Improving Robustness of Image Quality Measurement with Degradation Classification and Machine Learning

Tiago H. Falk, Yingchun Guo, and Wai-Yip Chan
Department of Electrical and Computer Engineering
Queen's University, Kingston, Ontario, Canada
Email: falkt@ee.queensu.ca, {guoy, chan}@post.queensu.ca

***Abstract*— Image quality metrics can be classified as generic or degradation specific. Degradation specific measures perform poorly under "mismatched" conditions. Generic measures, on the other hand, may compromise quality measurement accuracy while gaining robustness to variation in distortion conditions. To improve the accuracy-robustness tradeoff, we employ support-vector degradation classification and machine learning tools to judiciously combine generic and degradation specific measures. To test our algorithm, composite quality metrics are optimized for five different distortion classes. Experiment results show that the proposed algorithm achieves improved performance and robustness relative to two benchmark generic quality metrics.**

I. Introduction

The most reliable way to measure the quality of images is through the use of subjective quality assessment tests such as the commonly used mean opinion score (MOS) test. These tests, however, are expensive and time consuming, making them unsuitable for automatic quality measurement. Objective (machine-based) measurement methods have been the focus of more recent research. Machine-based measurement allows computer programs to automate image quality measurement in real time, thus playing a crucial role in modern image processing applications such as compression, steganalysis, and communication. Traditionally, error-based quality measures such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE) have been used. Such measures, however, have been shown to correlate poorly with subjective quality scores [1]. Current efforts have focused on devising features that incorporate characteristics of the human visual system.

When devising perceptually-relevant metrics for image quality measurement, one is faced with the option of devising a measure with a specific application in mind (e.g., [2]) or devising a generic measure which is applicable to various different degradation conditions (e.g., [3]). Degradation specific measures tend to perform poorly under conditions for which it was not optimized. Generic measures, on the other hand, may compromise quality measurement accuracy while gaining robustness to variation in distortion conditions. In this paper, we propose to improve the accuracy-robustness tradeoff by judiciously combining generic and degradation specific measures to form a composite quality measure.

Our experiments are carried out with the LIVE image database [4] where reference images are corrupted by five different distortion classes – JPEG and JPEG2000 compression, additive (white) noise, Gaussian blurring, and bit errors as simulated in a wireless fast fading environment. We employ machine learning methods to sift out the most salient features from a pool of 27 candidate features. Feature selection is performed for each of the five distortion classes described above. The selected features are linearly combined to form a composite measure that is optimal for a given distortion class. In order to devise robust (and generic) estimators of image quality, we propose a novel degradation-classification assisted image quality measurement paradigm. A degradation-type classifier is designed to detect which of the five degradation classes the test image belongs to. Once a degradation class is specified, one of the five composite measures is used. We show that improved robustness and improved quality estimation performance are attained with the proposed scheme.

The remainder of this paper is organized as follows. In Section II, a brief overview of subjective and objective image quality measurement methods is given. Section III describes the candidate features used in our experiments as well as the machine learning tools used for feature selection and degradation-type classification. Lastly, experimental results are reported in Section IV and conclusions in Section V.

II. Image Quality Measurement

In this section, common subjective and objective image quality measurement methods are briefly described.

A. Subjective Assessment Tests

Subjective image quality assessment tests can be classified as either double- or single-stimulus [5]. Double-stimulus (DS) tests can be further categorized as DS impairment scale (DSIS) tests or DS continuous quality scale (DSCQS) tests. In both cases, subjects are presented with two images: the unimpaired original (reference) and the impaired processed image. With DSIS tests, subjects are asked to rate the quality of the impaired image, relative to the quality of the original image using the impairment scale described in Table I. Subjects in DSCQS tests, on the other hand, are not told which of the two images is the original image. Instead, subjects are asked to rate the quality of *both* images using a continuous scale which is divided into five equal lengths. The five-point scale

TABLE I

SUBJECTIVE IMAGE QUALITY RATING SCALES.

Rating	MOS	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible, but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

corresponds to the so-called mean opinion score (MOS) scale described in Table I. Differences between subjective MOS for reference and impaired images can be computed and such measures are termed difference MOS (DMOS).

Subjects in single-stimulus (SS) tests, on the other hand, as the name suggests, are presented with only the impaired images. SS tests can be further classified as categorial or non-categorial. Categorial judgement tests often make use of the five-point MOS or impairment scales described in Table I. Analogous to the DSCQS method, non-categorial tests make use of continuous scales. As opposed to DS tests where subjects rate the quality relative to the original image, subjects in SS tests rate the image quality based on personal opinions of what high-quality images should look like.

As described in [6], the LIVE image database used here has been subjectively scored using the non-categorial SS methodology. Subjects provide their perception of quality on a continuous linear scale which was divided into five equal regions marked with the MOS-scale adjectives described in Table I. Original images were also evaluated in the same experimental session as the impaired images. Thus, despite being an SS test, quality difference scores (i.e., DMOS) were also computed. Moreover, as suggested in [7], in order to minimize the variation between individual quality scores, z-score transformation is used to normalize the scores. Variation in individual scores is caused by the fact that not all subjects use the full range of the continuous scale. After normalization, the quality scores are re-scaled to fill the range from 1 to 100 and the mean opinion score is obtained for each image.

As can be seen, subjective testing is reliable but is also very expensive and time consuming, making it unsuitable for frequent or rapid applications. As a consequence, current research focuses on devising objective measures that correlate well with subjective tests. Objective methods can be implemented by computer programs and can be used for real time quality measurement and control. Common objective image quality measurement methods are described next.

B. *Objective Measurement Methods*

Objective quality measurement methods can be classified into three broad classes: full-, no-, and reduced-reference. Full-reference (FR) measurement, sometimes called double-ended measurement, depends on some form of distance or similarity metric between the original and the impaired images. Traditional FR measures include peak signal-to-noise ratio (PSNR) and mean squared error (MSE). Such measures, however, have been shown to correlate poorly with subjective quality scores [1]. A recent study, however, using the LIVE database has shown that PSNR is a useful measure for additive white noise distortions [8]. Other more complex FR measures that take into account models of the human visual system (HVS) have been shown to attain more accurate quality predictions; some such measures will be presented in Section III-A.

No-reference (NR) measurement, on the other hand, is a more challenging problem as it only depends on the impaired image. Such methods are often called single-ended and distortions are detected and quantified using only the degraded image. Currently, NR image quality measurement is only feasible if prior knowledge of the image distortion process is available (e.g., JPEG compression [9]). To facilitate NR measurement, partial information can be extracted from the original signal and sent as side information to the NR measurement algorithm; such method is termed reduced-reference (RR). Examples of partial information obtained from the original signal include parameters describing natural image statistics [10] or parameters describing image blockiness [11].

Objective measurement methods can be further classified as degradation specific or generic. Degradation specific measures have prior knowledge of the degradation process and are able to quantify such distortions. As examples, blocking artifacts are often quantified for JPEG image quality estimation; ringing and blurring artifacts are quantified for JPEG2000 quality estimation, and, as mentioned previously, PSNR can be used for quality estimation of additive noise-corrupt images. Generic measures, on the other hand, assume no prior knowledge of the degradation process and, at the cost of possibly lower estimation performance, are applicable to different degradation types. In this study, we propose to use machine learning tools to judiciously combine generic and degradation specific measures to form improved-performance composite quality measures that are more robust to different degradation conditions.

III. MACHINE LEARNING FOR FEATURE SELECTION AND DEGRADATION-TYPE CLASSIFICATION

In this section, a brief overview of the features used in this paper is given. Machine learning methods used for feature mining and degradation-type classification, as well as the proposed composite measures, are also introduced.

A. *Feature Set*

In this paper, a pool of 27 full-reference image quality metrics (features) are investigated. The features fall into six different categories based on information extracted from:

1) pixel-differences (PD),
2) image correlation (IC),
3) edge stability (ES),
4) spectral distance (SD),
5) models of the human visual system (HVS), and
6) structural similarity (SS).

Due to space constraints, we do not describe each feature in detail; in lieu, Table II lists references that describe each feature. The table also provides a brief description as well as symbols for each feature. It is important to emphasize that the performance of each individual feature has been studied previously either in [8], [12], or [13]. Here, more

TABLE II

List, brief description, and references for the 27 features.

No.	Symbol	Description	Reference
1	PD.1	Mean square error (MSE)	[12]
2	PD.2	Peak signal-to-noise ratio	[12]
3	PD.3	Modified infinity norm	[12]
4	PD.4	L*a*p perceptual distortion	[12], [14]
5	PD.5	L1-norm distortion	[12], [15]
6	PD.6	Maximum difference	[12], [15]
7	PD.7	Image fidelity	[15]
8	IC.1	Czenakowski correlation	[12]
9	IC.2	Structural content	[12], [15]
10	IC.3	Normalized cross-correlation	[12], [15]
11	IC.4	Mean angle similarity	[12]
12	IC.5	Mean magnitude-angle similarity	[12]
13	ES.1	Edge stability MSE	[13]
14	ES.2	Pratt measure	[12]
15	ES.3	Mean square gradient error	[13]
16	SD.1	Spectral magnitude MSE	[12]
17	SD.2	Spectral phase MSE	[12]
18	SD.3	Spectral magnitude-phase MSE	[12]
19	SD.4	Block spectral magnitude error	[12]
20	SD.5	Block spectral phase error	[12]
21	SD.6	Block spectral magnitude-phase error	[12]
22	HVS.1	HVS-modified normalized absolute error	[12]
23	HVS.2	HVS-modified normalized MSE	[12]
24	HVS.3	HVS-modified MSE	[12]
25	HVS.4	Similarity weight	[12], [14]
26	SS.1	Structural similarity index	[6]
27	SS.2	Universal quality index	[3]

reliable and robust image quality estimators are proposed by combining multiple features to form a composite metric. To this end, automatic feature selection is used to select the top-seven features from the candidate feature pool. As a proof-of-concept experiment, we design composite measures for each of the five distortion types available with the LIVE image database. Online, degradation-type classification is used to select the optimal composite measure to be used. Feature selection, composite measures, and degradation classification are described in the remainder of this section.

B. Automatic Feature Selection

We investigate the effectiveness of two statistical data mining algorithms, namely, classification and regression trees (CART) [16] and multivariate adaptive regression splines (MARS) [17], and one classical feature selection algorithm – sequential forward selection (SFS). For CART and MARS, feature selection and estimator model optimization are performed jointly, thus the selected features maximize the correlation between estimated DMOS and subjective DMOS. The SFS algorithm, on the other hand, starts with the variable that is most correlated with subjective DMOS, and at each step adds a new variable that, together with the previous ones, most accurately predicts subjective DMOS via linear regression. A partial F-test is incorporated in the algorithm such that the variables chosen have small variance. The top-seven selected features for JPEG and JPEG2000 compression, for each feature selection algorithm, are given in Table III. In addition, Table IV exhibits the top-selected features for additive white noise, Gaussian blur, and bit errors.

TABLE III

Top features selected by each algorithm for JPEG and JPEG2000 compression.

Top feature	JPEG			JPEG2000		
	CART	MARS	SFS	CART	MARS	SFS
PD.3		✓	✓	✓	✓	✓
PD.6				✓		✓
PD.7			✓			✓
IC.1		✓			✓	
IC.5		✓				
ES.1			✓		✓	✓
ES.2	✓				✓	
ES.3				✓		
SD.2	✓					
SD.3					✓	
SD.5		✓				
SD.6			✓			✓
HVS.1	✓			✓		
HVS.2	✓			✓		✓
HVS.3	✓	✓	✓	✓	✓	
HVS.4	✓	✓	✓			
SS.1	✓	✓	✓		✓	✓
SS.2				✓		

As can be seen from Table III, for JPEG compression, all three algorithms select HVS-based features (HVS.3 and HVS.4) and structural similarity features (SS.1). The statistical analysis reported in [12] suggests that HVS.3 is in fact a "fundamental metric" for JPEG compression; other reported fundamental metrics include edge stability (ES.1 as selected by SFS) and spectral phase MSE (SD.3 as selected by CART). For JPEG2000 compression, all three algorithms selected the modified infinity norm (PD.3) as one of the top-selected features. Moreover, of the three selection algorithms, two selected maximum difference (PD.6) and edge stability (ES.1) measures. All three algorithms selected a feature from the structural similarity class; while MARS and SFS selected the structural similarity index (SS.1), CART selected the universal quality index (SS.2). Edge stability and HVS-modified MSE (HVS.3 as selected by CART and MARS) were shown in [12] to be fundamental metrics for the SPIHT image compression algorithm, a similar wavelet-based coding paradigm to JPEG2000.

From Table IV, it can be seen that for additive white noise, all three algorithms select PSNR (PD.2), L1-norm distortion (PD.5), and the Pratt measure (ES.2). Interestingly, the results reported in [12] suggest that MSE is the only fundamental metric for additive noise distortions. As can be seen from the tables, MSE (PD.1) is not selected by any of the three algorithms; in fact, feature PD.1 is not selected for any of the five distortion classes. For Gaussian blur distortions, the only feature selected by all three algorithms is the universal quality index (SS.2). Two of the three algorithms selected angle similarity (IC.4), gradient error (ES.3), and spectral magnitude-phase error (SD.3) as top features; the latter is also shown in [12] to be a fundamental metric for blur effects. In addition, edge stability (ES.1 as selected by CART) and HVS-modified MSE (HVS.3 as selected by MARS) were also

TABLE IV

TOP FEATURES SELECTED FOR EACH ALGORITHM FOR ADDITIVE WHITE NOISE, GAUSSIAN BLUR AND BIT ERROR DISTORTIONS.

Top feature	White noise			Gaussian blur			Bit errors		
	CART	MARS	SFS	CART	MARS	SFS	CART	MARS	SFS
PD.2	✓	✓	✓			✓	✓		✓
PD.3						✓			
PD.4		✓			✓				
PD.5	✓	✓	✓						
PD.6								✓	✓
PD.7						✓			
IC.2					✓				
IC.3					✓		✓	✓	
IC.4					✓	✓			
IC.5			✓						
ES.1	✓		✓	✓			✓		
ES.2	✓	✓	✓			✓		✓	
ES.3	✓			✓		✓	✓		
SD.2								✓	
SD.3	✓			✓	✓			✓	✓
SD.4	✓								
SD.5				✓			✓		
SD.6				✓			✓		
HVS.1								✓	✓
HVS.3					✓				
HVS.4			✓	✓					
SS.1		✓					✓	✓	
SS.2			✓	✓	✓	✓			✓

shown to be fundamental metrics for blurriness effects. Lastly, for distortions caused by bit errors, two of the three algorithms selected PSNR (PD.2), maximum difference (PD.6), normalized cross-correlation (IC.3), spectral magnitude-phase MSE (SD.3), HVS-normalized absolute error (HVS.1), and structural similarity index (SS.1) as top features.

C. Composite Measures

As mentioned previously, CART and MARS perform joint feature selection and model optimization. For SFS-selected features, composite measures are designed based on a linear combination of the selected features. Composite measures Q_i are designed for each of the five distortion classes by

$$Q_i = Q_0 + \sum_{j=1}^{N} \alpha_j F_j, \quad i = 1, \ldots, 5 \qquad (1)$$

where Q_0 is a constant term, α_j represents the coefficients (weights) for each feature and F_j are the selected features described in Tables III and IV. The parameter N describes the number of selected features, i.e., $N = 7$, except for the bit error degradation condition where only $N = 5$ features were selected. Subjective DMOS scores are used to train the composite measures. To this end, images belonging to each degradation class are separated into two disjoints sets and two-fold cross validation is performed. Table V exhibits the coefficients from the validation trial that resulted in superior quality estimation performance. In the table, features F_i refer to the features described in Tables III and IV in the order as they appear in the tables (from top to bottom); as an example, for JPEG compression, F_3 is equivalent to feature ES.1.

TABLE V

COEFFICIENTS FOR THE COMPOSITE MEASURES DESCRIBED BY (1)

F_i	Coefficients per distortion class				
	JPEG	JPEG2000	Noise	Blur	Bit error
Q_0	-77.83	-306.02	38.77	29.63	60.91
F_1	3.27	0.08	3.66	-59.89	-63.00
F_2	0.17	0.13	-0.01	0.35	0.07
F_3	0.51	7.06	-0.02	49.23	0.02
F_4	7.72	91.74	-23.69	-9.92	28.55
F_5	-0.01	-72.23	0.96	0.06	-0.01
F_6	-40.62	0.11	-219.28	-369.04	–
F_7	95.74	375.97	0.01	-0.02	–

TABLE VI

CONFUSION MATRIX FOR DEGRADATION-TYPE CLASSIFICATION.

True class	Predicted class				
	JPEG	JPEG2000	Noise	Blur	Bit Error
JPEG	86	0	0	0	1
JPEG2000	2	80	0	0	3
Noise	0	0	73	0	0
Blur	0	0	0	67	5
Bit error	3	2	0	0	68

D. Degradation Classification

As seen in Section III-B, different features are selected for different degradation types. As a consequence, degradation classification is needed such that an optimal composite measure is used. In this study, we employ radial basis function support vector classifiers (SVC), trained offline, for degradation-type classification [18]. Table VI presents the

TABLE VII

PERFORMANCE COMPARISON BETWEEN PROPOSED COMPOSITE METRICS AND SS.1 AND SS.2.

	Proposed		SS.1				SS.2			
	R	$RMSE$	R	%↑ R	$RMSE$	%↓ $RMSE$	R	%↑ R	$RMSE$	%↓ $RMSE$
JPEG	0.985	5.46	0.965	57.1	8.41	35.1	0.909	83.5	13.21	58.7
JPEG2000	0.968	6.26	0.960	20.0	7.34	14.8	0.885	72.1	12.12	48.4
Noise	0.989	4.22	0.977	50.0	5.85	27.9	0.946	78.7	9.07	53.5
Blur	0.979	3.82	0.934	67.4	6.81	43.9	0.957	50.0	5.26	27.4
Bit error	0.974	6.36	0.946	51.9	9.51	33.1	0.954	43.5	8.51	25.3
Average	–	–	–	49.3	–	31.0	–	65.6	–	42.7

(average) confusion matrix obtained for the cross-validation trials. An average misclassification rate of 4.1% is attained; as will be shown in Section IV, such classification errors are not detrimental to overall image quality measurement.

IV. EXPERIMENT RESULTS

Two-fold cross-validation is used to test the performance of the proposed composite measures; comparisons are carried out with SS.1 and SS.2, the two generic measures that attained the best overall performance. We report only the performance of the composite measures described by (1) since they resulted in superior performance relative to CART or MARS. Pearson correlation (R) and root-mean-square error ($RMSE$), averaged over the two cross validation trials, are used as performance figures to evaluate the three quality measures. To compute the performance of the benchmark metrics, 3^{rd} order polynomial fitting is applied to each metric in order to map it into the subjective DMOS scale.

Test results are reported in Table VII. Columns labeled "%↑ R" and "%↓ $RMSE$," describe the improvement in R and the decrease in $RMSE$, respectively, attained by the proposed composite measure over the benchmark. Since correlations attained by benchmark metrics are fairly high, we opt to report "improvement" in correlation, instead of "increase" in correlation. Correlation improvement is defined as:

$$\% \uparrow R = \frac{R_{comp} - R_{ind}}{1 - R_{ind}} * 100\% \qquad (2)$$

where R_{comp} and R_{ind} are the correlation attained by the proposed composite metrics and the individual benchmark metrics, respectively. The improvement indicates percentage reduction of the gap to perfect correlation. As can be seen, the proposed composite metrics are capable of improving correlation by an average 49.3% and 65.6% relative to SS.1 and SS.2, respectively. An average reduction in $RMSE$ of 31% and 42.7% is also attained. Careful analysis of the performance figures reported in Table VII also suggests that the proposed measures attain improved robustness against different degradation conditions.

V. CONCLUSIONS

We have investigated the use of three data mining algorithms to select the best seven features from a pool of 27 candidate image quality metrics. As an embodiment of the work, optimal feature subsets are selected for five different distortion classes and composite measures are designed for each class. Machine learning tools are used to detect, online, which degradation-specific composite measure to use. The proposed quality measurement method attains improved performance and is shown to be more robust against different degradation conditions.

REFERENCES

[1] Z. Wang and A. Bovik, *Modern Image Quality Assessment*, ser. Synthesis Lectures on Image, Video and Multimedia Processing. Morgan and Claypool, Feb. 2006.

[2] G. Fahmy and L. Karam, "Prediction of the quality of JPEG-compressed color images based on the SCIELAB metric," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, vol. 2, 2000, pp. 1054–1057.

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

[4] H. Sheikh, Z. Wang, L. Cormack, and A. Bovik, "LIVE image quality assessment database release 2." [Online]. Available: http://live.ece.utexas.edu/research/quality

[5] ITU-R BT-500.11, "Methodology for the subjective assessment of the quality of television pictures," Intl. Telecom. Union, 2002.

[6] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[7] A. van Dijk and J.-B. Martens, "Subjective quality assessment of compressed images," *Signal Processing*, vol. 58, pp. 235–252, 1997.

[8] H. Sheikh, M. Sabir, and A. 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.

[9] Z. Wang, H. Sheikh, and A. Bovik, "No-reference perceptual quality assessment of jpeg compressed images," in *Proc. IEEE Intl. Conf. on Image Processing*, Sept. 2002, pp. 477–480.

[10] Z. Wang and E. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Proc. SPIE Conf. Human Vision and Electronic Imaging*, Jan. 2005.

[11] I. Gunawan and M. Ghanbari, "Reduced-reference picture quality estimation using local harmonic amplitude information," in *Proc. London Communications Symposium*, Sept. 2003.

[12] I. Avcibas, B. Sankur, and K. Sayood, "Statistical evaluation of image quality measures," *Journal of Electronic Imaging*, vol. 11, no. 2, pp. 206–223, April 2002.

[13] D. Carevic and T. Caelli, "Region-based coding of color images using Karhunen-Loeve transform," *Graphical Models and Image Proc.*, vol. 59, no. 1, pp. 27–38, Jan. 1997.

[14] T. Frese, C. A. Bouman, and J. P. Allebach, "A methodology for designing image similarity metrics based on human visual system models," Purdue University, West Lafayette, IN, Tech. Rep. TR-ECE 97-2, Feb 1997.

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

[16] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks, 1984.

[17] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–141, March 1991.

[18] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995, New York.