Software to Stress Test Image Quality Estimators

He Liu and Amy R. Reibman
School of Electrical and Computer Engineering
Purdue University

***Abstract*—An image quality estimator (QE) can be used to improve the performance of a system, but only if its scores are easily interpretable. In this paper, we present software, entitled "Stress Testing Image Quality Estimators (STIQE)" that systematically explores the performance of a QE, with the goal of enabling users to interpret the QE's scores. Our software allows consistent and reproducible benchmarks of new QEs as they are developed, so the most effective QE for an application can be chosen. We demonstrate that results produced by the software provide new insights into hidden aspects of existing QEs.**

I. Introduction

For an image quality estimator (QE) to be useful in any application, the scores it produces must be interpretable. A user must be able to take one or two QE scores and draw correct conclusions about the visual quality of the associated image(s). Recently, a three-stage testing framework has been proposed to evaluate how well image and video QEs perform [1]. The first stage requires no subjective testing and instead relies on black-box computational tests [2]. Computational resources are applied to compute QE scores across a variety of scenarios, to identify situations in which the QE is inadequate. The second stage consists of small-scale targeted pairwise subjective tests designed to expose weaknesses in a QE [3], [4]. Finally the third-stage explores so-called specification-based subjective tests (SBSTs) [5], [6], [7], [8], [9], namely subjective tests designed to evaluate whether a QE performs according to its specifications. Statistical analysis [5] of how well a QE performs on pre-existing SBSTs has essentially become mandatory for a new QE. Moreover, there are publicly available software to analyze SBSTs, including IVQUEST [10] and the software at [11].

Because significant time and energy are necessary to create new subjective data, the existing subjective data [5], [6], [7], [8], [9] is reused repeatedly. Unfortunately, this extensive use quickly limits its effectiveness for statistical performance evaluation, due to overuse. Moreover, most existing subjective databases have been created with specific distortions that rely on (often hidden) assumptions. For example. JPEG compression artifacts almost always lie on regular 8x8-block boundaries. Therefore, while SBSTs are critical and we need to continue to expand the richness of the data available, relying on them exclusively is problematic.

Objective methods to assess the efficacy of QEs have been proposed [2]; however with the exception of [12], they are rarely incorporated in recent publications on new image QEs. Therefore, in this paper, we present a software package, written in Python, that enables researchers to easily incorporate an objective testing strategy into their design workflow [1]. Because MATLAB functions can be run by this Python program, the public availability of software will provide consistent evaluation, reports, and comparisons, enabling meaningful benchmarks of QE performance. This software can also assist in comparing among multiple QEs to understand which will provide the best performance for a given application.

[1] A link of the software is provided in: https://engineering.purdue.edu/VADL/software/QoMEX16/STIQE.zip

TABLE I. Summary of FR and NR QEs considered.

QE	Type	Runtime (sec/image) 512	1024	2048
ADM [13]	FR	0.146	.062	2.7
FSIM [14]	FR	0.34	0.54	1.44
MAD [6]	FR	1.52	6.3	29.2
PSNR	FR	0.036	0.108	0.4
PSNR-HVS-M [15]	FR	2.17	8.9	35
SSIM [16]	FR	0.046	0.14	0.66
BIQI [17]	NR	0.68	1.06	1.58
BRISQUE [18]	NR	0.24	0.49	1.5
CORNIA [19]	NR	3.5	4.2	7.8
IL-NIQE [20]	NR	9.8	9.8	9.85
NIQE [21]	NR	0.3	1.1	5.02

The software stress-tests an image QE, with the goal of providing answers to three questions about how the QE scores can be interpreted. These questions are discussed below in Section II. In addition, we demonstrate the power of the software by presenting its results for evaluating 6 full-reference (FR) and 5 no-reference (NR) QEs, which are summarized in Table I along with their relative execution time for three image resolutions.

We begin in Section II defining the questions considered by the software to assist in interpreting a QE's scores. The overall software system is described in Section III, including how distorted images are created. Section IV describes our experiments using the software. Finally, Sections V–VIII describe the procedures of the software to address the questions in Section II, including examples that explore existing weaknesses in recent QEs.

II. Interpreting a QE's scores

For a QE to be useful in a real application, the scores it produces should be interpretable. Our software is designed to evaluate the interpretability, and therefore the usefulness, of a QE for several tasks. In particular, the output of our tests inform a user how well (or how poorly) the QE answers three simple questions:

- Q1: Can the QE score partition a high-quality relatively undistorted image from a badly-distorted image?
- Q2: Does a difference in QE scores between two images indicate a difference in visual quality?
- Q3: Does a greater QE scores for one image than another correctly predict that the first is better visually?

The latter question is explored within two distinct contexts:

Q3a: Analyzing the performance of a single QE in isolation
Q3b: Analyzing the performance of one QE with respect to a set of other QEs.

Additional questions that are addressed by our analysis:

- Does the QE become less effective as the image size increases?
- Is the QE likely to be more effective comparing two images that have different spatial or different angular resolution?

Similar tests have been proposed previously to address these questions [2]; however, no software was provided to allow a consistent and reproducible comparison across research groups. Moreover, the reports generated by this software provide a more comprehensive interpretable summary than the data summarization in [2].

III. Software description

A. Overview

The software takes a folder of undistorted reference images and a list of one or more specific image QEs, and computes a series of tests on the QE using the collection of reference images, to determine how well it performs. Scores can be computed for QEs whose algorithms are already included by the software, or the user can easily add new QEs. Results of the tests are stored in a large Python dictionary and saved to disk. A summary report is stored in an Excel file, and graphs are stored to disk as PNG files. A variety of options are available to control the analysis; however, default operation will generate the tests described in this paper.

The software has three main modules: image impairment, QE computation and statistical analysis. In the image impairment module, the input images are treated as reference images, and sets of distorted images are created by impairing the reference images with different distortion types, according to the specific requirement of each analysis. In the QE computation module, QE scores are computed for each distorted image. Finally, the computed QE scores are analyzed in the statistical analysis module, and results are reported in an excel file.

B. Image distortion method

In the image impairment module, four distortion types are supported: Gaussian Blur, Gaussian Noise, JPEG compression and JPEG2000 compression. The strength of each distortion is controlled by a single parameter, and any level of distortion can be created. Gaussian blur is generated by convolving a Gaussian kernel with the reference images, where the variance of the kernel controls the severity of the distortion. Larger variance creates a blurrier image. Gaussian noise is generated by adding Gaussian white noise to the original image where the distortion severity is controlled by the variance of white noise. JPEG compression is generated by applying the PIL module in Python [22] for which the compression quality parameter controls severity, with 100 indicating the best and 1 the worst image. For JPEG-2000 compression, we use the Kakadu software [23], which controls compression rate through the bit-rate parameter; smaller bit-rate creates more distorted images.

TABLE II. Summary of Testing process, level 0 corresponds to reference image

	Test Goal	Distortion	Distortion levels	Performance Indicator
Q1	T1: quality separability of good from bad	ALL	{0, 50}	overlap fraction of images in overlap region
Q1	T2: comparable at equal angle or equal distance	{blur, noise}	{0,50}	p-value of KS test
Q2	T1: invariance to pixel shift	{JPEG, JP2K}	{30}	$95th\%\Delta QE_{max}$ after cropping
Q3a	T1: monotonicity of each distortion	ALL	{1,2,.. 50}	# monotonic images $80th\%\Delta QE_{max}$ $80th\%\Delta D\_level_{max}$
Q3b	T1: pairwise agreement with other QEs	ALL	{1,2,.. 50}	disagreement rate with other QEs

IV. Experimental conditions

We use 60 images [2] , captured from a variety of cameras at full resolution. One aspect we are particularly interested in, is how do the different QEs behave as the size of the image varies. Therefore, for each full-resolution image, we filter and crop to create versions at 3 spatial resolutions: 512*512, 1024*1024, and 2048*2048 pixels. All four distortions are introduced into these 60*3 reference images according to the default operation of the software. By default, the parameters of each distortion type are chosen to create 50 levels, each with roughly equal increments of distortion. Level 1 represents the lightest and 50 represents the most heavily distorted.

The generated distorted images are used to evaluate 6 Full-reference (FR) and 5 No-reference (NR) QEs summarized in Table I. These QEs have only minor overlap with the QEs considered in [2] and contain many QEs that have been published subsequently.

The collection of testing mechanisms is summarized in Table II. The first column indicates the basic question being considered. The second column, "Test Goal" indicates the attribute being tested, for example, Q_1T_1, indicates the degree to which a QE is able to separate good quality images from bad. The third column indicates which distortion the test can effectively be applied to, while the fourth column is the default distortion levels which are tested. The fifth column indicates performance indicators assess the QE for the particular questions and test. For example, $Q_{3a}T_1blurP_1$ is the number of source images for which the QE has monotonic performance for blur. More detail is provided below.

V. Question 1

A user should be able to rely on scores from a QE to effectively partition a badly degraded image from a nearly undegraded image. This is particularly challenging for a no-reference (NR) QE because it must distinguish desired content from undesired impairment. This question, "*Can the QE score be used to partition a badly degraded image from a relatively undistorted image?*", was considered in [2], where tables of percentiles were created to demonstrate a degree of overlap

[2] These 60 reference images can be downloaded at: https://engineering.purdue.edu/VADL/resources/ref_image_set/set1.zip

between the two sets of images. In this paper, we quantify the overlap of the two distributions to characterize the effectiveness of a QE on this question. While a significant amount of overlap may indicate the QE does not provide adequate performance, we note that a QE may still be useful even if there are overlapping scores between undegraded and heavily degraded images, provided the degree of overlap is known.

To implement this test, the software creates a "badly distorted" (level-50) distorted image for each distortion type and each input reference image. The QE scores for these badly distorted images and the corresponding undegraded images are all computed. The software finds the minimum and maximum value of undegraded images, A_{min} and A_{max}, and badly distorted images, B_{min} and B_{max}. If the QE uses larger value to represent good quality images, we swap the values of A and B so the QE direction does not affect the amount of overlap. Then the overlap is computed by

$$\text{overlap} = \frac{B_{min} - A_{max}}{B_{max} - A_{min}} \tag{1}$$

If this overlap is negative, the scores for the undegraded and badly-degraded images overlap; the more negative this value, the greater the overlap. A larger positive overlap shows a greater distance between high- and low-quality images.

When the input reference images are comprised of versions with different spatial resolutions, it is also possible to use these computed QE scores to learn when a QE is effective (or not) at comparing images of different sizes. Because it is rare for a description of a QE to describe its scope (including viewing distance, and spatial size of images it is effective for), this information can be invaluable to understand how to interpret a QE's score.

In the default operation of the software, the strongest (level-50) blur is chosen such that the width of the filter is a fixed ratio of the overall spatial resolution of the reference image. For example, for a 512-image, the variance is 4.06 while for a 2048-image, the variance is 16.2. Informal subjective tests indicate that the worst-case blurry images have nearly similar visual quality when the images are viewed at the same angular resolution, but the larger image is significantly blurrier than the smaller image when both are viewed with constant pixel-size (i.e., when pixels in each image have the same size). Moreover, in the default operation of the software, the strongest (level-50) noise is chosen to have identical variance, equal to 60, independent of the size of the reference image. These worst-case noisy images appear to have identical visual quality when viewed at constant pixel-size, but a larger image is significantly less noisy than a smaller one when both are viewed at the same viewing angle.

Thus, it is possible to explore whether a QE is more accurate at comparing two images viewed at identical viewing angle, or at constant pixel-size, using the objective scores computed here. We apply the Kolmogorov-Smirnov (KS) statistical test [24] to identify when two sets of QE scores are likely to come from the same distribution. The KS test computes a p-value and a larger p-value between two sizes of blurry images indicates they are more likely to come from the same distribution, therefore, that the QE is more effective at comparing images viewed at the same angular resolution than the same pixel size. Conversely, a larger p-value between two

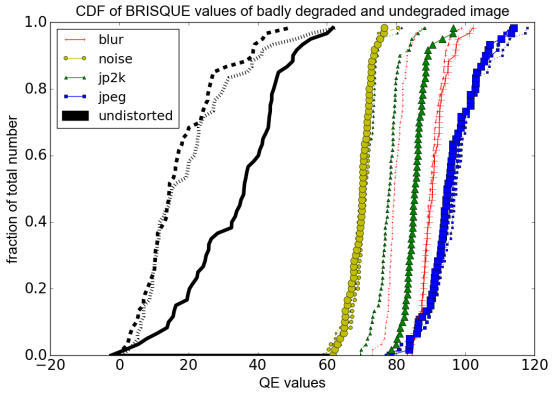

Fig. 1. Partition (overlap $Q_1T_1P_1$) between good images and bad: BRISQUE [18]. Blue squares: JPEG; red +: blur; yellow o: noise; green triangle: JPEG-2000, black: undegraded. The largest marker shows 2048 sized, middle sized is 1024 and smallest is 512.

sizes of noisy images indicates the QE is more effective at comparing images viewed at the same pixel size than at the same angular resolution.

Figure 1 demonstrates this for one exemplary QE, BRISQUE [18]. The cumulative distribution function (CDF) is shown for fifteen sets of images: the undegraded images of each of the three resolutions, and the badly-degraded JPEG, JPEG-2000, noise, and blur distortions for each resolution. The black symbols on the left denote the undegraded images for all sizes, and the solid black line on the x-axis demonstrates the range of QE scores for the undegraded image set.

Several conclusions are immediately apparently. First, the QE scores for undegraded images vary dramatically whether they are smaller (512 or 1024) images, or the larger 2048 images. Despite this, there is very little overlap between the undegraded and the worst-case noisy images, implying the QE may still be effective in this scenario. Second, among the four distortion types, the QE scores for noisy and JPEG-compressed images are relatively independent of image size, while JPEG-2000 and blur have distinctly different scores for small 512-sized images relative to larger 1024- or 2048-sized images. This indicates that BRISQUE is likely to be more effective when it is used to evaluate images with constant pixel sizes rather than images with identical viewing angle.

Table III summarizes the software reports for Question 1 for the 11 QEs considered, applied to the 60 reference images at 3 different spatial resolutions. From the "overlap range percent" column, corresponding to Eq. (1), it can be seen that FR QEs clearly partition the good from the bad images (i.e., have positive overlap ranges), while all NR QEs are unable to make a clear partition.

The KS-blur and KS-noise columns show the p-values from the KS 2-sample test for blur and noise distortions, respectively, when applied to the 512- and 2048-size images. As described above, a larger value for KS-blur indicates the QE is likely to be more effective comparing different sized images viewed with identical viewing angle, while a larger KS-noise value indicates it is more effective to compare images at constant pixel size. Thus we can conclude that among the FR QEs, ADM, FSIM, and SSIM are all more effective at

TABLE III. QE STATISTICS FOR QUESTION 1.

QE name	Overlap range percentage $Q_1T_1P_1$	Percent of images in overlap region $Q_1T_1P_2$	KS-blur $Q_1T_2P_1$	KS-noise $Q_1T_2P_1$
ADM	+9.6	0.0	0.0	0.0
FSIM	+3.5	0.0	0.477	0.0
MAD	+42.6	0.0	0.0	0.0
PSNR	+98.2	0.0	0.345	0.911
PSNR-HVS-M	+100.0	0.0	0.0	0.16
SSIM	+2.9	0.0	0.629	0.0
BIQI	-48.2	64.2	0.0	0.0
BRISQUE	-3.1	1.4	0.0	0.784
CORNIA	-10.0	3.6	0.0	0.0
IL-NIQE	-5.5	4.1	0.629	0.0
NIQE	-5.9	4.0	0.0	0.477

comparing different-sized images with identical viewing angle, while PSNR, PSNR-HVS-M, and MAD are more effective for constant pixel size. Similarly, among the NR QEs, IL-NIQE is unique, in that it will be more effective for identical viewing angle, while all other NR QEs will be more effective for constant pixel size.

VI. QUESTION 2

Another aspect of interpreting a QE is to understand *"When does a difference between QE scores indicate a meaningful visual difference between two images?"*. When subjective scores are also available, this can be measured using the resolving power, defined as the difference between two QE scores for which the corresponding subjective-score distributions have means that are statistically different from each other, typically at the 0.95 significance value [25].

When evaluating the performance of a QE in the absence of subjective scores, another approach is needed. Our software considers the same experiment as was proposed in [2], which assumes that the QE scores for a set of images that have nearly identical visual quality should be nearly identical. In particular, if we have a set of images with nearly identical visual quality, then the absolute difference between their maximum and minimum QE scores indicates how much a QE might vary when evaluating nearly identical images. Any QE variation smaller than this cannot be meaningfully interpreted as a difference in visual quality.

To implement this "invariance" test, each reference image is first impaired by JPEG or JPEG-2000 to produce a mid-quality (level-30) image. Then the reference image R and the two impaired images D_{jpeg} and D_{jp2k} are cropped by a total of 8 pixels in each direction but with 9 different pixel shifts, to create 9 pairs of (R, D_{jpeg}) and 9 pairs of (R, D_{jp2k}). All 9 cropped images in each set are visually very similar. Then the QE scores of these 9 cropped images are computed and the maximum difference among every 9 pairs of scores, $\Delta\mathrm{QE}_{max,i}$, is recorded to characterize the invariance behavior of this QE for reference image i. To align with the concept of 95-%-significance for resolving power, we report the 95-th percentile of these values. If this difference value is near 0, then this QE behaves well on the invariance test, and differences in this QE's scores are likely to be meaningful.

Table IV demonstrates the performance of each QE. Columns "Best" and "Worst" indicate the self-reported QE scores for best- and worst-quality images when available.

TABLE IV. QE STATISTICS FOR QUESTION 2. (OBSERVED BEST AND WORST VALUES ARE IN PARENTHESES WHEN THE PAPER DOES NOT INDICATE BEST/WORST.) ($Q_2T_1P_1$)

QE name	Best	Worst	JPEG 95%tile of $\Delta\mathrm{QE}_{max}$	JPEG-2000 95%tile of $\Delta\mathrm{QE}_{max}$
ADM	1	0	0.03	0.20
FSIM	1	0	0.0	0.01
MAD	0	(184.66)	5.42	4.67
PSNR	∞	0	0.26	0.25
PSNR-HVS-M	∞	0	2.70	0.31
SSIM	1	0	0.01	0.03
BIQI	0	100	23.57	44.40
BRISQUE	0	100	14.20	3.97
CORNIA	(-14.84)	(113.55)	19.78	17.81
IL-NIQE	0	(145.21)	5.77	9.03
NIQE	0	(22.00)	0.61	1.47

When these values are not available, we report our observed best- and worst-quality scores in parentheses. Of the FR QEs, ADM varies by about 20% of its range across images with visually similar quality. NIQE and IL-NIQE fare the best of the NR QEs on this test, each with variations less than 7% of their observed ranges. We note that it is highly likely that negative values for CORNIA are outside its desired range.

VII. QUESTION 3A: COMPARISON USING ONE QE

In this section, considering only a single QE in isolation, we explore the question *"Does a greater QE scores for one image than another correctly predict that the first image is better visually?"*. Without subjective tests, this question is difficult to answer definitively. However, an objective-only analysis can identify concerns that can be easily verified with small-scale subjective tests [3]. The basic approach, also considered by [26], [27], [28], [2], is to explore whether a QE correctly orders a set of distorted images that have all been created by applying different severities of the same distortion to the same reference image.

To implement this "monotonic QE" test, each reference image is impaired using 50 distortion levels for each distortion type, where the level-50 image has the greatest distortion. QE scores are computed for all distorted images. The software then identifies, within a single distortion and reference image, any pairs of images for which the distortion severity increases but the QE scores denote better quality. Once the number of reference images with such non-monotonicities are identified, the software next computes the severities of each non-monotonicity by computing, for each reference image, the maximum difference between QE scores and the maximum difference between distortion levels (as proposed in [2]). A scatter plot is generated indicating these values, with one point for each reference image. To easily interpret the scatter plot, a horizontal line is added to indicate the 80-th percentile of the maximum difference-level per reference image, and a vertical line is added to indicate the 80-th percentile of maximum QE-difference score.

Table V shows the number of reference images (out of 60) that have fully monotonic behavior for each reference image. It can be seen that most FR QEs have monotonic behavior for all distortion types except blur. FSIM has the best monotonic behavior, and ADM the worst among the FR QEs. However, because the NR QEs cannot use information from the reference

TABLE V. STATISTICS FOR QUESTION 3A (MONOTONICITY). NUMBER OF REFERENCE IMAGES FOR WHICH QE DEMONSTRATES FULLY MONOTONIC BEHAVIOR. ($Q_{3a}T_1P_1$)

QE name	Noise		Blur		JPEG		JP2K	
Size	512	2048	512	2048	512	2048	512	2048
ADM	59	60	0	19	20	28	25	54
FSIM	60	60	60	51	60	55	59	60
MAD	60	60	58	46	50	45	40	53
PSNR	60	60	38	0	60	60	48	60
PSNR-HVS-M	60	60	3	0	51	46	39	60
SSIM	60	59	60	44	59	56	60	60
BIQI	19	20	0	0	4	1	0	0
BRISQUE	51	15	14	6	0	1	0	5
CORNIA	0	0	0	0	0	0	0	0
IL-NIQE	38	0	0	0	0	0	0	0
NIQE	36	36	2	16	0	0	0	0

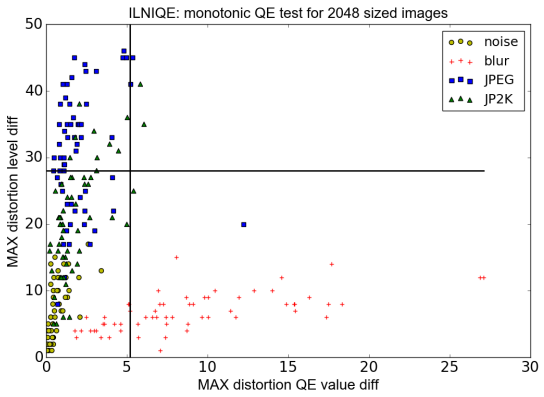

Fig. 2. Non-Monotonicity results for IL-NIQE with 2048-size images. 80% of the samples are below the horizontal line, and 80% of the samples are to the left of the vertical line.

image, they do not have good monotonic behavior. While BRISQUE and IL-NIQE perform reasonably well on 512-size noisy images, their performance drops dramatically for larger images. Finally, we see that CORNIA behaves poorly in this test, since all references exhibit non-monotonic behavior for all distortions.

It is possible to reduce the range of distortions over which this test is applied, to avoid ranges where more distortion may not monotonically influence quality. For example, a small amount of added noise is known to increase perceived quality for lightly blurred images.

From Table V, we see that IL-NIQE behaves non-monotonically for all 2048-size images. The scatter plot in Figure 2 explores this further, as described above. Points for noisy images are near the origin, indicating only minor non-monotonicities. For blur, many images have small distortion-level differences but relatively large differences in QE scores. As described in [2], these images are likely to produce False Differences (FDs) [25], where the QE scores imply a quality difference but the distortion-levels imply little difference. For the JPEG and JPEG-2000 distortions, IL-NIQE has many images with large distortion-level differences but relatively small differences in QE scores. These are likely to produce False Ties (FTs), where the QE scores imply nearly equal quality but the distortion levels imply a potential large visual difference. This analysis clearly pinpoints situations that should be tested subjectively to identify weaknesses in this QE.

TABLE VI. STATISTICS FOR QUESTION 3B, QE DISAGREEMENT PERCENTAGE ($Q_{3b}T_1P_1$)

QE name	512	1024	2048
ADM	13.9	11.0	13.8
FSIM	11.1	9.7	11.2
MAD	11.8	11.0	12.3
PSNR	15.2	15.6	17.5
PSNR-HVS-M	14.8	12.1	13.7
SSIM	11.8	10.3	11.4
BIQI	19.5	16.3	18.4
BRISQUE	15.0	13.6	17.0
CORNIA	16.9	15.1	16.0
IL-NIQE	14.4	16.1	20.7
NIQE	18.6	15.6	17.3

VIII. QUESTION 3B: COMPARISON ACROSS QES

In this section, we consider the same question as in Section VII, *"Does a greater QE scores for one image than another correctly predict that the first image is better visually?"*, by using a comparison across multiple QEs instead of just using one QE. With one QE, it is only meaningful to explore QE scores within the same reference image and the same distortion. However, comparisons across multiple QEs have been effectively used in [4], [29], [30] to explore performance objectively between different reference images and/or different distortions. We note that, as before, this objective analysis cannot replace subjective tests for these scenarios, although it is useful to uncover significant performance concerns.

The same distorted images and their QE scores are used in this section as were used in Section VII; only the method by which they are gathered into sets is different. In this test, the software selects images based on whether they have the same distortion type or same reference image. The selected images form an image pool and every two images in the pool are paired for comparison. After a list of image pairs is generated, the software computes a preference as to which image of each pair a given QE predicts has better quality. Preferences are compared across QEs, and the number of disagreements across all pairs of images and pairs of QEs is computed. Results are reported as the percentage of image pairs for which one QE disagrees with any other QE analyzed. Scatter plots of the disagreements between each pair of QEs are also generated (but not shown here).

Table VI shows the percentage of disagreement between each QE and all of the other 11 QEs, where comparisons are made between different reference images and different distortions. It can be seen that FR QEs have significantly less disagreement with the other QEs than NR QEs do. FSIM is the QE which agrees best with other QEs, and PSNR is the FR QE with the most disagreement, especially for the larger 2048-sized images. Of the NR QEs, IL-NIQE and BRISQUE perform best on 512-size images, having only slightly more disagreement than the FR QEs. However, for larger 2048-size images, IL-NIQE has the most disagreement of all QEs. While it is important to note that agreement does not mean correctness (since all QEs could agree with each other but not characterize human perception), these results provide additional insight into behavior of these QEs across a range of image sizes.

IX. Conclusions

In this paper, we present software STIQE that explores the performance of image QEs using a series of objective analyses, with the goal of improving the interpretability of an individual QE's scores. The software executes a series of tests that explore the performance of a single QE in isolation, as well as compares the performance of one QE relative to multiple other QEs. A series of reports are created, providing insight into each QE's performance, including the answers to the questions: (a) "Can the QE score be used to partition a badly degraded image from a relatively undistorted image?" (b) "Does a greater QE scores for one image than another correctly predict that the first image is better visually?", and (c) "Does a greater QE scores for one image than another correctly predict that the first image is better visually?".

We applied the software to analyze 6 FR QEs and 5 NR QEs. One key observation is that image size matters; performance of these QEs depends on the size of the input image. This is particularly true for those NR QEs which have been trained on existing subjective test images. Another observation is that because the software constructs specific distorted test images for different-sized images, it is possible to infer which QEs are likely to be more effective for comparing different-sized images viewed at identical viewing angle than at constant pixel size.

For the same set of input images, our software creates consistent performance reports that enable consistent performance benchmarks for newly designed QEs. However, a user is not constrained to an existing set of images. New experiments that may lead to new observations can be easily generated for any collection of high-quality images.

The current software supplements existing subjective tests for performance evaluation. We plan to extend the software in the future to simplify the creation of informative small-scale subjective tests from the results of the current analysis.

References

[1] A. R. Reibman, "Strategies for testing image and video quality estimators," *VQEG eLetter*, vol. 1, no. 2, pp. 14–19, December 2014, ftp://vqeg.its.bldrdoc.gov/eLetter/Issues/VQEG_eLetter_vol01_issue2.pdf.

[2] F. M. Ciaramello and A. R. Reibman, "Systematic stress testing of image quality estimators," in *IEEE Int. Conf. Image Proc.*, Sept. 2011.

[3] F. M. Ciaramello and A. R. Reibman, "Supplemental subjective testing to evaluate the performance of image and video quality estimators," in *Human Vision and Electronic Imaging XVI*, Jan. 2011.

[4] A. R. Reibman, "A strategy to jointly test image quality estimators subjectively," in *IEEE Int. Conf. Image Proc.*, Sept. 2012.

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

[6] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. of Electronic Imaging*, vol. 19, no. 1, Mar. 2010, http://vision.okstate.edu/index.php?loc=csiq.

[7] N. Ponomarenko et al., "TID2008 - a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, pp. 30–45, 2009.

[8] N. Ponomarenko et al. "Color image database TID2013: Peculiarities and preliminary results," in *Proceedings of 4th Europian Workshop on Visual Information Processing EUVIP2013*, 2013.

[9] A. Zaric et al., "VCL FER image quality assessment database," *Automatika*, vol. 53, no. 4, pp. 344–354, 2012, http://www.vcl.fer.hr/quality/vclfer.html.

[10] A. V. Murthy and L. J. Karam, "A MATLAB-based framework for image and video quality evaluation," in *International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2010, pp. 242–247.

[11] L. Zhang et al., "FSIM: A Feature SIMilarity Index for Image Quality Assessment (see section on evaluation results)," http://sse.tongji.edu.cn/linzhang/IQA/FSIM/FSIM.htm.

[12] A. Barri et al., "A locally adaptive system for the fusion of objective quality measures," *IEEE Transactions on Image Processing*, vol. 23, no. 6, pp. 2446–2458, 2014.

[13] S. Li et al., "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, 2011.

[14] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Trans. on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[15] N. Ponomarenko et al. "On between-coefficient contrast masking of DCT basis functions," in *Wkshp. on Video Proc. and Quality Metrics*, 2007.

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

[17] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, 2010.

[18] A. Mittal et al., "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[19] P. Ye et al., "Unsupervised Feature Learning Framework for No-reference Image Quality Assessment," in *Intl. Conf. on Computer Vision and Pattern Recognition (CVPR 2012)*, 2012, pp. 1098–1105.

[20] L. Zhang et al., "A feature-enriched completely blind image quality evaluator," *IEEE Trans. on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.

[21] A. Mittal et al., "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[22] "Python Imaging Library (PIL) 1.1.7 for python 2.7 (windows only)," http://www.pythonware.com/products/pil/#pil117.

[23] "Kakadu Software version 7.8 for win32," http://kakadusoftware.com/.

[24] H. W. Lilliefors, "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 399–402, 1967.

[25] International Telecommunication Union, "J.149: Method for specifying accuracy and cross-calibration of video quality metrics (VQM)," Mar. 2004.

[26] M. C. Q. Farias and S. K. Mitra, "A methodology for designing no-reference video quality metrics," in *VPQM*, 2009.

[27] A. Leontaris, P. C. Cosman, and A. R. Reibman, "Quality evaluation of motion-compensated edge artifacts in compressed video," *IEEE Trans. Image Proc.*, vol. 16, no. 4, pp. 943–956, April 2007.

[28] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)," *IEEE Trans. Image Proc.*, vol. 18, no. 4, pp. 717–728, April 2009.

[29] M. Barkowsky el al., "Objective video quality assessment?Towards large scale video database enhanced model development," *IEICE Transactions on Communications*, vol. 98, no. 1, pp. 2–11, 2015.

[30] V. Giotsas et al., "Perceptual video quality estimation by regression with myopic experts," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015, pp. 1–6.