SYSTEMATIC STRESS TESTING OF IMAGE QUALITY ESTIMATORS

Frank M. Ciaramello[*]
Cornell University, Ithaca, NY, USA

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

ABSTRACT

We present a methodology to systematically stress objective image quality estimators (QEs). Using computational results instead of expensive subjective tests, we obtain rigorous information of a QE's performance on a constrained but comprehensive set of degraded images. Our process quantifies many of a QE's potential vulnerabilities. Knowledge of these weaknesses can be used to improve a QE during its design process, to assist in selecting which QE to deploy in a real system, and to interpret the results of a chosen QE once deployed.

I. INTRODUCTION

Large-scale subjective tests are considered essential to adequately answer the question "Is this image quality estimator (QE) accurate in the required situations?". However, large-scale subjective tests are expensive and require careful construction to achieve an accurate answer to this question. Due to their high cost and high reward, the existence and increasing availability of subjective data for images [1], [2], [3] leads some researchers to train their QE based on the subjective data. Even when cross-validation is used, the applicability of the resulting QE is limited by the peculiarities of the subjective test set.

Recently, an unbiased process to evaluate the performance of a QE has been presented [4]. This independent validation process uses a secret test set of videos annotated with subjective ratings. QE designers can obtain the performance of their QE for a fee, so that the secrecy of the test set is maintained. Performance results are reported using a common template, allowing easy comparison across multiple QEs. However, using constrained, expensive subjective tests to evaluate QE performance is still limiting.

A QE deployed to evaluate image quality in a real system is exposed to a greater variety of images or videos than those envisioned by the subjective test designers. For example, a QE for broadcast news must be able to handle composited images (with words and graphics added), images acquired from handheld devices, graphics images, to name a few examples. To be sufficiently robust across this wide variety of images, a QE should be thoroughly tested, not only using images that are expected, but also those from unexpected scenarios [5].

In this paper, we present a methodology to find systematic weaknesses in the QE under test. Specifically, instead of answering the challenging question "Is this QE accurate?", we lower the burden of proof and address the question "Is this QE inaccurate?". For this paper, we consider only objective strategies to address this question. In [5] we explored a complementary strategy to consider this question using targeted small-scale subjective tests.

The current approach, relying solely on objective evaluation of a QE, is motivated by a cost-benefit analysis. We use low-cost but extensive computational search to find weaknesses without expensive subjective tests. Thus, we can evaluate a QE using many reference images and many more impairment levels than can be evaluated in a subjective test. Typical subjective tests [1], [2], [3] consider no more than 30 images and no more than 7 levels for a given impairment (for example, JPEG compression).

[*]This work was performed while the second author was a summer intern at AT&T Labs – Research.

Maker	Horiz.	Vert.
Kodak DX3600	1800	1200
Olympus X-2 C50-Z	2560	1920
Panasonic DMC FX-50	3072	2304
Casio EX-Z700	3072	2304

Table I. Summary of cameras

In this work, we consider over 400 reference images and at least 24 impairment levels, a collection for which reliable subjective evaluation is effectively impossible. The breadth of the analysis and the use of both expected and unexpected inputs allows us to identify a variety of potential weaknesses in many QEs.

These weaknesses are not only of academic interest. A QE that has a systematic weakness loses its interpretability, i.e., the QE scores cannot reliably distinguish high quality and low quality images. A system that relies upon such a QE is vulnerable to a motivated attacker who can obtain unfair advantage [6].

Therefore, the presented methodology is useful in three ways. First, during the design process of a QE, it can identify potential systematic weaknesses that can then be eliminated. Second, when evaluating which QE to use for a specific application, an understanding of how each QE performs in a variety of situations allows selection of the most appropriate QE. Finally, once a QE has been selected for deployment, its limitations can be quantified using this methodology. Even if a QE has known vulnerabilities, it may be the best choice due to cost or system constraints.

In Section II, we describe the data we gather to stress-test a number of QEs. Section III describes the five basic strategies we propose to evaluate QE performance. Sections IV and V describe both the methodology and results of our stress testing. Due to limited space, we present only a subset of our observations. We conclude in Section VI with a discussion of future work.

II. LARGE COLLECTION OF DEGRADED IMAGES

We create a large image collection comprised of part of one family's digital photo album. All reference images are recorded directly from a digital camera using "High Quality" JPEG, with no subsequent processing. Four different cameras are used, each with a different pixel resolution; see Table I. In addition, we include in this collection the reference images of three subjective test sets, LIVE [1], TID-2008 [2], and CSIQ [3].

We systematically degrade more than 450 images with a variety of degradations (including Gaussian blur, AWGN, JPEG and JPEG-2000 compression) using over 20 discrete severities of degradation. For example, for AWGN, we use 30 logarithmically spaced values of $\sigma^2 \in [0, 1000]$. Additionally, for JPEG and JPEG-2000 compression, we apply two quality-invariant image transformations: cropping and rotation.

This large collection of reference and degraded images provides a test set from which one can extract valuable information regarding the performance of any QE, *without* the need for subjective testing. For this work, we select both full-reference (FR) and no-reference (NR) QEs, which quantify image quality with and without using a reference image, respectively. Among the many QEs available, in this paper we choose BIQI [7], CPBD [8], JP2k-NR [9], JNBM [10], JQS [11], SSIM [12], VIF [13], for which implementations are

QE	Type	Min.	Max.	Best
BIQI	NR	0	100	0
CPBD	NR-blur	0	1	1
JP2k-NR	NR JPEG-2000	0	80	80
JNBM	NR-blur	0	∞	0
GBIM	NR-block	1	∞	1
JQS	NR-block	0	10	10
PSBIM	NR-block	1	∞	1
SSIM	FR	-1	1	1
VIF	FR	0	1	1

Table II. Summary of FR and NR QEs considered.

available from their authors, in addition to GBIM [14], and PSBIM [15]. There is no uniformity among these QEs regarding their designed maximum and minimum values, nor regarding whether visual quality improves as QE scores increase or decrease. In this paper, we choose to report the raw scores of each QE; therefore Table II indicates QE type and includes the QE score expected for a "best quality" image. NR-blur QEs are designed to measure only the impact of blur on image quality; NR-block QEs are designed to measure only the impact of blockiness. JP2k-NR is designed to measure the impact of JPEG-2000 compression, and BIQI is designed to measure a variety of impairments.

In the following sections, we present a methodology to evaluate the tremendous amount of data generated by applying each QE to the large collection of degraded and nearly undegraded images.

III. STRESS TESTING 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, the following performance measures of QEs are commonly used [16]: the root mean-squared error (RMSE), Pearson linear and Spearman rank-order correlation coefficients between the objective and subjective ratings, and the outlier ratio. 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, defined by Brill et al. [17].

These performance measures rely on subjective data, which is sparse and difficult to obtain accurately. Therefore, in this section, we describe five distinct scenarios in which objective testing can give us valuable information, both about how effectively a QE performs and how to interpret the resulting QE scores.

1) According to the QE under test, undegraded images should all have high quality.
2) According to the QE under test, heavily degraded images should all have poor quality.
3) Identical quality scores should be produced despite a simple transformation of a degraded image, like cropping by a few pixels or rotation by ninety degrees.
4) Monotonically increasing severity of a single degradation on a single reference image should, depending on the degradation and the type of QE, produce either a monotonic or an invariant response in the QE scores.
5) If two QEs disagree about the relative quality of two images, then one of the QEs is incorrect.

The first two correspond to absolute QE scores of a single image, while the remaining three correspond to relative QE scores between one or more pairs of images. The last case is particularly useful across multiple degradation types.

IV. INTERPRETING ABSOLUTE QE SCORES

Absolute QE scores are useful for product benchmarking, content acquisition, and system provisioning [18]. In these applications, the absolute QE score of a single image is typically compared to a fixed threshold to determine if the image has sufficient quality or not. While NR QEs must quantify how much a degradation affects image quality, one of their most challenging tasks is to recognize when an image is undegraded. The QE score for any undegraded image should indicate high quality. Conversely, if an image is heavily blurred or has strong additive noise, the subjective quality is certain to be low. For these images, a QE should produce a score that indicates poor quality.

QE	images	5%	25%	75%	95%
BIQI	473	16.25	25.02	40.90	60.90
CPBD	473	0.42	0.52	0.71	0.81
JP2k-NR	473	54.05	73.98	79.07	80.01
JNBM	473	4.14	10.63	18.63	25.33
GBIM	473	1.06	1.14	1.30	1.44
JQS	473	9.12	9.71	10.23	10.80
PSBIM	473	2.93	5.67	14.12	27.61

Table III. Statistics for NR QE, undegraded images.

QE	images	5%	25%	75%	95%
BIQI	473	60.58	77.62	95.45	111.68
CPBD	473	0.69	0.75	0.85	0.95
JP2k-NR	473	0.00	78.19	79.95	80.01
JNBM	473	3.18	14.73	21.49	34.13
GBIM	473	11.90	17.23	29.32	66.94
JQS	473	-13.80	-7.01	-2.98	-0.96
PSBIM	473	9.44	28.31	136.14	1211.28

Table IV. Statistics for badly degraded JPEG, Quality Factor 1.

To evaluate whether there is a reliable threshold that allows the tested NR QE to distinguish among low and high quality images, we apply the NR QEs to two subsets of the image collection: undegraded images and their corresponding heavily degraded images. For each image subset, we compute the distribution of the QE scores across the set of images and report the 5-th, 25-th, 75-th and 95-th percentile of this distribution. Table III shows the results for the undegraded images, while Tables IV, V, and VI show results for heavily degraded JPEG, Gaussian blur, and JPEG-2000, respectively.

As can be seen from Table III, scores for BIQI and CPBD on *undegraded* images span nearly the entire range of values for these QEs. Further, while JQS typically reports scores near the desired value of 10, close to 30% of its scores exceed its reported maximum.

The ranges of BIQI, CPBD, JNBM, and PSBIM in Tables III and IV show significant overlap; no one threshold can correctly partition the undegraded and badly degraded JPEG images. Of interest are the negative scores for JQS and the BIQI scores above 100 in Table IV. Scores for JNBM are actually *lower* for heavily blurred images than for the undegraded images. PSBIM, a blocking QE, responds nearly as strongly to blur as to JPEG.

While it is not apparent from these tables, BIQI, JP2k-NR, JNBM, and JQS all behave quite differently for the larger images than for the small images typically used in subjective test data. The broader set of undegraded images may not share the specific statistical characteristics of the images in these QEs' training set.

V. INTERPRETING RELATIVE QE SCORES

Relative QE scores are useful for algorithm optimization and product benchmarking [18]. As described in Section III, there are three cases where exploring relative QE scores for a *pair* of images can provide useful information about the accuracy of a QE: when a QE should give nearly similar scores; when a QE should respond monotonically as a degradation increases in severity [19], [10], [20]

QE	images	5%	25%	75%	95%
BIQI	440	57.91	62.58	70.10	80.70
CPBD	440	0.00	0.00	0.01	0.05
JP2k-NR	440	32.75	33.45	36.07	49.55
JNBM	440	1.43	5.50	8.85	12.62
GBIM	440	0.93	0.96	1.00	1.02
JQS	440	8.45	8.97	10.05	11.62
PSBIM	440	13.27	21.57	45.90	90.79

Table V. Statistics for heavy Gaussian blur, $\sigma^2 = 10$.

QE	images	5%	25%	75%	95%
BIQI	307	40.11	47.36	56.45	66.30
CPBD	307	0.06	0.13	0.26	0.40
JP2k-NR	307	32.65	33.05	36.71	48.66
JNBM	307	1.91	3.44	13.10	17.69
GBIM	307	0.93	0.97	1.04	1.11
JQS	307	8.75	9.40	10.30	11.23
PSBIM	307	5.34	9.10	20.26	40.22

Table VI. Statistics for JPEG-2000, compression factor 200.

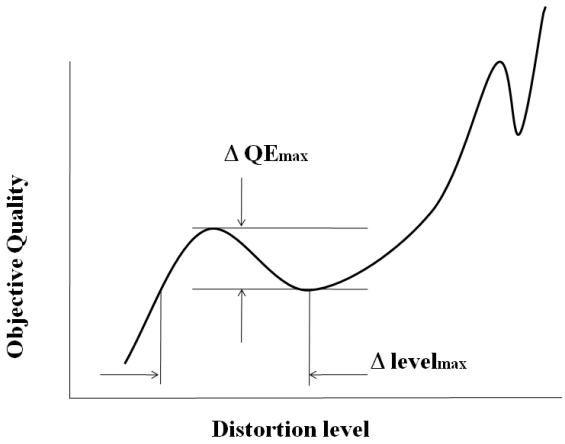

Fig. 1. An example non-monotonicity.

for the same reference image; and when two (or more) QEs disagree about which image of a pair of images is better.

To describe performance across pairs of images, we adapt the misclassification errors defined in [17] between a given QE and subjective data. Their misclassification errors include false rank or false ordering (FO) (the objective QE rates an image pair opposite to the humans), false differentiation (FD) (the objective QE rates an image pair as different but the humans do not), and false tie (FT) (the humans rate an image pair as having different quality but the objective QE does not). In our case subjective data does not exist; therefore, we adapt these by replacing the human ratings with information that serves as a *proxy*. One useful proxy is the knowledge that a QE should create equal scores for an image pair that has undergone a quality-invariant transform (such as cropping). Another is that as a single degradation (i.e., JPEG) increases in severity, the QE should respond monotonically [19], [10], [20].

First, we explore whether a QE produces similar scores when an image is cropped by a few pixels. To evaluate this property, for each reference image we choose a mid-level degradation and compute the maximum variation in QE scores when this degraded image is cropped by between 0-9 pixels. The distribution of this variation, across each reference image, is reported in Table VII. We see that significant FDs occurs for many QEs. JP2k-NR, BIQI, JQS, GBIM, and PSBIM all perform poorly on cropped images. The first two QEs rely on maximally decimated wavelets while the latter four assume known locations for block boundaries. CPBD is quite robust to cropping and rotation (not shown).

The vulnerabilities identified in Table VII are important given the threat models described in [6]. If a system uses one of the QEs with substantial fluctuation given crop, an attacker needs only choose the cropped image with the best (or worst) score; this will trick the system into behaving as though the image has better (or worse) quality than it actually has.

QE	images	min	5%	median	95%	max
BIQI	84	3.80	6.36	14.01	22.60	26.50
GBIM	84	0.73	0.92	1.47	2.76	4.29
JQS	84	3.02	3.26	4.59	7.06	9.32
PSBIM	84	0.84	1.04	2.74	11.12	39.83
QE	images	min	5%	median	95%	max
BIQI	84	3.90	5.25	15.44	36.75	49.39
CPBD	84	0.00	0.00	0.02	0.05	0.07
JP2k-NR	84	3.09	20.66	38.58	43.89	44.73
JNBM	84	0.06	0.15	0.48	1.10	2.24

Table VII. Invariance to cropping by between 0-9 pixels. JPEG Quality Factor 15; JPEG-2000 compression factor 140.

Second, we explore the requirement of monotonicity. Within a single degradation, FR QEs should respond monotonically as the severity of the degradation increases. NR-blocking QE should respond monotonically to JPEG; NR-blurring QE should respond monotonically to increasing blur and JPEG-2000 compression. The JP2k-NR QE should respond monotonically to JPEG-2000.

Figure 1 illustrates the performance measures we propose to evaluate non-monotonicities. It shows an example where a QE does not behave monotonically as one reference image is degraded with increasing severity. In general, any non-monotonicity will result in at least one pair of images that has a misclassification error. The type of misclassification depends on the relative position of the local minima and maxima in the non-monotonicity. In this example, the sharp decrease in the QE score on the right is likely to cause a FD, due to the small change in distortion level between the local minimum and maximum QE scores. The more gradual variation on the left is likely to lead to a pair with visually different quality and therefore a FO, due to the higher QE score associated with a much smaller distortion level. Without subjective testing, it is impossible to know which situations exists. Therefore, we denote these as *potential* misclassifications. $\Delta QE_{\max}$ quantifies the severity of the potential false ordering (PFO) or potential false difference (PFD), while $\Delta level_{\max}$ quantifies the severity of a Potential False Tie (PFT).

For each reference image, if there is a non-monotonicity for a given degradation, we search for potential misclassifications. Table VIII shows the 80% percentile of the distribution of $\Delta QE_{\max}$ and $\Delta level_{\max}$ across reference images. This table shows that many QEs exhibit some non-monotonicities. However, with the exception of BIQI, which has the hardest task since it is designed for multiple individual degradations, most of the PFOs are limited in either $\Delta QE_{\max}$ or $\Delta level_{\max}$.

Third, when comparing across multiple distortion types without subjective data, it is difficult to quantify correct QE performance. For a given reference image, what level of blur produces equivalent quality to a given level of JPEG? However, by examining the relationship between the scores of *two* QEs on such a pair of images, we are able to discover useful information about how the two QEs perform. In particular, we search across multiple QEs to identify cases of conflicting orderings (CO), in which two QEs disagree about how to rate an image pair. In the event of such a disagreement, one of the two QEs is necessarily inconsistent with human ratings. As was discussed above for Figure 1, one systematic weakness can result in a variety of misclassification

QE	images	images with PFO	80%-ile Δ QE	80%-ile Δ Level
BIQI	473	458	8.30	81.00
GBIM	473	167	0.04	15.00
JQS	473	153	0.14	9.70
PSBIM	473	458	1.33	28.00
BIQI	307	300	8.34	89.50
CPBD	307	281	0.01	35.00
JP2k-NR	307	273	0.58	48.00
JNBM	307	306	0.39	62.00
BIQI	440	417	10.70	0.67
CPBD	440	214	0.00	0.67
JP2k-NR	440	429	1.30	1.45
JNBM	440	122	0.08	0.46
BIQI	241	222	9.26	378.98
CPBD	241	113	0.03	744.38
JP2k-NR	241	241	0.73	167.03
JNBM	241	166	0.15	162.48

Table VIII. Potential misclassifications for monotonic degradations (JPEG, JPEG-2000, Gaussian blur, AWGN).

errors. Therefore, we describe the severity of the weakness using the *maximum PFD* for each QE.

For each reference image, given two degradation types, we search for all possible conflicting orderings. Next, we search for the image pair that one QE rates as having "equal quality" and the other QE rates as having maximally different quality. This pair has the maximum possible PFD for that QE.

In [5], we showed that SSIM had a systematic weakness when comparing noisy and blurry images. Using the same 10 reference images as [5], our systematic strategy indicates the $75th$ percentile of PFD for SSIM is 0.29; for VIF it is 0.30. Both are dramatically higher than the FD previously identified. Using 146 images, the severity of the PFD increases, with the 75%-tiles of 0.60 for SSIM and 0.53 for VIF. Subjective tests are necessary to determine if the PFD are actual FD; based on results in [5], it is likely that VIF correctly orders these images and SSIM does not. Comparing degradations of AWGN and JPEG, the corresponding numbers are 0.58 and 0.53 for SSIM and VIF across 146 images, while for degradations of blur and JPEG, they are 0.04 and 0.19 across 440 images. Thus, SSIM and VIF have much greater agreement between degradations of blur and JPEG than among the other pairs of degradations.

VI. CONCLUDING THOUGHTS

In this paper, we rely on extensive computational resources to identify inconsistencies and to search for potential vulnerabilities in existing QEs, without requiring any subjective experiments. We demonstrated a set of systematic stress tests using a large collection of undegraded images, composed of a family's digital photo album and commonly used, publicly available test images. We systematically applied a variety of degradations to the image collection using over 20 levels of severity per degradation. Inconsistencies in the performance of individual QEs, including BIQI and CPBD, were identified due to an overlap in the range of QE scores associated with collections undegraded and badly degraded images. Evidence for a vulnerability in a QE was provided when a quality-invariant transform produced images with disparate QE scores. Furthermore, applying multiple QEs to the image collection affords comparisons between QEs, leveraging more accurate QEs to identify weaknesses in less accurate QEs.

We suspect that the QEs examined in this paper are not alone in exhibiting such systematic weaknesses. This work is just one part of a framework for more rigorous testing of QEs and these tests should become part of the core analysis presented for every image QE. The proposed computational tests are complementary to the typical subjective tests; they do not replace the need for testing according to specifications. Furthermore, the tests discussed in this work explored only some approaches to systematic stress testing and are certainly not comprehensive. For example, this work considered images with only a single degradation type.

In future work, we will leverage this large-scale computational testing to facilitate targeted, small-scale subjective tests as described in [5]. We will continue to extend the collection of applicable tests within the proposed framework of systematic stress testing.

VII. REFERENCES

[1] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE image quality assessment database release 2," 2005, http://live.ece.utexas.edu/research/quality/subjective.htm.

[2] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008 - a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, pp. 30–45, 2009.

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

[4] R. S. Streijl, S. Winkler, and D. S. Hands, "Perceptual quality measurement: Towards a more efficient process for validating objective models," *IEEE Signal Processing Magazine*, pp. 136–140, July 2010.

[5] 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*, January 2011.

[6] W. Cheswick, D. Kormann, and A. R. Reibman, "Vulnerability assessment of image and video quality estimators," in *Wkshp. on Video Proc. and Quality Metrics*, January 2010.

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

[8] N. Narvekar and L. J. Karam, "An improved no-reference sharpness metric based on the probability of blur detection," in *Wkshp. on Video Proc. and Quality Metrics*, January 2010.

[9] H. R. Sheikh, A. C. Bovik, and L. Cormak, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. Image Proc.*, vol. 14, no. 11, pp. 1918–1927, Nov. 2005.

[10] 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.

[11] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *IEEE ICIP02*, 2002, http://www.cns.nyu.edu/ zwang/files/research/nr_jpeg_quality/index.html.

[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] 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.

[14] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Sig. Proc. Letters*, vol. 4, no. 11, pp. 317–320, November 1997.

[15] S. Suthaharan, "A perceptually significant block-edge impairment metric for digital video coding," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP*, 2003, pp. III–681–684.

[16] 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.

[17] M. H. Brill, J. Lubin, P. Costa, S. Wolf, and J. Pearson, "Accuracy and cross-calibration of video-quality metrics: new methods from ATIS/T1A1," *Signal Processing: Image Communication*, vol. 19, pp. 101–107, Feb. 2004.

[18] 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.

[19] 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.

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