A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms

Hamid Rahim Sheikh, *Member, IEEE*, Muhammad Farooq Sabir, and Alan Conrad Bovik, *Fellow, IEEE*

***Abstract*—Measurement of visual quality is of fundamental importance for numerous image and video processing applications, where the goal of quality assessment (QA) algorithms is to automatically assess the quality of images or videos in agreement with human quality judgments. Over the years, many researchers have taken different approaches to the problem and have contributed significant research in this area and claim to have made progress in their respective domains. It is important to evaluate the performance of these algorithms in a comparative setting and analyze the strengths and weaknesses of these methods. In this paper, we present results of an extensive subjective quality assessment study in which a total of 779 distorted images were evaluated by about two dozen human subjects. The "ground truth" image quality data obtained from about 25 000 individual human quality judgments is used to evaluate the performance of several prominent full-reference image quality assessment algorithms. To the best of our knowledge, apart from video quality studies conducted by the Video Quality Experts Group, the study presented in this paper is the largest subjective image quality study in the literature in terms of number of images, distortion types, and number of human judgments per image. Moreover, we have made the data from the study freely available to the research community [1]. This would allow other researchers to easily report comparative results in the future.**

***Index Terms*—Image quality assessment performance, image quality study, subjective quality assessment.**

I. Introduction

MACHINE evaluation of image and video quality is important for many image processing systems, such as those for acquisition, compression, restoration, enhancement, reproduction, etc. The goal of quality assessment research is to design algorithms for *objective* evaluation of quality in a way that is consistent with subjective human evaluation. By "consistent" we mean that the algorithm's assessments of quality should be in close agreement with human judgements, regardless of the type of distortion corrupting the image, the content of the image, or strength of the distortion.

Over the years, a number of researchers have contributed significant research in the design of full reference image quality assessment algorithms, claiming to have made headway in their respective domains. The QA research community realizes the importance of validating the performance of algorithms using extensive ground truth data, particularly against the backdrop of the fact that a recent validation study conducted by the video quality experts group (VQEG) discovered that the nine video QA methods that it tested, which contained some of the most sophisticated algorithms at that time, were "statistically indistinguishable" from the simple peak signal-to-noise ratio (PSNR) [2]. It is, therefore, imperative that QA algorithms be tested on extensive ground truth data if they are to become widely accepted. Furthermore, if this ground truth data, apart from being extensive in nature, is also publicly available, then other researchers can report their results on it for comparative analysis in the future.

Only a handful of QA validation literature has previously reported comparative performance of different image QA algorithms. In [3], [4], and [5], a number of mathematical measures of quality have been evaluated against subjective quality data. In [6], two famous visible difference predictors, by Daly [7] and Lubin [8], have been comparatively evaluated. In [9], three image quality assessment algorithms are evaluated against one another. In [10], an interesting new approach to IQM comparison is presented that compares two IQMs by using one IQM to expose the weaknesses of the other. The method is limited to differentiable IQMs only, and needs human subjective studies albeit of a different nature.

The reasons for conducting a new study were manifold. First, a number of interesting new QA algorithms have emerged since the work cited above, and it is interesting to evaluate the performance of these new algorithms as well. Second, previous studies did not contain some new, and important, image distortion types, such as JPEG2000 compression or wireless transmission errors, and were seldom diverse enough in terms of the distortion types or image content. In [3], the entire dataset was derived from only three reference images and distorted by compression distortion only, with a total of 84 distorted images. In [5], only 50 JPEG compressed images derived from a face database were used. The study presented in [9] also consisted of JPEG distortion only. The comparative study of [6] consisted of constructing visible difference maps only, and did not validate the ability of these algorithms to predict a graded loss of image quality. Third, few studies in the past have presented statistical significance testing, which has recently gained prominence in

Manuscript received November 10, 2005; revised May 2, 2006. This work was supported by the National Science Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhigang (Zeke) Fan.

H. R. Sheikh was with the Laboratory for Image and Video Engineering, Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712-1084 USA. He is now with Texas Instruments, Inc., Dallas, TX 75243 USA (e-mail: hamid.sheikh@ieee.org).

M. F. Sabir is with the Laboratory for Image and Video Engineering, Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712-1084 USA (e-mail: mfsabir@ece.utexas.edu).

A. C. Bovik is with the Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712-1084 USA (e-mail: bovik@ece.utexas.edu).

Color versions of Figs. 3–5 are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2006.881959

the QA research community. Fourth, in the context of statistical significance, the number of images in the study needs to be large so that QA algorithms can be discriminated with greater resolution. For example, if a QA metric *A* reports a linear correlation coefficient of, say, 0.93, on some dataset, while another metric *B* claims a correlation coefficient of 0.95 on the same set of images, then one can claim superiority of metric *B* over *A* with 95% confidence only if the data set had at least 260 images.[1] The number of images required is larger if the difference between correlation coefficients is smaller or if a greater degree of statistical confidence is required. Last, it is important to have large public-domain studies available so that researchers designing new QA algorithms can report the performance of their methods on them for comparative analysis against older methods. The public availability of VQEG Phase I data [12] has proven to be extremely useful for video QA research.

In this paper, we present our results of an extensive subjective quality assessment study, and evaluate the performance of ten prominent QA algorithms. The psychometric study contained 779 images distorted using five different distortion types and more than 25 000 human image quality evaluations. This study was diverse in terms of image content, distortion types, distortion strength, as well as the number of human subjects ranking each image. We have also made the data set publicly available [1] to facilitate future research in image quality assessment.

This paper is organized as follows. Section II gives the details of the experiment, including the processing of raw scores. Section III presents the results of the study, which are discussed in Section III-C. We conclude our paper in Section IV.

II. Details of the Experiment

A. Image Database

1) Source Image Content: The entire image database was derived from a set of source images that reflects adequate diversity in image content. Twenty-nine high resolution and high quality color images were collected from the Internet and photographic CD-ROMs. These images include pictures of faces, people, animals, closeup shots, wide-angle shots, nature scenes, man-made objects, images with distinct foreground/background configurations, and images without any specific object of interest. Fig. 1 shows a subset of the source images used in the study. Some images have high activity, while some are mostly smooth. These images were resized (using bicubic interpolation[2]) to a reasonable size for display on a screen resolution of 1024×768 that we had chosen for the experiments. Most images were 768×512 pixels in size. All distorted images were derived from the resized images.

2) Image Distortion Types: We chose to distort the source images using five different image distortion types that could occur in real-world applications. The distortion types are as follows.

- JPEG2000 compression: The distorted images were generated by compressing the reference images (full color) using JPEG2000 at bit rates ranging from 0.028 bits per pixel (bpp) to 3.15 bpp. Kakadu version 2.2 [13] was used to generated the JPEG2000 compressed images.
- JPEG compression: The distorted images were generated by compressing the reference images (full color) using JPEG at bit rates ranging from 0.15 bpp to 3.34 bpp. The implementation used was MATLAB's *imwrite* function.
- White Noise: White Gaussian noise of standard deviation σ_N was added to the RGB components of the images after scaling the three components between 0 and 1. The same σ_N was used for R, G, and B components. The values of σ_N used were between 0.012 and 2.0. The distorted components were clipped between 0 and 1, and rescaled to the range 0 to 255.
- Gaussian Blur: The R, G, and B components were filtered using a circular-symmetric 2-D Gaussian kernel of standard deviation σ_B pixels. The three color components of the image were blurred using the same kernel. The values of σ_B ranged from 0.42 to 15 pixels.
- Simulated Fast Fading Rayleigh (wireless) Channel: Images were distorted by bit errors during transmission of compressed JPEG2000 bitstream over a simulated wireless channel. Receiver SNR was varied to generate bitstreams corrupted with different proportion of bit errors. The source JPEG2000 bitstream was generated using the same codec as above, but with error resilience features enabled, and with 64×64 precincts. The source rate was fixed to 2.5 bits per pixel for all images, and no error concealment algorithm was employed. The receiver SNR used to vary the distortion strengths ranged from 15.5 to 26.1 dB.

Fig. 1. Some source images used in the study.

These distortions reflect a broad range of image impairments, from smoothing to structured distortion, image-dependent distortions, and random noise. The level of distortion was varied to generate images at a broad range of quality, from imperceptible levels to high levels of impairment. Fig. 4 shows how the subjective quality (after outlier removal and score processing as mentioned in Section II-C) varies with the distortion strength for each of the distortion types. Fig. 5 shows the histogram of the subjective scores for the entire dataset.

B. Test Methodology

The experimental setup that we used was a single-stimulus methodology in which the reference images were also evalu-

[1]This assumes a hypothesis test done using Fisher's Z transformation [11].

[2]Since we derive a quality difference score (DMOS) for each distorted image, any loss in quality due to resizing appears both in the reference and the test images and cancels out in the DMOS scores.

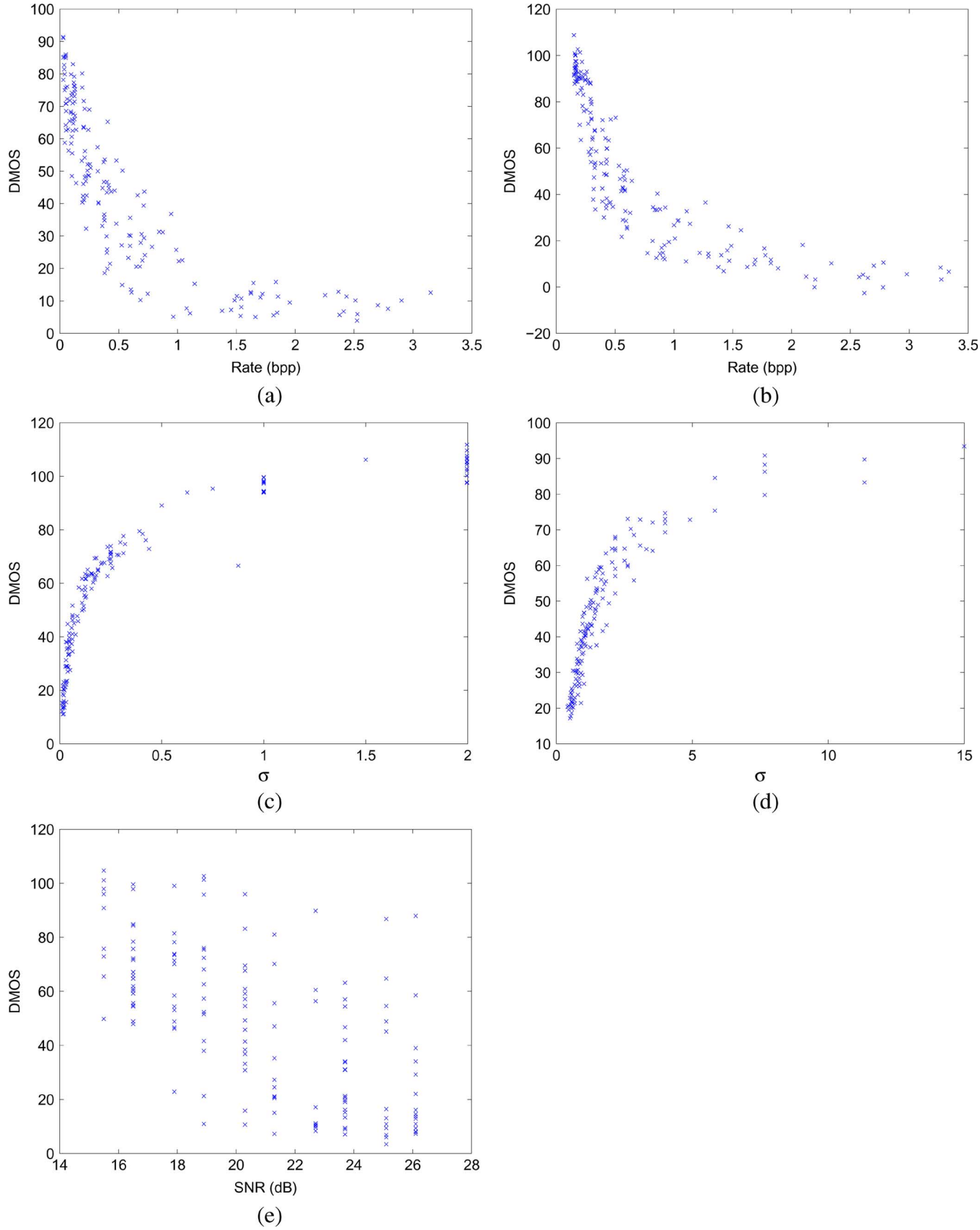


Fig. 4. Dependence of quality on distortion parameters for different distortion types. (a) JPEG2000: DMOS versus rate. (b) JPEG: DMOS versus rate. (c) Noise: DMOS versus σ_N. (d) Blur: DMOS versus σ_B. (e) FF: DMOS versus SNR.

ated in the same experimental session as the test images. A single-stimulus setup was chosen instead of a double-stimulus setup because the number of images to be evaluated was prohibitively large for a double-stimulus study (we evaluated a total

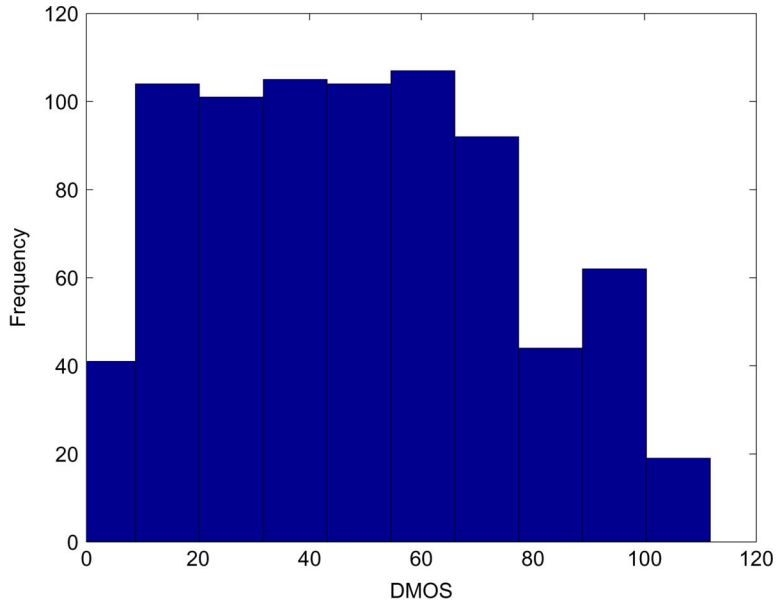


Fig. 5. Histogram of subjective quality (DMOS) of the distorted images in the database.

of 779 distorted images).[3] However, since the reference images were also evaluated by each subject in each session, a quality difference score can be derived for all distorted images and for all subjects.

1) Equipment and Display Configuration: The experiments were conducted using identical Microsoft Windows workstations. A web-based interface showing the image to be ranked and a Java scale-and-slider applet for assigning a quality score was used. The workstations were placed in an office environment with normal indoor illumination levels. The display monitors were all 21-in CRT monitors displaying at a resolution of 1024×768 pixels. Although the monitors were not calibrated, they were all approximately the same age, and set to the same display settings. Subjects viewed the monitors from an approximate viewing distance of 2–2.5 screen heights.

The experiments were conducted in seven sessions: two sessions for JPEG2000, two for JPEG, and one each for white noise, Gaussian blur, and channel errors. Each session included the full set of reference images randomly placed among the distorted images. The number of images in each session is shown in Table I.

2) Human Subjects, Training, and Testing: The bulk of the subjects taking part in the study were recruited from the Digital Image and Video Processing (undergraduate) and the Digital Signal Processing (graduate) classes at the University of Texas, Austin, over a course of two years. The subject pool consisted of mostly male students inexperienced with image quality assessment and image impairments. The subjects were not tested for vision problems, and their verbal expression of the soundness of their (corrected) vision was considered sufficient. The average number of subjects ranking each image was about 23 (see Table I).

Each subject was individually briefed about the goal of the experiment, and given a demonstration of the experimental procedure. A short training showing the approximate range of quality of the images in each session was also presented to each subject. Images in the training sessions were different from those used in the actual experiment. Generally, each subject participated in one session only. Subjects were shown images in a random order; the randomization was different for each subject. The subjects reported their judgments of quality by dragging a slider on a quality scale. The position of the slider was automatically reset after each evaluation. The quality scale was unmarked numerically. It was divided into five equal portions, which were labeled with adjectives: "Bad," "Poor," "Fair," "Good," and "Excellent." The position of the slider after the subject ranked the image was converted into a quality score by linearly mapping the entire scale to the interval [1 100], rounding to the nearest integer. In this way, the raw quality scores consisted of integers in the range 1–100.

3) Double Stimulus Study for Psychometric Scale Realignment: Ideally, all images in a subjective QA study should be evaluated in one session so that scale mismatches between subjects are minimized. Since the experiment needs to be limited to a recommended maximum of 30 min [14] to minimize effects of observer fatigue, the maximum number of images that can be evaluated is limited. The only way of increasing the number of images in the experiment is to use multiple sessions using different sets of images. In our experiment, we used seven such sessions. While we report the performance of IQMs on individual sessions, we also report their performance on aggregated datapoints from all sessions. The aggregation of datapoints from the seven sessions into one dataset requires scale realignment.

Since the seven sessions were conducted independently, there is a possibility of misalignment of their quality scales. Thus, it may happen that a quality score of, say, 30 from one session may not be subjectively similar to a score of 30 from another session. Such scale mismatch errors are introduced primarily because the distribution of quality of images in different

[3]A double-stimulus procedure typically requires 3–4 times more time per image than a single-stimulus procedure.

TABLE I
SUBJECTIVE EVALUATION SESSIONS: NUMBER OF IMAGES IN EACH SESSION AND THE NUMBER OF SUBJECTS PARTICIPATING IN EACH SESSION. THE REFERENCE IMAGES WERE INCLUDED IN EACH SESSION. THE ALIGNMENT STUDY WAS A DOUBLE-STIMULUS STUDY

Session	Number of images	Number of subjects
JPEG2000 #1	116	29
JPEG2000 #2	111	25
JPEG #1	116	20
JPEG #2	117	20
White noise	174	23
Gaussian blur	174	24
Fast-fading wireless	174	20
Total	982	22.8 (average)
Alignment study	50	32

sessions is different. Since these differences are virtually impossible to predict before the design of the experiment, they need to be compensated for by conducting scale realignment experiments *after* the experiment.

In our study, after completion of the seven sessions, a set of 50 images was collected from the seven session and used for a separate realignment experiment. The realignment experiment used a double-stimulus methodology for more accurate measurement of quality for realignment purposes. Five images were chosen from each session for JPEG2000 and JPEG distortions, and ten each from the other three distortion types. The images chosen from each session roughly covered the entire quality range for that session. The double-stimulus study consisted of *view A view B score A score B* sequence where A and B were (randomly) the reference or the corresponding test images. DMOS scores for the double-stimulus study were evaluated using the recommendations adapted from [15] on a scale of 0–100. Details of processing of scores for realignment are mentioned in Section II-C.

C. Processing of Raw Data

1) Outlier Detection and Subject Rejection: A simple outlier detection and subject rejection algorithm was chosen. Raw difference score for an image was considered to be an outlier if it was outside an interval of width Δ standard deviations about the mean score for that image, and for any session, all quality evaluations of a subject were rejected if more than R of his evaluations in that session were outliers. This outlier rejection algorithm was run twice. A numerical minimization algorithm was run that varied Δ and R to minimize the average width of the 95% confidence interval. Average values of Δ and R were 2.33 and 16, respectively. Overall, a total of four subjects were rejected, and about 4% of the difference scores were rejected as being outliers (where we count all datapoints of rejected subjects as outliers).

2) DMOS Scores: For calculation of DMOS scores, the raw scores were first converted to raw quality difference scores

$$d_{ij} = r_{i\mathrm{ref}(j)} - r_{ij} \tag{1}$$

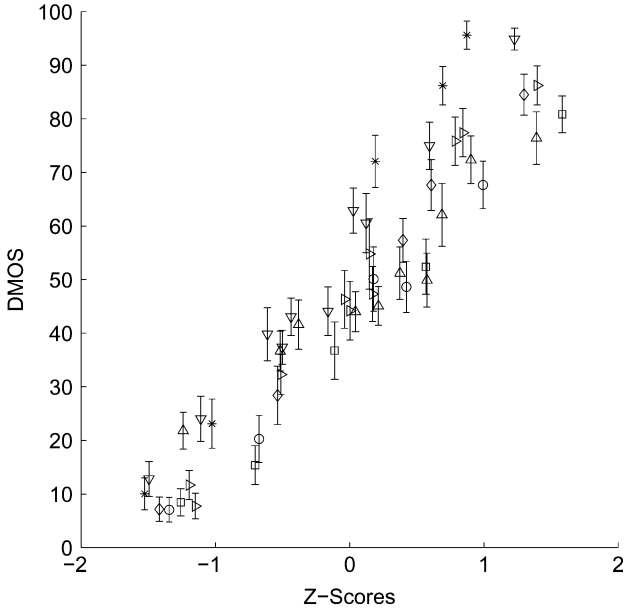


Fig. 2. Z scores and DMOS values for images used in the realignment study: JPEG2000 #1 ($\circ$), JPEG2000 #2 ($\diamond$), JPEG #1 ($\square$), JPEG #2 ($*$), white noise ($\triangledown$), Gaussian blur ($\triangle$), fast-fading wireless ($\triangleright$).

where r_{ij} is the raw score for the ith subject and jth image, and $r_{i\mathrm{ref}(j)}$ denotes the raw quality score assigned by the ith subject to the reference image corresponding to the jth distorted image. The raw difference scores d_{ij} for the ith subject and jth image were converted into Z scores (after outlier removal and subject rejection) [16]

$$z_{ij} = (d_{ij} - \bar{d}_i)/\sigma_i \tag{2}$$

where $\bar{d}_i$ is the mean of the raw difference scores over all images ranked by the subject i, and σ_i is the standard deviation. The Z scores were then averaged across subjects to yield $\bar{z}_j$ for the jth image.

The results of the realignment experiment were used to map the Z scores to DMOS. Fig. 2 shows the Z scores and the DMOS for the 50 images that were used in the realignment study, together with the confidence intervals. To obtain DMOS values for the entire database, we assume a linear mapping between Z scores and DMOS: $\mathrm{DMOS}(\bar{z}) = p_1\bar{z}+p_2$. The values for p_1 and p_2 are learned by minimizing the error between the $\mathrm{DMOS}(\bar{z}_j)$ and DMOS_j, where DMOS_j values were obtained from the realignment study. Seven such mappings were learned for the seven sessions and applied to the Z scores of all images in the respective sessions to yield the realigned DMOS for the entire database. Fig. 3 shows the mapping between the Z scores and the final realigned DMOS for the entire dataset.

D. Quality Assessment Algorithms

The image quality metrics (IQMs) tested in this paper are summarized in Table II. The implementations of the algorithms were either publicly available on the Internet or obtained from the authors. Although all of the methods in [5] were tested using the code available at the author's website [17], only the block-wise spectral distance measure (S4 and derivatives) performed better than PSNR. Similarly, all similarity measures presented

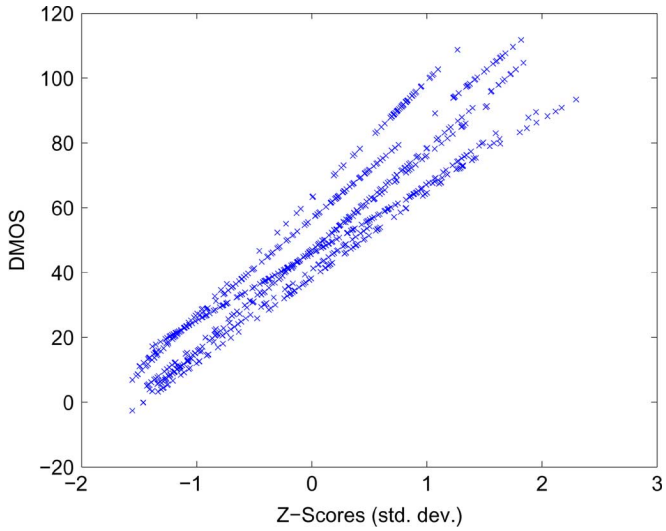


Fig. 3. Z scores to DMOS mapping.

TABLE II
QA ALGORITHMS TESTED IN THIS PAPER. EXTENDING IQMS THAT OPERATE ON LUMINANCE ONLY TO COLOR MAY IMPROVE THEIR PERFORMANCE, BUT THE INVENTORS REPORTED RESULTS ON LUMINANCE COMPONENTS ONLY

QA Algorithm	Remarks
PSNR	Luminance component only
Sarnoff JND Metrix	[19]. Based on the well-known work [8]. Works on color images.
DCTune	[20]. Originally designed for JPEG optimization. Works on color images.
PQS	[21]. Public C implementation available at [22]. Works with luminance only.
NQM	[23]. Works with luminance only.
Fuzzy S7	[18]. Works with luminance only.
BSDM (S4)	S4 in [5]. Works on color images.
Multiscale SSIM	[24]. Works with luminance only.
IFC	[25]. Works with luminance only.
VIF	[26]. Works with luminance only.

in [18] were tested, but only the highest performing similarity measure, S_7^n is presented here.

Although the list of algorithms reported in this paper is not exhaustive, it is representative of many classes of QA algorithms: HVS model based (JND), feature based (PQS), HVS models combined with application specific modeling (DCTune, NQM), mathematical distance formulations (Fuzzy S7, BSDM), structural [SSIM(MS)], and information-theoretic (IFC, VIF) frameworks. The public availability of the database easily allows future researchers to evaluate their methods and put them in context with prior work, including the methods evaluated in this paper.

III. RESULTS

A. Performance Metrics

We have chosen three performance metrics for this paper. The first metric is the linear correlation coefficient (CC) between DMOS and algorithm score after nonlinear regression. The nonlinearity chosen for regression for each of the methods tested was a five-parameter logistic function (a logistic function with an added linear term, constrained to be monotonic)

$$\text{Quality}(x) = \beta_1 \text{logistic}(\beta_2, (x - \beta_3)) + \beta_4 x + \beta_5 \quad (3)$$

$$\text{logistic}(\tau, x) = \frac{1}{2} - \frac{1}{1 + \exp(\tau x).} \quad (4)$$

This nonlinearity was applied to the IQM score or its logarithm, whichever gave a better fit for all data. The second metric is root-mean-squared error (RMSE) between DMOS and the algorithm score after nonlinear regression. The third metric is Spearman rank order correlation coefficient (SROCC). The results, presented in Tables III–V, are reported for the different methods on individual datasets as well as on the entire dataset.

B. Statistical Significance and Hypothesis Testing

In order to ascertain which differences between IQM performance are statistically significant, two hypothesis test were conducted. These tests tell us whether our confidence in the estimate of an IQM's performance, based on the number of sample points used, allows us to make a statistically sound conclusion of superiority or inferiority of an IQM. A variance-based hypothesis test (HT) using the residuals between DMOS and the quality predicted by the IQM (after the nonlinear mapping), and another variance-based test was conducted using the residuals between the IQM (after nonlinear mapping) and the individual subject ratings [15].

1) Hypothesis Testing Based on DMOS: The test is based on an assumption of Gaussianity of the residual differences between the IQM prediction and DMOS, and uses the F-statistic for comparing the variance of two sets of sample points [27]. The test statistic on which the hypothesis test is based is the ratio of variances (or, equivalently, the standard deviation), and the goal is to determine whether the two sample sets come from the same distribution or not. The Null Hypothesis is that the residuals from one IQM come from the same distribution and are statistically indistinguishable (with 95% confidence) from the residuals from another IQM. Based on the number of distorted images in each dataset, the threshold ratio of standard deviations below which the two sets of residuals are statistically indistinguishable can be derived using the F-distribution [27]. In this paper, we use a simple Kurtosis-based criterion for Gaussianity: if the residuals have a kurtosis between 2 and 4, they are taken to be Gaussian. The results of the hypothesis test using the residuals between DMOS and IQM predictions are presented in Table VI, whereas the results of the Gaussianity test are given in Table VII.

2) Hypothesis Testing Based on Individual Quality Scores: We also present results of statistical analysis based on individual quality scores, that is, without averaging across subjects. This allows us to compare IQM against the theoretical best method,

TABLE III
LINEAR CORRELATION COEFFICIENT AFTER NONLINEAR REGRESSION

	JP2K#1	JP2K#2	JPEG#1	JPEG#2	WN	GBlur	FF	All data
PSNR	0.9332	0.8740	0.8856	0.9167	0.9859	0.7834	0.8895	0.8709
JND	0.9649	0.9734	0.9605	0.9870	0.9687	0.9457	0.9129	0.9266
DCTune	0.8486	0.7834	0.8825	0.9418	0.9288	0.7095	0.7693	0.8046
PQS	0.9391	0.9364	0.9320	0.9777	0.9603	0.9216	0.9317	0.9243
NQM	0.9508	0.9463	0.9387	0.9783	0.9885	0.8858	0.8367	0.9075
Fuzzy S7	0.9360	0.9133	0.9134	0.9509	0.9038	0.5975	0.9120	0.8269
BSDM (S4)	0.9144	0.9450	0.9199	0.9747	0.9559	0.9619	0.9596	0.9335
SSIM(MS)	0.9702	0.9711	0.9699	0.9879	0.9737	0.9487	0.9304	0.9393
IFC	0.9421	0.9626	0.9209	0.9744	0.9766	0.9691	0.9631	0.9441
VIF	0.9791	0.9787	0.9714	0.9885	0.9877	0.9762	0.9704	0.9533

TABLE IV
ROOT-MEAN-SQUARED ERROR AFTER NONLINEAR REGRESSION

	JP2K#1	JP2K#2	JPEG#1	JPEG#2	WN	GBlur	FF	All data
PSNR	8.4858	12.7437	10.7793	13.5633	4.6689	11.4402	12.9716	13.4265
JND	6.1989	6.0092	6.4629	5.4507	6.9217	5.9812	11.5860	10.2649
DCTune	12.4881	16.2971	10.9168	11.4089	10.3340	12.9712	18.1366	16.2150
PQS	8.1135	9.2064	8.4146	7.1305	7.7803	7.1445	10.3119	10.4221
NQM	7.3167	8.4762	8.0004	7.0336	4.2250	8.5413	15.5467	11.4698
Fuzzy S7	8.3102	10.6828	9.4481	10.5070	11.9316	14.7596	11.6437	15.3544
BSDM (S4)	9.5533	8.5797	9.1009	7.5826	8.1888	5.0336	7.9907	9.7934
SSIM(MS)	5.7214	6.2584	5.6575	5.2729	6.3577	5.8225	10.4023	9.3691
IFC	7.9184	7.1050	9.0470	7.6267	5.9990	4.5380	7.6398	9.0007
VIF	4.7986	5.3864	5.5083	5.1274	4.3598	3.9905	6.8553	8.2459

TABLE V
SPEARMAN RANK ORDER CORRELATION COEFFICIENT

	JP2K#1	JP2K#2	JPEG#1	JPEG#2	WN	GBlur	FF	All data
PSNR	0.9263	0.8549	0.8779	0.7708	0.9854	0.7823	0.8907	0.8755
JND	0.9646	0.9608	0.9599	0.9150	0.9487	0.9389	0.9045	0.9291
DCTune	0.8335	0.7209	0.8702	0.8200	0.9324	0.6721	0.7675	0.8032
PQS	0.9372	0.9147	0.9387	0.8987	0.9535	0.9291	0.9388	0.9304
NQM	0.9465	0.9393	0.9360	0.8988	0.9854	0.8467	0.8171	0.9049
Fuzzy S7	0.9316	0.9000	0.9077	0.8012	0.9199	0.6056	0.9074	0.8291
BSDM (S4)	0.9130	0.9378	0.9128	0.9231	0.9327	0.9600	0.9372	0.9271
SSIM(MS)	0.9645	0.9648	0.9702	0.9454	0.9805	0.9519	0.9395	0.9527
IFC	0.9386	0.9534	0.9107	0.9005	0.9625	0.9637	0.9556	0.9459
VIF	0.9721	0.9719	0.9699	0.9439	0.9828	0.9706	0.9649	0.9584

or the Null Model, which predicts the DMOS values exactly, that is, $\mathrm{IQM} = \mathrm{DMOS}$. The goal is to determine whether any IQM is statistically indistinguishable from the Null Model. Table VIII presents the standard deviation of the residuals between different IQMs (after nonlinear mapping) and individual subjective scores. The individual scores were obtained by applying the same scale transformations (see Section II-C) on the Z scores z_{ij} as were used for the averaged Z scores $\bar{z}_j$. A variance based Hypothesis test was conducted using the standard deviations shown in Table VIII, while Table X shows which residuals can be assumed to have a Gaussian distribution. Table IX shows the results of the hypothesis tests.

C. Discussion

1) IQM Evaluation Criteria: The results presented in Tables III and IV are essentially identical, since they both are measuring the correlation of an IQM after an identical nonlinear fit.

TABLE VI
STATISTICAL SIGNIFICANCE MATRIX BASED ON IQM-DMOS RESIDUALS. EACH ENTRY IN THE TABLE IS A CODEWORD CONSISTING OF EIGHT SYMBOLS. THE POSITION OF THE SYMBOL IN THE CODEWORD REPRESENTS THE FOLLOWING DATASETS (FROM LEFT TO RIGHT): JPEG2000 #1, JPEG2000 #2, JPEG #1, JPEG #2, WHITE NOISE, GAUSSIAN BLUR, FAST-FADING WIRELESS, ALL DATA. EACH SYMBOL GIVES THE RESULTS OF THE HYPOTHESIS TEST ON THE DATASET REPRESENTED BY THE SYMBOL'S POSITION: 1 MEANS THAT THE IQM FOR THE ROW IS STATISTICALLY BETTER THAN THE IQM FOR THE COLUMN, 0 MEANS THAT IT IS STATISTICALLY WORSE, AND "–" MEANS THAT IT IS STATISTICALLY INDISTINGUISHABLE

	PSNR	JND	DCTune	PQS	NQM	Fuzzy S7	BSDM (S4)	SSIM(MS)	IFC	VIF
PSNR	--------	000010-0	11--1-11	-0001000	-000-010	---011-1	-0-01000	00001000	-0-01000	0000-000
JND	111101-1	--------	11111111	1111-1--	-1110111	111111-1	1111100-	-------0	1-110000	0---0000
DCTune	00--0-00	00000000	--------	00000000	00000000	00--1-0-	00000000	00000000	00000000	00000000
PQS	-1110111	0000-0--	11111111	--------	----0111	---111-1	-----000	000000-0	-0--0000	00000000
NQM	-111-101	-0001000	11111111	----1000	--------	-1-11101	1---1000	00001000	----1000	0000-000
Fuzzy S7	---100-0	000000-0	11--0-1-	---000-0	-0-00010	--------	-0-00000	000000-0	-0-00000	00000000
BSDM (S4)	-1-10111	0000011-	11111111	-----111	0---0111	-1-11111	--------	0000011-	00--0--0	00000000
SSIM(MS)	11110111	-------1	11111111	111111-1	11110111	111111-1	1111100-	--------	1-11-00-	----0000
IFC	-1-10111	0-001111	11111111	-1--1111	----0111	-1-11111	11--1--1	0-00-11-	--------	00000--0
VIF	1111-111	1---1111	11111111	11111111	1111-111	11111111	11111111	----1111	11111--1	--------

TABLE VII
GAUSSIANITY OF IQM-DMOS RESIDUALS. "1" MEANS THAT THE RESIDUALS CAN BE ASSUMED TO HAVE A GAUSSIAN DISTRIBUTION SINCE THE KURTOSIS LIES BETWEEN 2 AND 4

	JP2K#1	JP2K#2	JPEG#1	JPEG#2	WN	GBlur	FF	All data
PSNR	1	1	1	0	1	1	1	1
JND	1	1	1	1	1	1	0	0
DCTune	0	1	1	0	1	0	1	0
PQS	1	1	0	0	1	1	1	1
NQM	1	1	1	1	1	1	1	0
Fuzzy S7	1	1	1	0	1	1	1	1
BSDM (S4)	1	1	1	1	1	1	0	1
SSIM(MS)	1	1	1	1	1	1	0	1
IFC	1	1	1	1	1	1	0	1
VIF	1	1	0	1	1	1	0	1

However, RMSE is a more intuitive criterion of IQM comparison than the linear correlation coefficient because the latter is a nonuniform metric, and it is not easy to judge the relative performance of IQMs especially if the metrics are all doing quite well. The QA community is more accustomed to using the correlation coefficient, but RMSE gives a more intuitive feel of the relative improvement of one IQM over another. Nevertheless, one can derive identical conclusions from the two tables.

The Spearman rank order correlation coefficient merits further discussion. SROCC belongs to the family of nonparametric correlation measures, as it does not make any assumptions about the underlying statistics of the data. For example, SROCC is independent of the monotonic nonlinearity used to fit the IQM to DMOS. However, SROCC operates only on the rank of the data points, and ignores the relative distance between datapoints. For this reason, it is generally considered to be a less sensitive measure of correlation, and is typically used only when the number of datapoints is small [28]. In Table V, although the relative performance of IQMs is essentially the same as in Table III, there are some interesting exceptions. For example, one can see that on JPEG #2 dataset, Sarnoff JNDMetrix performs almost as good as SSIM (MS) or VIF [in fact, it is statistically indistinguishable from VIF or SSIM (MS) on this dataset], but the SROCC shows a different conclusion, that is, JND is much worse than SSIM (MS) or VIF. A further analysis of the data and the nonlinear fits reveals that the fits of the three methods are nearly identical (see Fig. 6), and the linear correlation coefficient, or RMSE is indeed a good criterion for comparison. It is also easy to see why SROCC is not a good criterion for this dataset. One can note that many datapoints lie in that region of the graph where the DMOS is less than 20 (high quality), and all of these images are high quality with approximately the same quality rank. The difference between DMOS of these images is essentially measurement noise. Since SROCC ignores the relative magnitude of the residuals, it treats this measurement noise with the same importance as datapoints in other parts of the graph, where the quality rank spans a larger range of the measurement scale. Thus, in this case, the nature of the data is such that SROCC is not a good measure of IQM performance, and leads to misleading conclusions.

We believe that when the underlying quantity exhibits a qualitative saturation, such as when image quality is perceptually the same for many distortion strengths, SROCC is not a good validation criterion, and RMSE is more suitable.

2) Cross-Distortion Validation: Tables III–V also allow us to compare different IQM against one another on individual distortion types as well as on all data. We see that while some metrics perform very well for some distortions, their performance is much worse for other distortion types. PSNR for example, is among the best methods for the white noise distortion, but its performance for other distortion types is much worse than other IQMs. Moreover, all of the methods perform better than PSNR on at least one of the datasets using at least one of the criteria! This underlies the importance of extensive validation of full reference IQM over distortion types as well as statistical significance testing.

TABLE VIII
STANDARD DEVIATIONS OF THE RESIDUALS BETWEEN DIFFERENT IQMS AND SUBJECT SCORES. THE TABLE ALSO GIVES THE NUMBER OF INDIVIDUAL SCORES IN EACH DATASET

	JP2K#1	JP2K#2	JPEG#1	JPEG#2	WN	GBlur	FF	All data
Data points	2398	1879	1504	1579	3311	3403	2878	16952
PSNR	14.0187	17.7418	16.0393	17.6145	12.5156	14.7832	18.4089	17.4095
JND	12.7536	13.6865	13.4591	12.4137	13.5250	11.0678	17.4220	15.2179
DCTune	16.8368	20.4284	16.1153	15.9592	15.5692	15.9808	22.3555	19.6881
PQS	13.8184	15.4409	14.5328	13.2059	13.9788	11.7365	16.6407	15.3918
NQM	13.3480	14.9805	14.2746	13.1875	12.3555	12.6605	20.2826	16.0460
Fuzzy S7	13.9155	16.2779	15.1627	15.3378	16.6603	17.4870	17.5017	19.0501
BSDM (S4)	14.7145	15.0257	14.9488	13.4766	14.2101	10.5723	15.3048	14.9935
SSIM(MS)	12.5328	13.8039	13.0988	12.3362	13.2394	10.9694	16.6708	14.7366
IFC	13.7056	14.2344	14.9091	13.4828	13.0657	10.3416	15.1184	14.4437
VIF	12.1360	13.4389	13.0252	12.2638	12.4004	10.1183	14.7365	14.0411
NULL	11.1208	12.2761	11.7833	11.1437	11.6034	9.2886	13.0302	11.4227

TABLE IX
IQM COMPARISON AND STATISTICAL SIGNIFICANCE MATRIX USING A 95% CONFIDENCE CRITERION ON IQM-SUBJECT SCORE RESIDUALS. SEE CAPTION OF TABLE VI FOR INTERPRETATION OF THE TABLE ENTRIES. THE NULL MODEL IS THE DMOS

	PSNR	JND	DCTune	PQS	NQM	Fuzzy S7	BSDM (S4)	SSIM(MS)	IFC	VIF	NULL
PSNR	--------	00001000	11-01111	-0001000	0000-010	-0001101	10001000	00001000	-0001000	0000-000	00000000
JND	11110111	--------	11111111	1111110-	11110111	111111-1	11111000	------00	11110000	0---0000	00000000
DCTune	00-10000	00000000	--------	00000000	00000000	000-1100	00000000	00000000	00000000	00000000	00000000
PQS	-1110111	0000001-	11111111	--------	0---0111	-1-11111	1----000	000000-0	-0--0000	00000000	00000000
NQM	1111-101	00001000	11111111	1---1000	--------	11111101	1-1-1000	00001000	-01-1000	0000-000	00000000
Fuzzy S7	-1110010	000000-0	111-0011	-0-00000	00000010	--------	10-00000	00000000	-0-00000	00000000	00000000
BSDM (S4)	01110111	00000111	11111111	0----111	0-0-0111	01-11111	--------	00000110	00--0--0	00000000	00000000
SSIM(MS)	11110111	------11	11111111	111111-1	11110111	11111111	11111001	--------	1-11-000	----0000	00000000
IFC	-1110111	00001111	11111111	-1--1111	-10-0111	-1-11111	11--1--1	0-00-111	--------	00000--0	00000000
VIF	1111-111	1---1111	11111111	11111111	1111-111	11111111	11111111	----1111	11111--1	--------	00000000
NULL	11111111	11111111	11111111	11111111	11111111	11111111	11111111	11111111	11111111	11111111	--------

TABLE X
GAUSSIANITY OF IQM-SUBJECT SCORE RESIDUALS. "1" MEANS THAT THE RESIDUALS CAN BE ASSUMED TO HAVE A GAUSSIAN DISTRIBUTION SINCE THE KURTOSIS LIES BETWEEN 2 AND 4

	JP2K#1	JP2K#2	JPEG#1	JPEG#2	WN	GBlur	FF	All data
PSNR	1	1	0	0	1	1	1	1
JND	1	1	0	1	1	1	1	1
DCTune	1	1	0	0	1	1	1	1
PQS	1	1	0	1	1	1	1	1
NQM	1	1	0	1	1	1	1	1
Fuzzy S7	1	1	0	1	1	1	1	1
BSDM (S4)	1	1	0	1	1	1	1	1
SSIM(MS)	1	1	0	1	1	1	1	1
IFC	1	1	0	1	1	1	1	1
VIF	1	1	0	1	1	1	1	1
NULL	1	1	0	1	1	1	1	1

3) Statistical Significance: We have devoted a large portion of our paper to presenting results of statistical significance tests as this topic has recently gained importance in the quality assessment research community. The two types of tests presented in this paper are the same as used by the Video Quality Experts Group (VQEG) in their Phase-II report [15], which also discusses the merits and demerits of tests based on DMOS versus tests based on subject score residuals. The results allow us to conclude whether the numerical differences between IQM performance is statistically significant or not. Note that this assessment is purely from the standpoint of the number of samples used in the experiments.

Statistical significance testing allows one to make quantitative statements about the soundness of the conclusion of the experiment solely from the point of view of the number of sample points used in making the conclusion. It is different from *practical* significance, that is, whether a difference in performance is of any practical importance. One IQM may be statistically inferior, but it may be computationally more suited for a particular application. Another IQM may be statistically inferior only when tested on thousands of images. Moreover, the selection of the confidence criterion is also somewhat arbitrary and it obviously affects the conclusions being drawn.

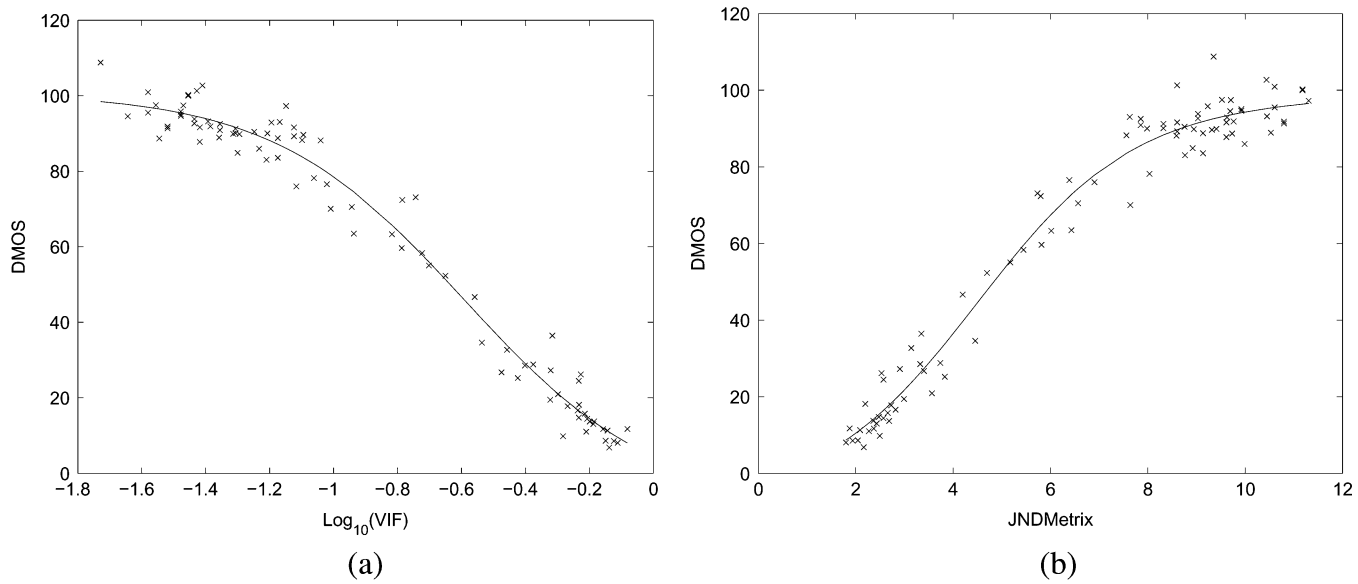


Fig. 6. Nonlinear fits for VIF and Sarnoff's JND-Metrix for the JPEG #2 dataset. Note that while the fits are nearly identically good, as also quantified by CC and RMSE in Tables III and VI, the SROCC leads to a much different conclusion (in Table V). (a) VIF. (b) JNDMetrix.

For example, if one uses a 90% confidence criterion, then many more IQMs could be statistically distinguished from one another.

From Tables VI and IX, it may seem that the HT using DMOS should always be less sensitive than HT using IQM-subject score residuals since the latter has far greater number of data points. This is not necessarily the case since the averaging operation in the calculation of DMOS reduces the contribution of the variance of subject scores. Thus, depending upon the standard deviation of the subject scores, either of the two variance based HT may be more sensitive.

The HTs presented in the paper rely on the assumption of Gaussianity of the residuals. We see from Tables VII and X that this assumption is not always met. However, we still believe that the results presented in Tables VI and IX are quite representative. This is because, for the large number of sample points used in the HT, the Central Limit Theorem comes into play and the distribution of the variance estimates (on which the HTs are based) approximates the Gaussian distribution. We verified this hypothesis by running two Monte-Carlo simulations. In the first simulation, we drew two sets of DMOS residual samples randomly from two Gaussian distributions corresponding to two simulated IQM. The variance of the two sets was chosen to approximately reflect SSIM versus VIF comparison for the JPEG2000 #1 dataset (36 and 25, respectively). With 90 samples in both sets, we ran the HT multiple times and discovered that the Null Hypothesis (the two variances are equal) was rejected about 51% of the time. In the second simulation, we repeated the same procedure with samples drawn from a Uniform distribution with the same variance, and found that the Null hypothesis was rejected about 47% of the time, thereby suggesting that the underlying distribution had little effect due to the large number of sample points used. However, with only ten sample points, the Null hypothesis was rejected 38% of the time in the case of Normal distribution versus about 20% for the Uniform distribution.

This leads us into a discussion of the design of experiments and the probability of detection of an underlying difference in IQM performance. As it is apparent from the above simulations, it is not always possible to detect a difference in the IQM performance using a limited number of observations, and one can only increase the likelihood of doing so by using a large number of samples in the design of the experiment. Thus, anyone designing an experiment to detect the difference between the performance of SSIM and VIF for the JPEG2000 #1 dataset (assuming that the numbers in Table IV represent the standard deviations of the true distributions of residuals) has only about 50% chance of detecting the difference if he chooses to use 100 sample points (distorted images in the dataset). He or she may increase this probability to about 75% by using 200 samples in the study [27]. Obviously, it is not possible to increase this number arbitrarily in one session due to human fatigue considerations. Experts agree that the duration of the entire session, including training, should not exceed 30 min to minimize the effects of fatigue in the measurements. This poses an upper limit on how many images can be evaluated in one session. More sessions may be needed to construct larger datasets.

4) Comparing IQMS Using Data Aggregated From Multiple Sessions: The use of aggregated data from multiple sessions has two advantages. It allows greater resolution in statistical analysis due to a large number of datapoints, and it allows us to determine if an IQM's evaluation for one distortion type is consistent with another. For example, an IQM should give similar score to two images that have the same DMOS but for which the underlying distortion types are different.

Such aggregation of sample points has its limitations as well. Each set of images used in a session establishes its own subjective quality scale which subjects map into quantitative

judgments. This scale depends upon the distribution of quality of images in the set. Since it is virtually impossible to generate two different sets of images with the same distribution of subjective quality (note that subjective quality cannot be known until *after* the experiment, a classic chicken-and-egg problem!), the quality scales corresponding to two data sets will not properly align by default. Without this realignment, experts generally do not agree that data from multiple sessions can be combined for statistical analysis [2], and only the results on the seven individual sessions presented in this paper would be strictly considered justified. In this paper, we attempt a new approach to compensate for the difference between quality scales by conducting a scale-realignment experiment, and it is quite obvious from Fig. 3 that the realignment procedure does modify the quality scales to a significant degree. Nevertheless, the residual lack of alignment between scales or the variability in the realignment coefficients is difficult to quantify statistically.

There could be other reasons for conducting multiple sessions for subjective quality studies apart from the need for greater statistical resolution, such as resource constraints, and the need to augment current studies by adding in new data reflecting possible different distortion types or image content. Calibration and realignment experiments should, in our view, be given due resources in future studies.

5) Final Note on IQM Performance: One important finding of this paper is that none of the IQMs evaluated in this study was statistically at par with the Null model on any of the datasets using a 95% confidence criterion, suggesting that more needs to be done in reducing the gap between machine and human evaluation of image quality.

Upon analyzing Tables VI and IX, one can roughly classify the IQM performance based on the overall results as follows. DCTune performs statistically worse than PSNR,[4] while all other methods perform statistically better than PSNR on a majority of the datasets. JND, SSIM (MS), IFC, and VIF perform much better than the rest of the algorithms, VIF being the best in this class. The performance of PQS, NQM, and BSDM is noticeably superior to PSNR, but generally worse than the best performing algorithms. The performance of VIF is the best in the group, never 'losing' to any algorithm in any testing on any dataset.

Another interesting observation is that PSNR is an excellent measure of quality for white noise distortion, its performance on this dataset being indistinguishable from the best IQMs evaluated in this paper.

IV. Conclusion

In this paper, we presented a performance evaluation study of ten image quality assessment algorithms. The study involved 779 distorted images that were derived from 29 source images using five distortion types. The data set was diverse in terms of image content, as well as distortion types. We have also made the data publicly available to the research community to further scientific study in the field of image quality assessment.

[4]Note that DCTune was designed as part of JPEG optimization algorithm.

References

[1] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, LIVE Image Quality Assessment Database, Release 2 2005 [Online]. Available: http://live.ece.utexas.edu/research/quality

[2] Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment VQEG, Mar. 2000 [Online]. Available: http://www.vqeg.org/

[3] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.

[4] D. R. Fuhrmann, J. A. Baro, and J. R. Cox, Jr., "Experimental evaluation of psychophysical distortion metris for JPEG- encoded images," *J. Electron. Imag.*, vol. 4, pp. 397–406, Oct. 1995.

[5] I. Avcibaş, B. Sankur, and K. Sayood, "Statistical evaluation of image quality measures," *J. Electron. Imag.*, vol. 11, no. 2, pp. 206–23, Apr. 2002.

[6] B. Li, G. W. Meyer, and R. V. Klassen, "A comparison of two image quality models," *Proc. SPIE*, vol. 3299, pp. 98–109, 1998.

[7] S. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: The MIT Press, 1993, pp. 179–206.

[8] J. Lubin, "A visual discrimination mode for image system design and evaluation," in *Visual Models for Target Detection and Recognition*, E. Peli, Ed. Singapore:: World Scientific, 1995, pp. 207–220.

[9] A. Mayache, T. Eude, and H. Cherifi, "A comparison of image quality models and metrics based on human visual sensitivity," in *Proc. IEEE Int. Conf. Image Processing*, 1998, pp. 409–413.

[10] Z. Wang and E. Simoncelli, "Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics," *Proc. SPIE*, vol. 5292, pp. 99–108, 2004.

[11] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, 2000.

[12] *VQEG: The Video Quality Experts Group*, [Online]. Available: http://www.vqeg.org/

[13] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards, and Practice*. Norwell, MA: Kluwer, 2001.

[14] *Methodology for the Subjective Assessment of the Quality for Television Pictures*, ITU-R Rec. BT. 500-11.

[15] Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II VQEG, Aug. 2003 [Online]. Available: http://www.vqeg.org/

[16] A. M. van Dijk, J. B. Martens, and A. B. Watson, "Quality assessment of coded images using numerical category scaling," *Proc. SPIE*, vol. 2451, pp. 90–101, Mar. 1995.

[17] I. Avcibaş, B. Sankur, and K. Sayood [Online]. Available: http://www.busim.ee.boun.edu.tr/image/Image'and'Video.html

[18] D. V. Weken, M. Nachtegael, and E. E. Kerre, "Using similarity measures and homogeneity for the comparison of images," *Image Vis. Comput.*, vol. 22, pp. 695–702, 2004.

[19] JNDmetrix Technology Sarnoff Corp., evaluation Version, 2003 [Online]. Available: http://www.sarnoff.com/products-services/video-vision/jndmetrix/downloads.asp

[20] A. B. Watson, "DCTune: A technique for visual optimization of DCT quantization matrices for individual images," *Soc. Inf. Display Dig. Tech. Papers*, vol. XXIV, pp. 946–949, 1993.

[21] M. Miyahara, K. Kotani, and V. R. Algazi, "Objective Picture Quality Scale (PQS) for image coding," *IEEE Trans. Commun.*, vol. 46, no. 9, pp. 1215–1225, Sep. 1998.

[22] *CIPIC PQS ver. 1.0*, [Online]. Available: http://msp.cipic.ucdavis.edurestes/ftp/cipic/code/pqs, available at

[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 Trans. Image Process.*, vol. 4, no. 4, pp. 636–650, Apr. 2000.

[24] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," presented at the IEEE Asilomar Conf. Signals, Systems, and Computers, Nov. 2003.

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

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

[27] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*. New York: Wiley-Interscience, 1999.

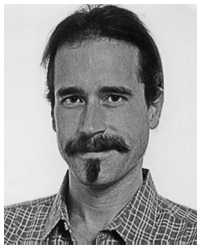
[28] *Electronic Statistics Textbook, WEB*. Tulsa, OK: StatSoft, 2004 [Online]. Available: http://www.statsoft.com/textbook/stathome.html

Hamid Rahim Sheikh (S'93–M'04) received the B.S. degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 1998, and the M.S. and Ph.D. degrees from The University of Texas, Austin, in 2001 and 2004, respectively.

He is currently with Texas Instruments, Inc., Dallas, TX, in the Imaging and Audio Group. His research interests include image acquisition, full-reference and no-reference quality assessment, application of natural scene statistics models and human visual system models for solving image and video processing problems, and image and video codecs and their embedded implementation.

Muhammad Farooq Sabir received the B.S. degree in electronics engineering from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan, in 1999, and the M.S. degree in electrical engineering from The University of Texas, Austin in 2002. He is currently pursuing the Ph.D. degree at The University of Texas, Austin.

His research interests include joint source-channel coding, real-time image and video communication, multiple-input multiple-output systems, and space-time coding. He is a member of the Laboratory for Image and Video Engineering and the Wireless Networking and Communications Group at The University of Texas, Austin.



Alan Conrad Bovik (S'80–M'81–SM'89–F'96) received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from the University of Illinois, Urbana-Champaign, in 1980, 1982, and 1984, respectively.

He is currently the Curry/Cullen Trust Endowed Chair in the Department of Electrical and Computer Engineering, The University of Texas, Austin, where he is the Director of the Laboratory for Image and Video Engineering (LIVE) in the Center for Perceptual Systems. During the Spring of 1992, he held a visiting position in the Division of Applied Sciences, Harvard University, Cambridge, MA. He is the editor/author of the *Handbook of Image and Video Processing* (New York: Academic, 2000). His research interests include digital video, image processing, and computational aspects of visual perception, and he has published over 350 technical articles in these areas and holds two U.S. patents.

Dr. Bovik received the Technical Achievement Award of the IEEE Signal Processing Society in 2005, was named Distinguished Lecturer of the IEEE Signal Processing Society in 2000, received the IEEE Signal Processing Society Meritorious Service Award in 1998, the IEEE Third Millennium Medal in 2000, the University of Texas Engineering Foundation Halliburton Award in 1991, and is a two-time Honorable Mention winner of the International Pattern Recognition Society Award for Outstanding Contribution (1988 and 1993). He was named a Dean's Fellow in the College of Engineering in 2001. He has been involved in numerous professional society activities, including (among many others) Board of Governors, IEEE Signal Processing Society, 1996 to 1998; Editor-in-Chief, IEEE TRANSACTIONS ON IMAGE PROCESSING, 1996 to 2002; Editorial Board, PROCEEDINGS OF THE IEEE, 1998 to 2004; Senior Editorial Board, IEEE JOURNAL ON SPECIAL TOPICS IN SIGNAL PROCESSING, 2005 to present; Series Editor for Image, Video, and Multimedia Processing, Morgan and Claypool Publishing Company, 2003 to present; and Founding General Chairman, First IEEE International Conference on Image Processing, held in Austin, TX, November 1994. He is a registered Professional Engineer in the State of Texas and a frequent consultant to legal, industrial, and academic institutions.