A STRATEGY TO JOINTLY TEST IMAGE QUALITY ESTIMATORS SUBJECTIVELY

Amy R. Reibman
AT&T Labs – Research, Florham Park, NJ, USA

ABSTRACT

We present an automated algorithm to design subjective tests that have a high likelihood of finding misclassification errors in many image quality estimators (QEs). In our algorithm, a collection of existing QEs collaboratively determine the best pairs of images that will test the accuracy of each individual QE. We demonstrate that the resulting subjective test provides valuable information regarding the accuracy of the cooperating QEs. The proposed strategy is particularly useful for comparing efficacy of QEs across multiple distortion types and multiple reference images.

1. INTRODUCTION

Accurate image and video quality estimators (QEs) can be deployed into real systems to assist with a variety of applications. Absolute quality scores rate one image on an absolute scale, and are useful for content acquisition and delivery, and system provisioning. Relative QE scores, which rate the quality of one image relative to another, are useful for algorithm optimization and product benchmarking [1].

Large-scale subjective tests have been relied upon to address the question "Is this image quality estimator (QE) accurate in the required situations?" [2, 3, 4]. However, large-scale subjective tests are expensive and require careful construction to achieve an accurate answer to this question. Furthermore, it is difficult for subjective test designers to anticipate the wide variety of images encountered in a real system.

To be sufficiently robust across a wide variety of images, a QE should be thoroughly tested. Therefore, we are developing a set of testing methodologies for image and video QEs that build upon the well-established strategies of software testing [5]. Instead of attempting to answer the challenging question "Is this QE accurate?", we lower the burden of proof and consider "Is this QE inaccurate?" [6, 7].

The following principles lie at the core of software testing [5] and are directly applicable to testing image and video QEs [6]. The goal of testing should be to find errors, not demonstrate that the system satisfies its specifications. To find errors, it is important to include both positive and negative tests, and to consider conditions that are anticipated and unanticipated. Test cases should be generated automatically and, because exhaustive testing is impossible, it is desirable to maximize the expected number of errors using a fixed number of test cases.

In [6], we apply these principles to systematically create targeted, small-scale subjective tests. The test cases are comprised of image pairs designed to have a high probability of creating potential misclassification errors. Subsequently in [7], we consider a complementary strategy which relies on extensive computational search without expensive subjective tests. In both, we demonstrate approaches that successfully identify a variety of systematic weaknesses in many QEs.

In this paper, we present an automated approach to design subjective tests with a high likelihood of finding one type of misclassification error. In our design algorithm, a number of existing QEs "vote", first to create a list of individual pairs of degraded images, and second to select the best collection of pairs to test. The goal of this collaborative process is to build a pairwise subjective test, whose results will provide valuable information regarding the accuracy of all cooperating QEs.

Section 2 describes the misclassification errors that can occur between an objective QE and the subjective assessment it strives to mimic. Section 3 presents challenges of designing subjective tests to identify systematic weaknesses. Our collaborative approach to design the subjective test appears in Section 4, while Section 5 demonstrates that this testing strategy identifies significant vulnerabilities in a variety of QEs. We conclude with a discussion of future work.

2. ACCURACY OF QUALITY ESTIMATORS

An ideal QE will produce values that are in perfect agreement with subjective test scores. To characterize the deviations of the actual QE scores from this ideal, it is typical to report the root mean-squared error (RMSE), Pearson linear and Spearman rank-order correlation coefficients between the objective and subjective ratings, and the outlier ratio [2]. Two additional measures that assess the ability of a QE to specify relative quality among *pairs* of images are the resolving power and the misclassification error [8].

In this paper, we focus on identifying misclassification errors between a given QE and subjective data [8], [9]. Defined for a *pair* of images, there are three categories of misclassification errors: false ranking or false ordering (FO) (the objective QE rates an image pair opposite to the viewers), false differentiation (FD) (the objective QE rates an image pair as different but the viewers do not), and false tie (FT) (the viewers rate an image pair as having different quality but the objective QE does not).

Table 1 indicates the conditions necessary for the different classification results. Here W_s, E_s, and B_s indicate the statistical decisions that the first image has worse, equal, or better

	W_s	E_s	B_s
W_o	correct decision	false differentiation	false ordering
E_o	false tie	correct decision	false tie
B_o	false ordering	false differentiation	correct decision

Table 1. Classification results based on relative objective QE and subjective results.

subjective visual quality than the second image, respectively. Deciding B_s or W_s is achieved when subjective test participants agree sufficiently that we can reject the null hypothesis that the two images have identical quality. The decision E_s is actually inconclusive; it is the default decision when we cannot reject the null hypothesis.

The decisions W_o, E_o, and B_o indicate that the *objective* QE rates the first image to have worse, equal, or better quality than the second image, respectively. These objective decisions depend on a threshold, Δo, which is the necessary absolute difference between the objective QE scores before they are considered to be unequal.

3. CHALLENGES OF SUBJECTIVE TEST DESIGN

Finding a "bug" in a QE corresponds to finding a *systematic* method to create image pairs that cause misclassification. The traditional subjective testing strategies will generate samples with misclassifications only randomly. The strategies in [10, 11] can only identify potential false ties. Various strategies in [6] create either potential FTs or potential FOs. The goal of this paper is to create as many False Orderings as possible. In particular, given $N = 2M + 1$ QEs, we would like to design a subjective test of K image pairs that maximizes the number of FOs across the test.

3.1. Challenges of identifying FOs

Suppose we want to expose a suspected systematic weakness in a specific QE. For example, suppose the QE cannot accurately assess the relationship between blurry and noisy images. To do so, we need to select a pair of test images which, when tested subjectively, will yield one of the three types of misclassifications.

Figure 1 shows an example of a systematic weakness. A star represents the first image of a potential test pair. The curve indicates a set of potential images associated with a systematic weakness in the tested QE. We will choose the second image of a potential test pair somewhere along this curve. The dotted lines partition the plane into objective and subjective decision regions.

The boundaries of the regions W_o, E_o and B_o are easy to identify using the QE we are testing; it is straightforward to select an image that has a potential FT [10, 11, 6]. However, a FT does not provide much information about the severity of a systematic weakness; it only informs us that one exists.

Without subjective testing, the boundaries of the regions W_s, E_s, and B_s are unknown. Before testing, it is unknown

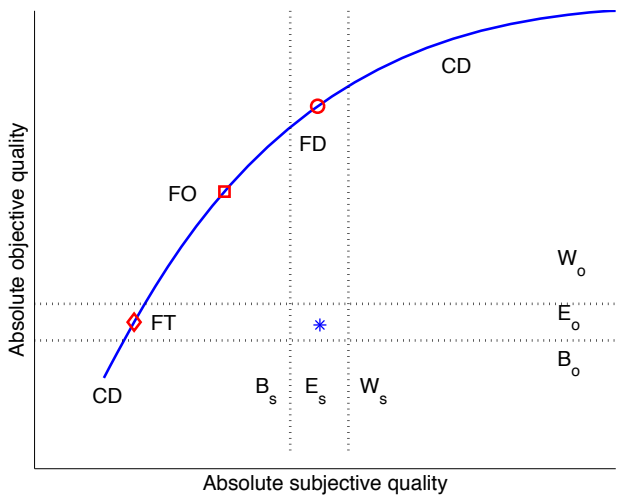

Fig. 1. Misclassification regions and a systematic weakness.

whether a specific image on the curve in the W_o region will produce a FO, a FD, or a Correct Decision (CD). Herein lies the challenge of exposing a suspected systematic weakness: how shall we choose the second image of a test pair so that it has a high probability of being in one of the regions labeled FO or FD, and not in the regions labeled CD?

3.2. Proxies for subjective quality

In [6], we introduce the notion of using one more accurate QE as a *proxy* for subjective quality when designing test pairs for a second less accurate QE. Occasionally the proxy QE identifies its own weakness. Using only one proxy finds no more than one misclassification error in the two QEs.

In the current paper, we extend this concept to explore the ability of a collection of QEs to provide mutual proxy information to each other, with the goal of improving the effectiveness of the test pair selection process. In our algorithm, the QEs collaboratively determine the best test pairs so as to explore the accuracy of each individual QE.

In our subjective test design process, multiple QEs "vote" to decide which image pairs should be tested subjectively. The voting procedure creates a collection of image pairs based on the following goals:

1. The process should be fully automatic.
2. Each test pair should create as many FOs as possible.
3. The identified FOs should be as large as possible.
4. No test pair should produce a FD.

The second goal clearly conflicts with the third. For a pair to achieve the second goal, the proxy QEs should disagree as many times as possible about whether the second image lies in region B_o or W_o relative to the first image. However, requiring many simultaneous FOs limits the size of each FO.

It is also difficult to achieve the third and fourth goals simultaneously. The fourth goal is only achieved if the viewers always decide either B_s or W_s. Such a subjective decision is

more likely if the second image is located close to the boundary between E_o and W_o in Figure 1. However, the third goal is achieved for a given QE when the second image is just outside the unknown boundary of the E_s region.

Finally, the fourth goal is most likely to be achieved when the images in the pair are as visually dissimilar as possible. However, if the two images are very dissimilar in quality, then it is likely that all QEs are accurate enough to agree. Thus, the second goal also conflicts with the fourth.

4. ALGORITHM FOR SUBJECTIVE TEST DESIGN

4.1. Problem definition and notation

Given $N = 2M + 1$ QEs, we would like to design a subjective test of K image pairs that maximizes the number of FOs across the test. K should be large enough to facilitate a statistical comparison of the QEs, and is constrained only by the number of reference images available. All QEs should be monotonic in the desired range.

By design, each image pair is constructed so that at least M QEs decide W_o, at least M QEs decide B_o, and no QEs decide E_o. Ideally, viewers will decide either B_s or W_s for each of the K pairs selected by our algorithm.

Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two (possibly different) distortions, and let x_1 and x_2 be two (possibly different) reference images. We assume here that exactly one of $\mathcal{D}_1 = \mathcal{D}_2$ or $x_1 = x_2$ is true, although this is not necessary. We describe the algorithm to find pairs for which $x_1 = x_2$; the algorithm to find pairs with $\mathcal{D}_1 = \mathcal{D}_2$ is nearly identical.

Our proposed algorithm has two phases. In the first phase, the QEs collaborate to form a collection of potential pairs, two pairs for each reference image and combination of distortions. The first image of each pair is a fixed initial degraded image. The two second images lie at either endpoint of the interval where the QEs maximally disagree. We choose two pairs, because without subjective testing it is impossible to know which side of the interval is most likely to create a decisive visual difference.

In the second phase, the QEs collaborate to select a total number of pairs from the collection of potential pairs created in the first phase. Longer intervals over which the QEs maximally disagree are preferred to shorter intervals, because these will more likely create at least one image pair that has distinct visual quality.

In both phases, the QEs vote to identify maximal disagreement. In Phase One, they vote on the *existence* of disagreement. In Phase Two, they vote on its *severity*.

4.2. Algorithm

Begin by choosing $N = 2M+1$ distinct QEs to collaborate, a collection of reference images, and a collection of distortions.

Phase one: Choose two pairs, (y_1, y_2^a) and (y_1, y_2^b), for each reference image x and each distortion pair $\{\mathcal{D}_1, \mathcal{D}_2\}$ in the collection, using the following algorithm.

1. Pick a severity p_1 for $\mathcal{D}_1$ to be applied to reference image x_1 to create test image $y_1 = \mathcal{D}_1(x_1, p_1)$.
2. Create a dense sampling of distortion severities for $\mathcal{D}_2$ to create a pool of potential test images $\{\hat{y}_2\}$.
3. Apply the N QEs to y_1 and all images in $\{\hat{y}_2\}$.
4. Compute a scaled QE value for each QE computed in Step 3 using a nonlinear fitting function trained on a subjective dataset.
5. For each scaled QE and degraded image y_2 in $\{\hat{y}_2\}$, collect votes of preferred image inside pair (y_1, y_2) .
6. Determine interval of distortion severities in $\{\hat{y}_2\}$ for which at least M QEs prefer y_1 and at least M QEs prefer y_2 and no QEs predicts equal quality.
7. Form two pairs, (y_1, y_2^a) and (y_1, y_2^b), where y_2^a and y_2^b are the images from $\{\hat{y}_2\}$ that lie at the endpoints of the interval identified in Step 6.

The two pairs generated in Step 7 are the input to Phase 2 of the algorithm. There are two pairs for each combination of reference image and distortion.

Phase Two: Choose K pairs from among the pairs generated in Step 7, using the following algorithm.

8. For each pair identified in Step 7, compute the sum across the N QEs of the absolute value of the difference of the scaled QE values for each image in the pair.
9. Pick the $K/2$ largest sums. Add the two image pairs for each sum onto the list of pairs to test subjectively.

We use scaled QEs [9] in Step 4 and Step 8 to facilitate equal treatment of all QEs and to select a common Δo across the QEs for their decision about E_o.

5. SUBJECTIVE TEST

This section presents the design, implementation, and results of a subjective test that implements the proposed algorithm. For this example, we select seven full-reference (FR) QEs, which quantify image quality using a reference image: Structural Similarity index (SSIM) [12], Information content Weighted SSIM (IW-SSIM) [13], PSNR-HVS-M [14], Visual Information Fidelity (VIF) [15], and Visual SNR (VSNR) [16], for which implementations are available from their authors, in addition to Peak Signal-to-Noise Ratio (PSNR) and PSNR_A [17], which computes PSNR on the approximation subband of a Haar wavelet decomposition. To compute the scaled QEs, we train the nonlinear fitting function of [2] to the subjective data in the CSIQ database [4], which results in scaled QEs ranging between approximately 0 and 1. Choosing a strict $\Delta o = 0.01$ for all scaled QEs sets a lower bound on the minimum FO range.

We implement the algorithm of the previous section using the reference images in the CSIQ database [4] and the seven FR QEs above. We choose 70 image pairs that share a reference image but have different distortions among the four we consider: Gaussian blur, JPEG-2000, JPEG, and Gaussian noise. The voting algorithm results in all 70 pairs with blurring, 20 pairs with JPEG-2000, 24 pairs with JPEG, and 26

Same ref. image	FO	FD	CD	FO range	FD range
IW-SSIM	40	13	17	0.01 – 0.23	0.01 – 0.31
PSNR	24	13	33	0.01 – 0.14	0.01 – 0.15
PSNR-A	15	13	42	0.01 – 0.10	0.01 – 0.13
PSNR-HVS-M	15	13	42	0.02 – 0.13	0.02 – 0.18
SSIM	42	13	15	0.01 – 0.39	0.03 – 0.46
VIF	31	13	26	0.01 – 0.21	0.01 – 0.23
VSNR	15	13	42	0.01 – 0.09	0.01 – 0.14
Same distortion					
IW-SSIM	3	3	24	0.02 – 0.13	0.02 – 0.10
PSNR	24	3	3	0.01 – 0.18	0.06 – 0.10
PSNR-A	23	3	4	0.01 – 0.13	0.03 – 0.06
PSNR-HVS-M	16	3	11	0.01 – 0.07	0.01 – 0.05
SSIM	20	3	7	0.01 – 0.17	0.02 – 0.07
VIF	3	3	24	0.01 – 0.11	0.03 – 0.09
VSNR	7	3	20	0.02 – 0.13	0.01 – 0.06

Table 2. Number of actual misclassifications: False Orderings, False Differences, and Correct Decisions. There are no False Ties. Ranges are for scaled QEs.

pairs with noise. We also choose 30 image pairs that have different reference images but share either blurring or JPEG-2000 distortion. The voting algorithm results in 8 pairs with blurring and 22 with JPEG-2000.

We conducted a pair comparison subjective test using all 100 image pairs. Pairs were presented in random order, and thirty viewers naive to the purposes of the experiment selected which image of each pair they preferred. Viewers were instructed to rate the pairs based not on a preference for the image *content* but instead on the technical *quality* of each image. Each viewer took less than 15 minutes to complete the subjective test.

Table 2 summarizes the number of actual misclassifications produced by each of the seven QEs using our subjective test. By design, there are no FT for any of the QEs, and all the correct decisions are correct orderings. Among all 100 pairs, we have 84 pairs for which viewers expressed a statistically significant preference for one image relative to another. Thus 16 pairs will each create a FD in every QE, not a FO.

The subjective test design uncovered FOs in each of the QEs. The most FOs were uncovered for SSIM, the fewest for VSNR. Repeating the process for a different combination of QEs may identify additional FOs in VSNR.

6. CONCLUDING THOUGHTS

We presented a collaborative algorithm where multiple image QEs vote to determine which image pairs should be incorporated in a subjective test. We designed and implemented such a test, which demonstrated that there are still clear inaccuracies in existing QEs, even FR QEs. We suspect that the QEs examined here are not alone. It would also be valuable to apply the methodology to evaluate No-Reference (NR) QEs in their domain of interest. For example, we could test NR blur QEs on blurry images, and NR blocky QEs on JPEG compressed images.

This work is just one part of a framework that provides more rigorous testing of QEs. Our testing indicates the need to improve the current methodology of evaluating QEs. The typical method of testing on generic specification-based subjective databases is insufficient to expose weaknesses in QEs. In future work, we will continue to extend the collection of applicable tests within our developing framework.

7. REFERENCES

[1] S. S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," *Signal Processing: Image Communication*, Aug. 2010.

[2] 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.

[3] 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.

[4] 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.

[5] G. J. Myers, T. Badgett, C. Sandler, and T. M. Thomas, *The art of software testing*, John Wiley and Sons, 1979.

[6] 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.

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

[8] M. H. et al. Brill, "Accuracy and cross-calibration of video-quality metrics: new methods from ATIS/T1A1," *Signal Processing: Image Communication*, vol. 19, pp. 101–107, Feb. 2004.

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

[10] Z. Wang and E. Simoncelli, "Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities," *J. of Vision*, vol. 8, no. 12, pp. 1–13, Sept. 2008.

[11] A. C Brooks, X. Zhao, and T. N. Pappas, "Structural similarity quality metrics in a coding context: exploring the space of realistic distortions," *IEEE Trans. Image Proc.*, vol. 17, no. 8, pp. 1261 – 1273, Aug. 2008.

[12] 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.

[13] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Proc.*, vol. 0, no. 5, pp. 1185–1198, May 2011.

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

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

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

[17] S. Rezazadeh and S. Coulombe, "Method and system for determining a quality measure for an image using multi-level decomposition of images," U. S. Patent application No. 2011/0038548A1, 2011.