varied at the receiver from 14 to 17 dB to introduce different degrees of transmission errors.

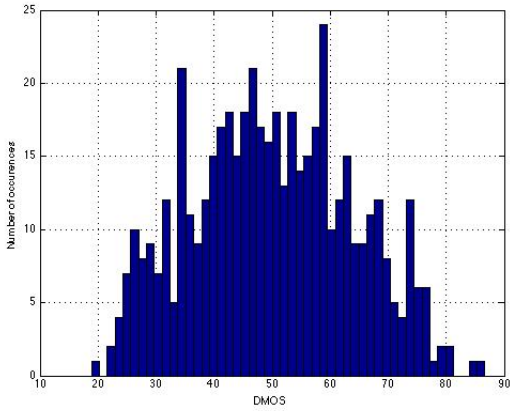

Fig. 2: Histogram of differential mean opinion scores (DMOS) for distorted images in the study. DMOS span most of the quality range.

C. Subjective Testing Methodology

Since the number of images to be evaluated (525) was prohibitively high for a double stimulus setup, a single stimulus setup was used. We followed the single stimulus

[1] All images are copyright of their rightful owners, and the authors do not claim ownership. No copyright infringement is intended. The database is to be used strictly for non-profit educational purposes.

continuous evaluation testing procedure in [10]. Subjects evaluated the reference undistorted images in the same session as the distorted images. This enabled us to derive the differential score for all of the test images.

Every image in the database was viewed by each subject, over three sessions of an hour each, separated by roughly 24 hours. Each session was divided into two sub-sessions of 25 minutes each separated by a break of five minutes in order to minimize visual fatigue. The 64 subjects who participated in the test were graduate and undergraduate students at The University of Texas at Austin (Fall 2014), with ages ranging from 18-30 years, mostly without prior experience in participation of subjective tests or image quality assessment. The gender ratio of the subject pool was roughly 1:1.

Before the start of the experiment, the purpose of the experiment was explained to each subject. A verbal confirmation of 20/20 (corrected) vision was also obtained. Subjects viewed roughly 175 test images during each session which were randomly ordered using a random number generator, and randomized for each subject. Each testing session was preceded by a short training session comprising of around 10 images in order to familiarize themselves with the testing setup. The training images were of different content, but had the same type of distortions as the test images.

D. Subjective Testing Display

The user interface for the study was designed on two identical PCs, one running Window and the other Linux, on MATLAB, using the Psychology Toolbox [11]. Both the PCs used identical NVIDIA Quadro NVS 285 GPUs and were interfaced with identical Dell 24 inches U2412M displays, which were roughly of the same age and having identical display settings. Each image was displayed on the screen for 12 seconds and the experiment was carried out under normal office illumination conditions. The subjects viewed the images from roughly 2 - 2.25 times of the display height.

The screen resolution was set at 1920×1200 pixels, but the images were displayed at their normal resolution (1920×1080) without any distortion introduced by interpolation. The top and bottom portions of the display were gray. At the end of the image display duration, a continuous quality scale was displayed on the screen, the default location of the slider was at the center of the scale. It was marked with five qualitative adjectives: "Bad", "Poor", "Fair", "Good", and "Excellent" placed at equal distances along the scale. After the subject has entered the rating for the image, the location of the slider along the scale was converted into a numerical score lying between [0,100], after rounding to the nearest integer. The subject could take as much time as needed to decide the score, but there was no provision of changing the score once entered or view the image again. The next image was displayed once the score was entered.

E. Processing of Raw Scores

The raw human subject ratings were processed using the same method as outlined in [10], [12]. In total 64 subjects participated in the study and each subject evaluated 525 images. 12 subjects were treated as outliers and the ratings obtained from the remaining 52 subjects were considered in the calculation of the final differential mean opinion score (DMOS) for each image. In order to take into account any variability in assigning the quality by the human observers per session, the difference scores were computed per session by subtracting the rating assigned by the subject to a distorted image from the rating assigned by the same subject to the corresponding reference video per session. The standard error in the DMOS scores was 0.6212 across distorted images.

III. Performance of objective IQA algorithms

In the paper, we have evaluated the performance of 23 state-of-art FR-IQA algorithms on the ESPL Synthetic Image Database, with the source code for the FR-IQA algorithms coming from [7] and [31]. Table **??** lists the IQA algorithms considered by name and paper citation. It also shows the Spearman's rank ordered correlation coefficient (ROCC) between the DMOS and the IQA predictions (after non-linear regression using a logistic function in [3]) and the outlier ratio (OR), defined as the percentage of the number of predictions outside the range of ±2 times of DMOS standard deviations [32]. Our study aims at benchmarking the performance of different categories of IQA algorithms over different distortion categories. Also, statistics of natural images are somewhat different from synthetic images [2], which should be kept in mind while studying the performance of IQA algorithms typically used for natural images in evaluating synthetic image distortions.

Overall, PSNR (row 22) is outperformed by other objective IQA algorithms (except for SSIM on row 23), but PSNR performs reasonably well for additive noise and fast-fading artifacts since it captures high-frequency distortions. The SSIM and MS-SSIM IQA algorithms, which perform exceedingly well on the LIVE database [3], shows a less impressive performance on our database, primarily due to the very low degree of correlation with human judgment on certain classes of distortions, such as interpolation, which has not been studied in any of the existing databases of natural images before. Almost all of the existing IQA algorithms fail to predict the subjective ratings for the interpolation artifact. Only MAD [5] achieves a reasonable performance, which advocates multiple strategies for determining the overall image quality, based on whether the distortions are near-threshold or supra-threshold. Low down-sampling factors result in near-threshold artifacts, which might appear almost imperceptible, especially at normal viewing distances. Blurred images also show a

	IQA	Interpolation		Blur		Additive Noise		JPEG Blocking		Fast Fading		Overall	
		ROCC	OR	ROCC	OR	ROCC	OR	ROCC	OR	ROCC	OR	ROCC	OR
1	Spectral Residual Based Similarity (SR-SIM) [13]	0.752	0	0.823	1	0.916	0	0.925	2	0.920	14	**0.880**	7.6
2	Feature Similarity Index (Color) (FSIMc) [14]	0.694	0	0.802	0	0.902	0	0.938	0	0.911	6	0.877	4.2
3	Feature Similarity Index (FSIM) [14]	0.692	0	0.801	0	0.902	0	**0.940**	0	0.907	5	0.876	4.6
4	Visual Saliency-Induced Index (VSI) [15]	0.692	1	0.811	0	0.914	0	0.880	0	0.923	13	0.872	5.6
5	Most Apparent Distortion (MAD) [5]	**0.788**	0	0.813	0	0.909	0	0.933	0	**0.927**	0	0.863	0.4
6	Gradient Similarity Measure (GSM) [16]	0.676	0	0.780	1	0.919	0	0.903	0	0.921	17	0.839	7.6
7	Information Content Weighted SSIM (IW-SSIM) [17]	0.761	0	0.823	0	0.902	0	0.933	1	0.925	3	0.827	1.0
8	Riesz-transform based Feature SIM (RFSIM) [18]	0.706	0	0.763	0	0.906	0	0.907	1	0.891	0	0.825	2.6
9	Gradient Magnitude Similarity Deviation (GMSD) [19]	0.716	0	0.791	0	**0.930**	0	0.862	0	0.920	1	0.821	0.8
10	PSNR-HVS(modified) (PHVSM) [20]	0.657	0	0.712	1	0.896	0	0.849	2	0.898	1	0.809	3.0
11	PSNR-HVS(modified)-A (PHMA) [21]	0.661	0	0.713	1	0.859	0	0.842	2	0.896	1	0.806	2.8
12	Information Fidelity Criterion (IFC) [22]	0.728	2	0.792	0	0.837	0	0.913	0	0.850	3	0.791	8.0
13	Noise Quality Measure (NQM) [23]	0.753	0	**0.837**	0	0.880	0	0.919	0	0.859	2	0.790	2.4
14	Weighted Signal-to-Noise ratio (WSNR) [24]	0.617	1	0.745	0	0.845	1	0.873	2	0.886	0	0.783	4.2
15	PSNR-HVS (PHVS) [25]	0.652	0	0.651	1	0.865	0	0.817	3	0.896	1	0.769	3.4
16	PSNR-HVS-A (PHA) [21]	0.639	0	0.652	1	0.834	0	0.808	3	0.890	1	0.765	3.4
17	Visual Information Fidelity (VIF) [26]	0.716	0	0.788	0	0.874	0	0.901	0	0.761	7	0.755	4.2
18	Multi-Scale SSIM (MS-SSIM) [27]	0.623	0	0.646	0	0.908	0	0.871	0	0.903	5	0.699	8.4
19	Pixel Domain Visual Information Fidelity (VIFP) [26]	0.651	0	0.624	1	0.895	0	0.878	1	0.791	9	0.693	5.4
20	Visual Signal-to-Noise ratio (VSNR) [28]	0.607	1	0.611	1	0.848	0	0.756	6	0.884	1	0.690	6.8
21	Universal Quality Index (UQI) [29]	0.703	0	0.673	0	0.815	0	0.918	2	0.840	2	0.682	5.0
22	PSNR	0.565	0	0.481	1	0.864	0	0.695	8	0.846	1	0.590	9.2
23	Structural Similarity Index (SSIM) [30]	0.463	2	0.440	0	0.909	0	0.633	11	0.797	6	0.542	11.2

TABLE I: Spearman's Rank Ordered Correlation Coefficient (ROCC) between the algorithm scores and the DMOS for various IQA Algorithms and the Outlier Ratio (OR). PSNR is Peak Signal-to-Noise Ratio. HVS stands for Human Visual System. OR for each distortion category has been calculated with 100 images and the overall OR has been calculated with 500 images. The bold values indicate the best performing algorithm for that category.

	PSNR	MS-SSIM	VIF	NQM	FSIM	IW-SSIM	SR-SIM	GMSD	MAD	PHVSM
PSNR	- - - - - -	- - - 0 - -	- 0 - 0 1 0	0 0 - 0 - 0	- 0 - 0 - 0	0 0 - 0 - 0	0 0 - 0 1 -	- 0 0 0 0 0	0 0 - 0 0 0	- 0 - 0 - 0
MSSIM	- - - 1 - -	- - - - - -	- 0 - - - 0	- 0 - - - 0	- 0 - - - 0	0 0 - - - 0	- - - - 1 -	- - - 0 0 0	0 0 - 0 0 0	- - - - 0 0
VIF	- 1 - 1 0 1	- 1 - - - 1	- - - - - -	- - - - - -	- - - 1 - -	- - - - - 0	- - - 1 - 1	- - 0 - 0 0	- - - - 0 0	- - - 1 0 0
NQM	1 1 - 1 - 1	- 1 - - - 1	- - - - - -	- - - - - -	- - - - - -	- - - - - 0	- 1 - 1 1 1	- - 0 - 0 -	- - - - 0 0	- 1 - 1 - -
FSIM	- 1 - 1 - 1	- 1 - - - 1	- - - 0 - -	- - - - - -	- - - - - -	- - - - - 0	- - - - - 1	- - - 0 0 0	- - - 0 0 0	- - - - 0 0
IW-SSIM	1 1 - 1 - 1	1 1 - - - 1	- - - - - 1	- - - - - 1	- - - - - 1	- - - - - -	- 1 1 1 1 1	- - - - 0 -	- - - 0 0 0	- 1 - - 0 -
SR-SIM	1 1 - 1 0 -	- - - - 0 -	- - - 0 - 0	- 0 - 0 0 0	- - - - - 0	- 0 0 0 0 0	- - - - - -	- - 0 0 0 0	- - - 0 0 0	- - - - 0 0
GMSD	- 1 1 1 1 1	- - - 1 1 1	- - 1 - 1 1	- - 1 - 1 -	- - - 1 1 1	- - - - 1 -	- - 1 1 1 1	- - - - - -	- - - - - 0	- - - 1 - -
MAD	1 1 - 1 1 1	1 1 - 1 1 1	- - - - 1 1	- - - - 1 1	- - - 1 1 1	- - - 1 1 1	- - - 1 1 1	- - - - - 1	- - - - - -	1 - - 1 - 1
PSNR-HVSM	- 1 - 1 - 1	- - - - 1 1	- - - 0 1 1	- 0 - 0 - -	- - - - 1 1	- 0 - - 1 -	- - - - 1 1	- - - 0 - -	0 - - 0 - 0	- - - - - -

TABLE II: Results of the F-test performed on the residuals between model predictions and DMOS values at 95% confidence intervals. In each cell, the symbol of 6 entries indicates "Interpolation', "Blur", "Additive Noise", "JPEG Blocking" , "Fast Fading" and "Overall" respectively. '1' ('0') indicates that the row IQA is statistically superior (inferior) than the column IQA, '-' implies statistical equivalence of the row and the column.

lower correlation with human scores. This indicates two avenues of future research. Firstly we would like to study the effects of varying display sizes on error visibility and secondly, we find a significant performance gap for this distortion category on which future researchers can work.

Overall, some of the recently proposed IQA algorithms, such as FSIM [14], VSI [15], SR-SIM [13] and MAD [5] are some of the algorithms that correlate best with human perception in terms of ROCC. FSIM takes into account image gradient magnitude and the phase congruency (a dimensionless measure of significance of local structure) and also uses it as a pooling strategy. VSI and SR-SIM uses more sophisticated pooling strategies based on visual fixations. Irrespective of whether the image is natural or synthetic, IQA algorithms that use more efficient pooling strategies by taking into account the localized distortions perform better than other IQA algorithms, as corroborated by [7]. Some of the IQA algorithms which model different aspects of the human visual system (HVS), such as NQM, VSNR, PSNR-HVSM, perform worse than the top performing signal driven IQA algorithms.

To determine whether the IQA algorithms are significantly different from each other, the F-statistic, as in [3] [12], was used to determine the statistical significance between the variances of the residuals after a non-linear logistic mapping between the two IQA algorithms. Table II shows the results for ten selected IQA algorithms and all distortions. The value of '1' ('0') indicates that the row IQA is statistically superior (inferior) than the column IQA, '-' implies statistical equivalence of the row and the column. Some of the best performing IQA algorithms, such as NQM, FSIM, IW-SSIM, GMSD, and MAD are found to be statistically superior to PSNR.

REFERENCES

[1] D. Kundu and B. L. Evans, "ESPL Synthetic Image Database Release 2," January 2015, http://signal.ece.utexas.edu/~bevans/synthetic/.

[2] D. Kundu and B. L. Evans, "Spatial domain synthetic scene statistics," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Nov 2014.

[3] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov 2006.

[4] N. N. Ponomarenko, O. Ieremeiev, V. V. Lukin, L. Jin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "A new color image database TID2013: Innovations and results," in *Advanced Concepts for Intelligent Vision Systems*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2013, vol. 8192, pp. 402–413.

[5] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, 2010.

[6] F. De Simone, L. Goldmann, V. Baroncini, and T. Ebrahimi, "Subjective evaluation of JPEG XR image compression," in *Proceedings of SPIE*, vol. 7443, 2009.

[7] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," in *Proc. IEEE International Conference on Image Processing*, Sept 2012, pp. 1477–1480.

[8] M. Čadík, R. Herzog, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts," *ACM Transactions on Graphics*, vol. 31, no. 6, pp. 147:1–147:10, Nov. 2012.

[9] R. Herzog, M. Čadík, T. O. Aydin, K. I. Kim, K. Myszkowski, and H.-P. Seidel, "NoRM: No-Reference image quality metric for realistic image synthesis." *Computer Graphics Forum*, vol. 31, no. 2, pp. 545–554, 2012.

[10] ITU-r BT.500-12 methodology for the subjective assessment of the quality of television pictures - IHS, inc. [Online]. Available: http://www.dii.unisi.it/~menegaz/DoctoralSchool2004/papers/ITU-R_BT.500-11.pdf

[11] M. Kleiner, D. Brainard, D. Pelli, C. Broussard, T. Wolf, and D. Niehorster, "The Psychology Toolbox," http://psychtoolbox.org/.

[12] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, June 2010.

[13] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in *Proc. IEEE International Conference on Image Processing*, Sept 2012, pp. 1473–1476.

[14] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, Aug 2011.

[15] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, Oct 2014.

[16] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1500–1512, April 2012.

[17] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[18] L. Zhang, D. Zhang, and X. Mou, "RFSIM: A feature based image quality assessment metric using riesz transforms," in *Proc. IEEE International Conference on Image Processing*, Sept 2010, pp. 321–324.

[19] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, Feb 2014.

[20] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On Between-Coefficient Contrast Masking of DCT Basis Functions," *Proc. of the Third International Workshop on Video Processing and Quality Metrics*, 2007.

[21] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, and M. Carli, "Modified image visual quality metrics for contrast change and mean shift accounting," in *Proc. of the International Conference The Experience of Designing and Application of CAD Systems in Microelectronics*, Feb 2011, pp. 305–311.

[22] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, Dec 2005.

[23] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Transactions on Image Processing*, vol. 9, pp. 636–650, 2000.

[24] T. Mitsa and K. L. Varkur, "Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, April 1993, pp. 301–304 vol.5.

[25] K. Egiazarian, J. Astola, V. Lukin, F. Battisti, and M. Carli, "New Full-Reference Quality Metrics based on HVS," *Proc. of the Second International Workshop on Video Processing and Quality Metrics*, 2006.

[26] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, Feb 2006.

[27] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. Asilomar Conference on Signals, Systems and Computers*, vol. 2, Nov 2003, pp. 1398–1402 Vol.2.

[28] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, Sept 2007.

[29] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, March 2002.

[30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[31] M. Gaubatz, "Metrix mux visual quality assessment package," http://foulard.ece.cornell.edu/gaubatz/metrix_mux/.

[32] "Final report from the video quality experts group on the validation of objective models of video quality assessment," ftp://vqeg.its.bldrdoc.gov/Documents/Meetings/Hillsboro_VQEG_Mar_03/VQEGIIDraftReportv2a.pdf, 2003.