This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

A ParaBoost Method to Image Quality Assessment

Tsung-Jung Liu, *Member, IEEE*, Kuan-Hsien Liu, *Member, IEEE*, Joe Yuchieh Lin, Weisi Lin, *Senior Member, IEEE*, and C.-C. Jay Kuo, *Fellow, IEEE*

***Abstract*— An ensemble method for full-reference image quality assessment (IQA) based on the parallel boosting (ParaBoost) idea is proposed in this paper. We first extract features from existing image quality metrics and train them to form basic image quality scorers (BIQSs). Then, we select additional features to address specific distortion types and train them to construct auxiliary image quality scorers (AIQSs). Both BIQSs and AIQSs are trained on small image subsets of certain distortion types and, as a result, they are weak performers with respect to a wide variety of distortions. Finally, we adopt the ParaBoost framework, which is a statistical scorer selection scheme for support vector regression (SVR), to fuse the scores of BIQSs and AIQSs to evaluate the images containing a wide range of distortion types. This ParaBoost methodology can be easily extended to images of new distortion types. Extensive experiments are conducted to demonstrate the superior performance of the ParaBoost method, which outperforms existing IQA methods by a significant margin. Specifically, the Spearman rank order correlation coefficients (SROCCs) of the ParaBoost method with respect to the LIVE, CSIQ, TID2008, and TID2013 image quality databases are 0.98, 0.97, 0.98, and 0.96, respectively.**

***Index Terms*— Auxiliary image quality scorer (AIQS), basic image quality scorer (BIQS), ensemble, parallel boosting (ParaBoost), scorer.**

I. Introduction

FOR decades, the peak signal-to-noise ratio (PSNR) has been the most well-known method to gauge the quality of an image or a video. Although researchers have doubts on the credibility of PSNR (or its relative mean-squared error (MSE)) [1], [2], it is still widely used nowadays. The main reason is that the PSNR is easy to compute. Wang *et al.* [3] proposed to use the structural similarity (SSIM) index to measure image quality in 2004, which has attracted a lot of attention due to its simplicity and good performance. With increasing demand of image quality assurance and assessment, more and more databases are made publicly available in recent years, such as LIVE [4], TID2008 [5], CSIQ [6], and TID2013 [7], to facilitate the development of image quality metrics. The SSIM and its variant IW-SSIM [8] work well across databases. The feature similarity (FSIM) index [9] outperforms the SSIM in several databases.

Recently, the learning-based approach emerges as a strong competitor in the image quality assessment (IQA) field since it is difficult to predict visual quality under various distortion types and rich image contents using a single formula [10], [11]. Examples of learning-based IQA methods can be found in [12]–[19], among many others. Simply speaking, they extract features from images, and use a machine learning approach to build a score prediction model (called a scorer), which is used to predict the perceived quality of test images. However, there exist many distortion types and it is difficult to find a single prediction model to cover all of them. Liu *et al.* [18] proposed a fusion approach called multi-method fusion (MMF) that fuses the scores of a couple of IQA methods to generate a new score using a machine learning method. These IQA methods include PSNR, SSIM, FSIM, and so on. They are called strong (or universal) scorers since they are not designed for specific distortion types. Intuitively, the fusion of stronger scorers will result in an even stronger scorer. Thus, it is not surprising that MMF outperforms each individual scorer in the ensemble.

In this paper, we still adopt an ensemble approach for full-reference IQA, yet examine the fusion of scores from a larger number of weak scorers. To derive weak scorers, we extract features from existing image quality metrics and train them to form basic image quality scorers (BIQSs). Next, we select additional features, which are useful in characterizing specific distortion types, and train them to construct auxiliary image quality scorers (AIQSs). Since BIQSs and AIQSs are only trained on small image subsets with certain distortion types, they are weak scorers with respect to a wide variety of distortions in an IQA database. Finally, we propose a parallel boosting (ParaBoost) scheme, which is a statistical scorer selection method for support vector regression (SVR), to fuse BIQSs and AIQSs to form an ensemble system to cope with a wide range of distortion types. The main advantage of the ParaBoost method is that we can design an image quality scorer (IQS) tailored to a specific distortion type and add it to the ensemble system. Thus, the corresponding IQA scoring system can be easily extended to images with new distortion types. Extensive experiments are conducted to demonstrate the superior performance of the ParaBoost method. Experimental results show that it outperforms existing IQA methods by a significant margin. Specifically, the Spearman rank order correlation coefficients (SROCCs) of the ParaBoost method with respect to the LIVE, CSIQ, TID2008, and

Manuscript received December 6, 2014; revised April 20, 2015, August 4, 2015, and November 2, 2015; accepted November 2, 2015. *(Corresponding author: Tsung-Jung Liu.)*

T.-J. Liu is with the Department of Electrical Engineering, National Chung Hsing University, Taichung 40227, Taiwan (e-mail: tjliu@dragon.nchu.edu.tw).

K.-H. Liu, J. Y. Lin, and C.-C. J. Kuo are with the Ming Hsieh Department of Electrical Engineering, Signal and Image Processing Institute, University of Southern California, Los Angeles, CA 90089 USA (e-mail: liuk@usc.edu; lin650@usc.edu; cckuo@sipi.usc.edu).

W. Lin is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2015.2500268

TID2013 image quality databases are 0.98, 0.97, 0.98, and 0.96, respectively.

The rest of this paper is organized as follows. We give a brief review on recent learning-based IQA methods in Section II. Then, BIQSs and AIQSs are presented in Section III. The ParaBoost method is described in Section IV. The process of selecting a suitable subset of scorers is discussed in Section V. Experimental results are reported in Section VI, where we conduct extensive performance comparisons with four image quality databases. Finally, concluding remarks are given in Section VII.

II. Review of Previous Work

The machine learning methodology has been applied to image quality evaluation. Narwaria and Lin [15] used the singular value decomposition (SVD) to quantify the major structural information in images and then adopt SVR to learn complex data patterns and map the detected features to scores for image quality prediction. The result is better than those obtained by formula-based methods.

Liu *et al.* [17], [18] proposed an MMF method for IQA. It is motivated by the observation that no single method gives the best performance in all situations. A regression approach is used to combine the scores of multiple IQA methods in the MMF. First, a large number of image samples are collected, each of which has a score labeled by human observers and scores associated with different IQA methods. The MMF score is obtained by a nonlinear combination of scores computed by multiple methods (including SSIM [3], FSIM [9], and so on) with suitable weights obtained by a training process. To improve the predicted scores furthermore, distorted images are classified into five groups based on distortion types, and regression is performed within each group, which is called the context-dependent MMF (CD-MMF). So far, MMF offers one of the best IQA results in several popular databases, such as LIVE, CSIQ, and TID 2008.

A block-based MMF [19] method was also proposed for IQA. First, an image is decomposed into small blocks. Blocks are then classified into three types (smooth, edge, and texture), while distortions are classified into five groups. Finally, one proper IQA metric is selected for each block based on the block type and the distortion group. Pooling over all blocks leads to the final quality score of a test image. It offers competitive performance with the MMF for the TID2008 database.

As compared with the previous work, the ParaBoost method proposed in this paper has several unique characteristics.

1) It fuses scores from a set of weak IQSs, where each IQS can be designed to predict the quality of some specific image distortion types. The proposed ParaBoost system can perform well in situations where individual IQS cannot perform well.
2) An IQS bank structure is adopted to optimize the overall performance of the IQA system. The structure of the ensemble system is modular so that we can add or discard an IQS easily depending on the application need.
3) Each IQS is built by training images on different distortion types to increase the diversity among all scorers, which can help optimize the ensemble performance.
4) No need for distortion classification stage, which can degrade the overall performance when the classification rate is low.
5) The statistical methods for scorer selection can help the overall system achieve the optimal performance with the smallest number of scorers.

III. Feature Extraction for Image Quality Scorers

In this paper, we use the term IQS, to denote a method that can give a quality score to an image. Two IQS types are examined in this section: 1) BIQSs and 2) AIQSs. BIQSs are derived from well-known IQA metrics, while AIQSs are designed to tailor to specific distortion types. In this section, we focus on feature extraction for BIQSs and AIQSs.

A. Features for Basic Image Quality Scorers

We derive features for BIQSs from several well-known image quality metrics by decomposing their contributing factors. To be more specific, we choose three components (i.e., luminance (L), contrast (C), and structure (S)) of SSIM [3] as features of the first three BIQSs, respectively. Furthermore, we extract two other components, namely, phase congruency (PC) and gradient magnitude (GM), from FSIM [9] and use them as features for the fourth and fifth scorers since PC and GM have totally different characteristics with L, C, and S. Finally, PSNR is selected as the sixth basic scorer because of its simplicity and superior capability in predicting quality of images with additive noise [17], [18]. These six BIQSs are detailed below.

As aforementioned, the features of the first three BIQSs are the similarity measures of luminance (L), contrast (C), and structure (S) between reference and distorted images, respectively. Suppose $\mathbf{x}$ represents both image patches extracted from the same spatial location of reference and distorted images, and $\mu_r(\mathbf{x})$, $\mu_d(\mathbf{x})$, $\sigma_r^2(\mathbf{x})$, $\sigma_d^2(\mathbf{x})$, and $\sigma_{rd}(\mathbf{x})$ are the means, the variances, the covariance of $\mathbf{x}$ from the reference and distorted images, respectively. When there are N such image patches for the whole image I, the luminance similarity measure between the two images is selected as the feature for BIQS #1

$$\text{BIQS \#1}: S_L = \frac{1}{N}\sum_{\mathbf{x}\in I}\frac{2\mu_r(\mathbf{x})\mu_d(\mathbf{x}) + C_1}{\mu_r^2(\mathbf{x}) + \mu_d^2(\mathbf{x}) + C_1}. \tag{1}$$

The contrast similarity measure is selected as the feature for BIQS #2

$$\text{BIQS \#2}: S_C = \frac{1}{N}\sum_{\mathbf{x}\in I}\frac{2\sigma_r(\mathbf{x})\sigma_d(\mathbf{x}) + C_2}{\sigma_r^2(\mathbf{x}) + \sigma_d^2(\mathbf{x}) + C_2}. \tag{2}$$

The structure similarity measure is selected as the feature for BIQS #3

$$\text{BIQS \#3}: S_S = \frac{1}{N}\sum_{\mathbf{x}\in I}\frac{\sigma_{rd}(\mathbf{x}) + C_3}{\sigma_r(\mathbf{x})\sigma_d(\mathbf{x}) + C_3}. \tag{3}$$

TABLE I
IMAGE DISTORTION TYPES IN TID2008 DATABASE

Type	Type of distortion
1	Additive Gaussian Noise
2	Different additive noise in color components
3	Spatially correlated noise
4	Masked noise
5	High-frequency noise
6	Impulse noise
7	Quantization noise
8	Gaussian blur
9	Image denoising
10	JPEG compression
11	JPEG2000 compression
12	JPEG transmission errors
13	JPEG2000 transmission errors
14	Non-eccentricity pattern noise
15	Local block-wise distortions of different intensity
16	Mean shift (intensity shift)
17	Contrast change

Constants in (1)–(3) are $C_1 = (K_1 D)^2, C_2 = (K_2 D)^2, C_3 = C_2/2, K_1 = 0.01$, and $K_2 = 0.03$, to avoid instability [3], and D is the dynamic range of pixel values (i.e., $D = 255$ for the 8-bit pixel representation).

The fourth and fifth BIQSs measure the similarity of PC and GM between the reference and the distorted images. With the same notation above, we assume $\mathbf{x}$ represents both image patches extracted from the same spatial location of reference and distorted images, and $PC_r(\mathbf{x})$, $PC_d(\mathbf{x})$, $GM_r(\mathbf{x})$, and $GM_d(\mathbf{x})$ are PCs and GMs of $\mathbf{x}$ from the reference and the distorted images, respectively. For an image I with N image patches, the PC similarity measure between the two images is selected as the feature for BIQS #4

$$\text{BIQS \#4: } S_{PC} = \frac{1}{N}\sum_{\mathbf{x}\in I}\frac{2PC_r(\mathbf{x})PC_d(\mathbf{x}) + T_1}{PC_r^2(\mathbf{x}) + PC_d^2(\mathbf{x}) + T_1} \quad (4)$$

and the GM similarity between two images is selected as the feature for BIQS #5

$$\text{BIQS \#5: } S_{GM} = \frac{1}{N}\sum_{\mathbf{x}\in I}\frac{2GM_r(\mathbf{x})GM_d(\mathbf{x}) + T_2}{GM_r^2(\mathbf{x}) + GM_d^2(\mathbf{x}) + T_2} \quad (5)$$

where T_1 and T_2 are the positive constants which are added to avoid instability of S_{PC} and S_{GM}.

The sixth BIQS is based on the PSNR value, which is related to the mean-squared error (MSE). For two images I_r and I_d, of size $X \times Y$, the MSE can be computed via

$$\text{MSE} = \frac{1}{XY}\sum_x\sum_y [I_r(x,y) - I_d(x,y)]^2. \quad (6)$$

Then, the PSNR value in decibels is used as the feature for BIQS #6

$$\text{BIQS \#6: } PSNR = 10\log\frac{D^2}{\text{MSE}} \quad (7)$$

where D is the maximum value that a pixel can take (e.g., 255 for 8-bit images), as aforementioned.

B. Features of Auxiliary Image Quality Scorers

There are 17 image distortion types in the TID2008 database [5], [20], which are listed in Table I for easy reference.

TABLE II
SROCC PERFORMANCE OF BIQSS VERSUS DISTORTION TYPES IN TID2008 DATABASE

Distortion type	BIQS #1 (S_L)	BIQS #2 (S_C)	BIQS #3 (S_S)	BIQS #4 (S_{PC})	BIQS #5 (S_{GM})	BIQS #6 ($PSNR$)
1	0.0710	0.7052	0.7958	0.4896	0.8737	0.8917
2	0.0520	0.5340	0.7688	0.0431	0.8202	0.8814
3	0.0735	0.7304	0.8112	0.6090	0.8581	0.9089
4	-0.1183	0.0534	0.6877	0.0504	0.2824	0.8274
5	0.1334	0.7885	0.8796	0.4593	0.9036	0.9182
6	0.1077	0.4458	0.6942	0.3724	0.7113	0.9041
7	0.5365	0.8079	0.7788	0.7197	0.8347	0.8615
8	0.1384	0.9265	0.9575	0.8115	0.9460	0.8668
9	0.5729	0.9483	0.9424	0.8310	0.9552	0.9340
10	0.2217	0.8688	0.9097	0.7747	0.9405	0.8945
11	0.1187	0.9451	0.9520	0.9255	0.9720	0.8033
12	-0.0244	0.7597	0.8360	0.5758	0.8570	0.7142
13	0.1143	0.7543	0.8444	0.7117	0.8695	0.7687
14	0.1445	0.4260	0.6587	0.5335	0.6902	0.5551
15	0.0449	0.7749	0.8264	0.6534	0.8365	0.5456
16	0.6778	0.1590	0.0651	0.0443	0.2910	0.6750
17	-0.0509	0.6604	0.0316	0.0509	0.7658	0.6097

In Table II, we summarize the SROCC between objective and subjective scores for the performance of each BIQS with respect to all distortion types. The higher the SROCC is, the better match between these two scores. As indicated in Table II, each BIQS has its respective advantage in predicting image quality scores for certain distortion types. For example, BIQS #5 can predict the image quality quite well for distortion types 8–15 and 17, while BIQS #6 has the best performance for distortion types 1–7 among six BIQSs.

As we can see in Table II, several distortion types (e.g., types 14, 16, and 17) cannot be handled well even with all six BIQSs. Hence, we need to find more features to design new scorers to boost the performance. These scorers are called AIQSs since they are designed to support BIQSs in addressing specific distortion types.

The feature of the first AIQS is the zero-crossing (ZC) rate [21], which is defined as

$$z_h(i,j) = \begin{cases} 1, & \text{ZC happens at } d_h(i,j) \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where $d_h(i,j) = x(i,j+1) - x(i,j)$, $j \in [1, N-1]$, is the difference signal along the horizontal line, and $x(i,j)$, $i \in [1, M]$, $j \in [1, N]$ for an image of size $M \times N$. The horizontal ZC rate can be written as

$$Z_h = \frac{1}{M(N-2)}\sum_{i=1}^{M}\sum_{j=1}^{N-2} z_h(i,j). \quad (9)$$

We can calculate the vertical component Z_v in a similar fashion. Finally, the overall ZC rate is selected as the feature for AIQS #1 as shown

$$\text{AIQS \#1: } ZC = \frac{Z_h + Z_v}{2}. \quad (10)$$

From [21], we know ZC rate can be used as an index to measure the signal activity of images. Since high-frequency noise and JPEG2000 compression can cause the reduction of signal activity, AIQS #1 is particularly useful in evaluating distortion type 5 (high-frequency noise) and distortion type 11 (JPEG 2000 compression).

The feature of the second AIQS is derived from the gray-level co-occurrence matrix (GLCM), which is also known

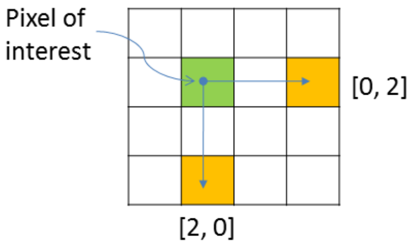


Fig. 1. Spatial relationship of pixel of interest in GLCM.

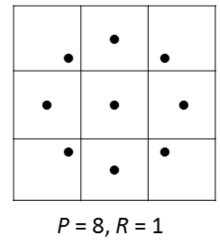


Fig. 2. Circularly symmetric neighbor set in LBP.

as the gray-tone spatial-dependence matrix [22]. The GLCM characterizes the texture of an image by calculating how often a pixel with intensity (gray level) value l occurs in a specific spatial relationship to a pixel with value m. Here, we are interested in two spatial relationships, as shown in Fig. 1. Each element at (l, m) in the resultant GLCM is simply the sum of frequencies for one pixel with value l and its neighbor pixel satisfying the desired spatial relationship with value m. Then, the contrast difference feature of GLCM is selected as the feature of AIQS #2 as given by

$$\text{AIQS \#2}: \sum_{l,m} |l-m|^2 p(l,m) \tag{11}$$

where $p(l, m)$ is the joint probability for the occurrence of pixel pairs having gray level values l and m with a defined spatial relationship (say, Fig. 1) in the image. In essence, the second AIQS in (11) returns a measure of the intensity contrast between a pixel and its horizontal and vertical neighbors over the whole image. AIQS #2 is also useful for distortion type 11 (JPEG 2000 compression).

The features of the third AIQS are derived from the rotation-invariant and uniform local binary pattern (LBP) operator [23], which is in the form of

$$\mathrm{LBP}_{P,R}^{\mathrm{riu2}} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c), & \text{if } U(\mathrm{LBP}_{P,R}) \le 2 \\ P+1, & \text{otherwise} \end{cases} \tag{12}$$

where

$$U(\mathrm{LBP}_{P.R}) = |s(g_{p-1} - g_c) - s(g_0 - g_c)| + \sum_{p=1}^{P-1} |s(g_p - g_c) - s(g_{p-1} - g_c)| \tag{13}$$

and

$$s(g_p - g_c) = \begin{cases} 1, & \text{if } g_p \ge g_c \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

with g_c corresponding to the gray value of the center pixel of the local neighborhood and $g_p (p = 0, \ldots, P-1)$ corresponding to the gray values of P equally spaced pixels on a circle of radius $R (R > 0)$ that form a circularly symmetric neighbor set. The superscript $^{\mathrm{riu2}}$ stands for the use of rotation-invariant uniform patterns that have a U value of at most 2. Then, the features of AIQS #3 can be written as

$$\text{AIQS \#3}: \sqrt{|nh_r(b) - nh_d(b)|} \tag{15}$$

where $b = 0, 1, \ldots, P+1$ represents the bin of the histogram, and nh_r and nh_d denote the normalized histograms of (12) for reference and distorted images, respectively. As shown in Fig. 2, we choose $P = 8$ and $R = 1$ for simplicity of the LBP operator. Therefore, we have ten (i.e., $P+2$) values to represent the scorer in (15). If a larger P value is chosen, then the complexity of LBP operator will be higher because the implementation needs a lookup table of 2^P elements [23]. AIQS #3 is useful in evaluating distortion types 6–11.

The features of the fourth and the fifth AIQSs are used to characterize the edge structure of an image. An image is first divided into M non-overlapping 16×16 patches, and the Sobel edge operator [24], [25] is used in each patch to generate horizontal gradient g_h and vertical gradient g_v. Then, we can obtain the edge magnitude and edge orientation via $(g_v^2 + g_h^2)^{1/2}$ and $\tan^{-1}(g_v/g_h)$, respectively. Here, we chose the gradient orientation since it was shown to be robust to illumination change and also successfully applied to many areas, such as disparity estimation [26], video quality assessment (VQA) [27], and age estimation [28] tasks.

Suppose that $h_{1,i}(b)$ and $h_{2,i}(b)$ represent the n-bin histograms of edge magnitude from the i-th image patch of the reference and distorted images, respectively, and we can compute the root MSE (RMSE) between histograms $h_{1,i}(b)$ and $h_{2,i}(b)$ by

$$\mathrm{RMSE}_{\mathrm{mag},i} = \sqrt{\frac{1}{n}\sum_{b=1}^{n}[h_{1,i}(b) - h_{2,i}(b)]^2}, \quad i = 1, \ldots, M \tag{16}$$

where b denotes the bin of the histogram, n is the number of bins ($n = 10$ in this paper for low computation), and M represents the number of non-overlapping 16×16 image patches, which also depends on the image size. Similarly, $h_{3,i}(b)$ and $h_{4,i}(b)$ represent the histograms of edge orientation from the i-th image patch of the reference and distorted images, respectively, and then the RMSE between histograms $h_{3,i}(b)$ and $h_{4,i}(b)$ can be computed in a similar way as (16)

$$\mathrm{RMSE}_{\mathrm{ori},i} = \sqrt{\frac{1}{n}\sum_{b=1}^{n}[h_{3,i}(b) - h_{4,i}(b)]^2}, \quad i = 1, \ldots, M. \tag{17}$$

Finally, the features of the fourth and fifth AIQSs are given by

$$\text{AIQS \#4}: [\mathrm{RMSE}_{\mathrm{mag},1}, \ldots, \mathrm{RMSE}_{\mathrm{mag},M}]^T \tag{18}$$

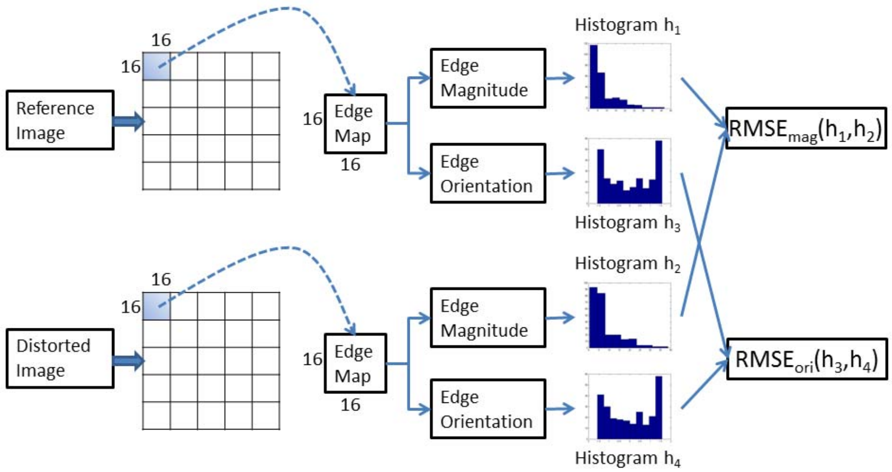


Fig. 3. Extracting features of the fourth and fifth AIQSs.

and

$$\text{AIQS \#5}: [\text{RMSE}_{\text{ori},1}, \ldots, \text{RMSE}_{\text{ori},M}]^T. \quad (19)$$

We show the procedure of extracting features for the fourth and the fifth AIQSs in Fig. 3. These two AIQSs are useful in evaluating distortion type 14 since the non-eccentricity pattern noise can be easily spotted on the edge map, where the features can be captured by the AIQSs #4 and #5.

Features of other AIQSs are also derived from a local region. We divide an image into N non-overlapping blocks of size 8×8 and compute three quantities, known as the MSE, the mean difference ratio (MDR), and the contrast difference ratio (CDR), respectively, for each block

$$\text{MSE}_j = \frac{1}{64}\sum_{x=1}^{8}\sum_{y=1}^{8}[\text{IB}_r(x,y) - \text{IB}_d(x,y)]^2 \quad j = 1, \ldots, N \quad (20)$$

$$\text{MDR}_j = \frac{\mu_d - \mu_r}{\mu_r}, \quad j = 1, \ldots, N \quad (21)$$

$$\text{CDR}_j = \frac{c_d - c_r}{c_r}, \quad j = 1, \ldots, N \quad (22)$$

where $\mu_d = (1/64)\sum_{x=1}^{8}\sum_{y=1}^{8}[\text{IB}_d(x,y)]$, $\mu_r = (1/64) \sum_{x=1}^{8}\sum_{y=1}^{8}[\text{IB}_r(x,y)]$, $c_d = \max\{\text{IB}_d(x,y)\} - \min\{\text{IB}_d(x,y)\}$, $c_r = \max\{\text{IB}_r(x,y)\} - \min\{\text{IB}_r(x,y)\}$, and where IB_d and IB_r represent the 8×8 image blocks of the distorted and the reference images, respectively. Then, the features of the sixth to the eighth AIQSs are the 10-bin histograms of (20)–(22)

$$\text{AIQS \#6}: h_{10-\text{bin}}(\text{MSE}_j | j = 1, \ldots, N) \quad (23)$$

$$\text{AIQS \#7}: h_{10-\text{bin}}(\text{MDR}_j | j = 1, \ldots, N) \quad (24)$$

$$\text{AIQS \#8}: h_{10-\text{bin}}(\text{CDR}_j | j = 1, \ldots, N). \quad (25)$$

AIQS #6 and #8 are designed to address distortion types 15 and 17, respectively, because the local MSE can capture the blockwise distortions, and local contrast difference can test if there is a contrast change; while AIQS #7 can be used to boost the performance with respect to distortion type 16 since the local mean difference can be used to detect the shift in mean.

In above, AIQSs #1 and #2 extract global features, but AIQSs #3–8 belong to local feature extractors. Thus, more AIQSs need to be designed to evaluate global distortions. In addition, images with distortion types 14 and 16 are the most difficult ones to be assessed well. Therefore, using only AIQSs #1–8 are still not good enough to deal with all types of distortions and three more AIQSs (i.e., AIQSs #9–11), which are designed to target at distortion types 14 and 16, become necessary for this purpose.

For the ninth AIQS, we consider the mean absolute difference (mAD) between reference (I_r) and distorted (I_d) images. For reference and distorted images of size $M \times N$, we use three components of the YIQ color space simultaneously to account for differences in different color dimensions [24], [25]. The mAD in Y, I, and Q components can be described

$$\text{mAD}_i = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}|I_{r,i}(m,n) - I_{d,i}(m,n)|, \quad i = Y, I, Q. \quad (26)$$

Then, the feature vector of the ninth AIQS is formed by concatenating the three components of mAD in (26)

$$\text{AIQS \#9}: [\text{mAD}_Y, \text{mAD}_I, \text{mAD}_Q]^T. \quad (27)$$

AIQS #9 is a good scorer for distortion types 1–13. Besides, it can be used to boost the performance with respect to distortion type 14.

The mean shift distortion (i.e., distortion type 16 in TID2008) for color images is one of the most challenging distortion types for BIQSs, as shown in Table II. We design two more AIQSs (i.e., AIQS #10 and AIQS #11) to boost the overall performance. The tenth AIQS measures the histogram range (or the dynamic range (DR)) of distorted images in three components of YIQ, respectively. We use the 256-bin histograms for Y, I, and Q components of distorted image I_d, respectively, as shown in Fig. 4. The DR of Y, I, and Q components can be written as

$$\text{DR}_i(I_d) = b_{1,i} - b_{0,i}, \quad i = Y, I, Q \quad (28)$$

where $b_{0,i}$ and $b_{1,i}$ are, respectively, the first and the last bins in the histogram with values significantly larger than zero

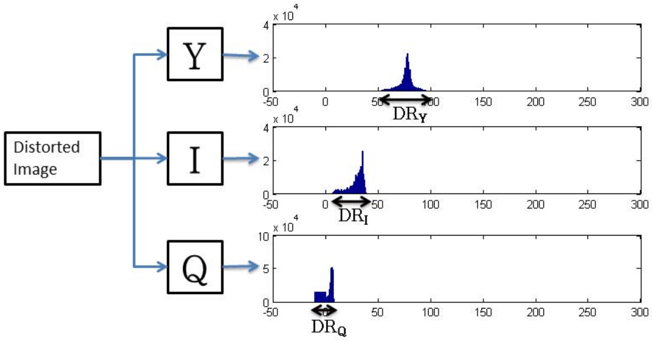


Fig. 4. Extracting features of the tenth AIQS.

(i.e., the value should be at least greater than 1 to reflect the existence of that pixel value). Then, the feature vector of the tenth AIQS is

$$\text{AIQS \#10}: [\text{DR}_Y(I_d), \text{DR}_I(I_d), \text{DR}_Q(I_d)]^T. \quad (29)$$

Finally, the feature vector of AIQS #11 is a 15-element vector consisting of the global mean shift and the DR in YIQ, YCbCr, and RGB color spaces. It is in a form of

$$\text{AIQS \#11}: \left[\frac{\Delta\mu_Y}{\mu_Y(I_r)}, \frac{\Delta\mu_I}{\mu_I(I_r)}, \frac{\Delta\mu_Q}{\mu_Q(I_r)}, \frac{\Delta\mu_R}{\mu_R(I_r)}, \frac{\Delta\mu_G}{\mu_G(I_r)}, \frac{\Delta\mu_B}{\mu_B(I_r)}, \text{DR}_Y(I_r), \text{DR}_I(I_r), \text{DR}_Q(I_r), \text{DR}_R(I_r), \text{DR}_G(I_r), \text{DR}_B(I_r), \frac{\Delta\mu_Y}{\text{DR}_Y(I_r)}, \frac{\Delta\mu_{Cb}}{\text{DR}_{Cb}(I_r)}, \frac{\Delta\mu_{Cr}}{\text{DR}_{Cr}(I_r)}\right]^T \quad (30)$$

where $\Delta\mu_i = \mu_i(I_d) - \mu_i(I_r)$, $i = Y, I, Q, R, G, B, Cb, Cr$, and $\mu_i(I_d)$ and $\mu_i(I_r)$ denote the global mean of distorted and reference images, respectively, on color component i, and $\text{DR}_i(I_r)$ has the same definition as that in (28) except that it is applied to reference image I_r instead of distorted image I_d.

The 11 AIQSs as described in this section are designed to complement BIQSs to account for some distortion types that are difficult to be assessed. We will elaborate this point in Section IV. Moreover, AIQS #1, 2, and 10 only need to extract features for distorted images. Thus, the complexity of these three AIQSs is lower. In addition, another advantage is that they can be used to develop no-reference (blind) IQA methods.

IV. IQS Evaluation, Training, and ParaBoost

A. Contribution Evaluation of IQSs

We list the individual performance of all BIQSs and AIQSs for the TID2008 database in Table III. Most BIQSs (except for BIQS #1, this is understandable since it is just a similarity measures of luminance) have better performance than the AIQSs. AIQSs are not suitable to work alone because of their poor performance (with SROCC < 0.5). However, they can be used together with BIQSs to boost the overall performance of the entire IQA system.

TABLE III
Performance of Each IQS in TID2008 Database

IQS	PCC	SROCC	RMSE
BIQS #1	0.1444	0.4748	1.3279
BIQS #2	0.7556	0.7798	0.8790
BIQS #3	0.7766	0.7683	0.8454
BIQS #4	0.8126	0.8016	0.7821
BIQS #5	0.8246	0.8353	0.7591
BIQS #6	0.5422	0.5477	1.1276
AIQS #1	0.5057	0.2844	1.1577
AIQS #2	0.4006	0.1626	1.2295
AIQS #3	0.3579	0.3187	1.2530
AIQS #4	0.4394	0.3479	1.2054
AIQS #5	0.5189	0.4886	1.1471
AIQS #6	0.2649	0.2268	1.2940
AIQS #7	0.1578	0.2906	1.3251
AIQS #8	0.4026	0.2534	1.2284
AIQS #9	0.3451	0.2679	1.2595
AIQS #10	0.1714	0.4536	1.3221
AIQS #11	0.0936	0.0705	1.3360

To demonstrate this point, we show the SROCC performance of 11 AIQSs for 17 distortion types in the TID2008 database [5] in Table IV. We are particularly interested in the use of AIQSs to boost the overall system performance in evaluating images with distortion types 14 (i.e., non-eccentricity pattern noise), 16 (mean shift), and 17 (contrast change) with BIQSs in a parallel configuration.

Apparently, AIQS #4 has superior performance against distortion type 14. In addition, as compared with Table II, AIQS #8 and AIQS #11 can significantly improve the correlation performance for distortion types 17 and 16, respectively. We will give a brief discussion on why these AIQSs work well on these special distortion types below.

The non-eccentricity pattern noise can be differentiated more easily when an image is transformed into its edge map, as shown in Fig. 5, where the non-eccentricity pattern noise occurs in the logo region of the hats and significant differences can be observed by comparing Fig. 5(c) and (d). As a result, the difference between the distorted and the reference images can be captured by comparing the histogram difference of their GMs as done in the feature extraction of AIQS #4. The mean shift difference in images can be quantified via features of AIQS #11, which considers the global mean shift and the DR of histograms in several color spaces. The contrast change can be described by computing the contrast difference between the distorted and reference images as done in deriving the features of AIQS #8.

We show the use of AIQSs to boost the overall performance for distortion types 14, 16, and 17 in Table V. We see that the inclusion of AIQS #8 can boost the SROCC performance by over 0.05 on distortion type 17. Furthermore, by including three more AIQSs (#4, 6, 9) and four more AIQSs (#3, 7, 10, 11), we are able to boost the performance for distortion types 14 and 16, respectively.

Generally speaking, we can classify IQSs into two categories based on their feature types as given in Table VI, i.e., global features that grasp viewers' quality impression of the whole image and local features that capture fine details in local regions. A good IQA system should contain both of them.

TABLE IV
SROCC PERFORMANCE OF AIQSS VERSUS DISTORTION TYPES IN TID2008 DATABASE

Distortion type	AIQS #1	AIQS #2	AIQS #3	AIQS #4	AIQS #5	AIQS #6	AIQS #7	AIQS #8	AIQS #9	AIQS #10	AIQS #11
1	0.7184	0.3652	0.5019	0.5695	0.4783	0.2539	0.2774	0.3088	0.8870	0.5694	0.2305
2	0.7570	0.5334	0.4340	0.4031	0.4140	-0.1069	0.3155	0.5132	0.8806	0.5764	-0.0042
3	0.3884	0.2313	0.4500	0.2318	0.5161	-0.0138	0.5135	0.0687	0.9073	0.5844	-0.0440
4	0.7084	0.1674	0.6409	0.4535	0.4843	0.2118	-0.0293	0.1851	0.8209	0.6795	-0.0187
5	0.8437	0.7557	0.7080	0.2265	0.4811	0.5020	0.6710	0.3127	0.9182	0.8115	0.2220
6	0.3992	0.5363	0.8767	-0.0353	0.7002	0.8531	0.4903	0.2363	0.8809	0.6653	0.3722
7	0.5954	-0.0935	0.6679	-0.0243	0.0436	0.6057	0.4506	0.6224	0.8593	0.8419	-0.0456
8	0.6956	0.3625	0.8617	0.4958	0.4580	0.3183	0.4996	0.3001	0.8506	0.2191	-0.0504
9	0.4301	0.5636	0.7815	0.1094	0.1301	0.3947	0.6497	0.7247	0.9164	0.7819	0.5321
10	0.7422	0.2103	0.7936	0.3010	0.5044	0.0540	-0.0292	0.7255	0.9234	0.3789	0.1443
11	0.8874	0.7989	0.8728	0.2133	0.1438	0.2869	0.7203	0.7061	0.8697	-0.1221	0.1997
12	-0.0890	0.4201	0.7974	0.4382	0.6974	0.4743	0.5328	0.5672	0.8247	0.3857	0.1152
13	0.3790	0.1668	0.5022	0.5380	0.4248	0.3105	0.1709	0.3213	0.8756	0.1237	-0.1000
14	0.3041	-0.1553	0.7011	0.8737	0.7489	0.0324	0.4853	0.6194	0.5992	-0.0290	-0.0197
15	-0.1225	0.4097	0.7440	0.2972	0.7534	0.8208	0.4836	0.2471	0.7695	0.3776	0.0458
16	0.0942	0.3180	0.3763	0.3345	0.2740	0.2873	0.4294	0.2864	0.6785	0.4506	0.7207
17	0.0215	0.4746	0.7040	0.2956	0.1644	0.1861	0.7749	0.8324	0.5826	0.8259	0.7426

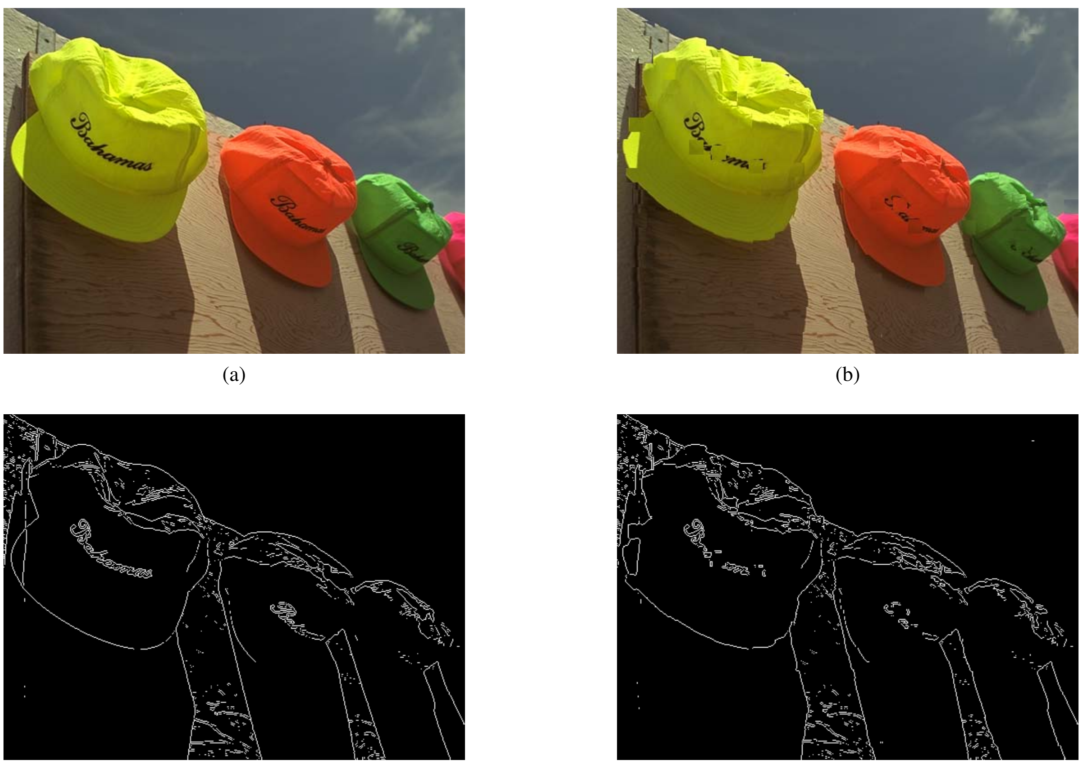

(a) (b) (c) (d)

Fig. 5. Original and distorted Hat images and their corresponding Sobel edge maps. (a) Image (reference). (b) Image (non-eccentricity pattern noise). (c) Sobel edge map of (a). (d) Sobel edge map of (b).

B. Training of BIQS and AIQS Models

We train a model for each BIQS with proper image subsets, as shown in Table VII. The training image subset is chosen according to the following criterion. If the BIQS works well for some specific distortion types, then the images associated with those distortion types will be chosen to train the respective BIQS. For instance, in Table II, BIQS #2 has better SROCC performance with respect to distortion types 8–11 than the others. Therefore, we choose the image subset belonging to distortion types 8–11 to train the model of BIQS #2. The detailed procedure is described as follows. First, the images are divided into 5 folds and the 5-fold cross-validation is applied. The training is only conducted on four of the 5 folds each time, with the images having distortion types specified in Table VII. The images with the same corresponding distortion types in the 5-th fold are used for testing. For example, the model of BIQS #6 is trained on the images with distortion types 1–7. We can obtain a PSNR value for each trained image. Then, each PSNR value can be treated as a feature to train and build the model of BIQS #6. This strategy is necessary

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

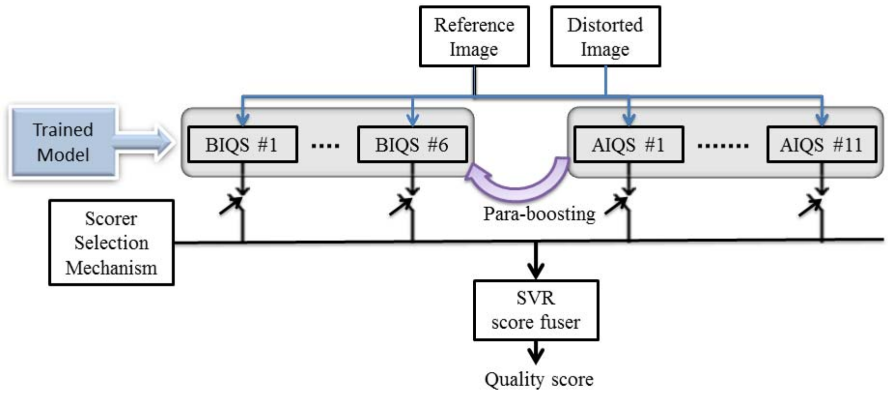


Fig. 6. The ParaBoost IQA system.

TABLE V
SROCC PERFORMANCE BOOSTING BY CONSIDERING BOTH BIQS AND AIQS IN TID2008 DATABASE

Distortion type	All BIQSs	All BIQSs + AIQS #4,6,9	All BIQSs + AIQS #3,7,10,11	All BIQSs + AIQS #8
14	0.9182	0.9345	-	-
16	0.9446	-	0.9477	-
17	0.8412	-	-	0.8978

TABLE VI
IQS CATEGORIZATION

	IQS
Global features	BIQS #6 AIQS #1,2,9,10,11
Local features	BIQS #1,2,3,4,5 AIQS #3,4,5,6,7,8

TABLE VII
TRAINING IMAGE SETS (DISTORTION TYPES) FOR BIQSs

BIQS #	1	2	3	4	5	6
Training image distortion type	16	8-11	8-15	8-11	8-15,17	1-7

because it offers some advantages, including: 1) the saving of training time since training is only conducted on several distortion types instead of all distortion types; 2) increased diversity among BIQSs because they are trained on different image subsets [29]; and 3) the enhanced performance for a small number of distortion types since each BIQS is trained for the distortion types where it is supposed to perform well.

Similarly, we train a model for each AIQS with only one distortion type, as shown in Table VIII since each AIQS is designed to target at one difficult distortion type only. For example, we have low SROCC performance for distortion types 14 and 16, as shown in Table II. To boost the performance, we add three AIQSs (#4, 6, 9) for distortion type 14 and four AIQSs (#3, 7, 10, 11) for distortion type 16. Thus, the training time can be saved and each AIQS can target at one distortion type, becoming an expert for one specific type of image distortion. Moreover, several scorers can cooperate and conquer the extremely difficult ones.

TABLE VIII
TRAINING IMAGE SETS (DISTORTION TYPES) FOR AIQSs

AIQS #	1	2	3	4	5	6	7	8	9	10	11
Training image distortion type	13	15	16	14	12	14	16	17	14	16	16

C. ParaBoost

The final IQA system, including both BIQS and AIQS, is shown in Fig. 6. We call it a ParaBoost system, since AIQSs are used to boost the overall system performance along with BIQSs in a parallel configuration.

In general, we consider a ParaBoost system consisting of n BIQSs and r AIQSs. Given m training images, we can obtain the quality score of the i-th training image for each individual IQS, which is denoted by $s_{i,j}$, where $i = 1, 2, \ldots, m$, and $j = 1, 2, \ldots, n + r$. Then, the quality score of the ParaBoost system can be modeled as

$$\text{ParaBoost}(\mathbf{s}_i) = \mathbf{w}^T \varphi(\mathbf{s}_i) + b \tag{31}$$

where $\mathbf{s}_i = (s_{i,1}, \ldots, s_{i,n+r})^T$ is the quality score vector for the i-th image, $\mathbf{w} = (w_1, \ldots, w_{n+r})^T$ is the weighting vector, $\varphi(\cdot)$ denotes a fixed feature-space transformation, and b is the bias.

In the training stage, we determine weight vector $\mathbf{w}$ and bias b from the training data that minimize the difference between ParaBoost($\mathbf{s}_i$) and the (differential) mean opinion score ($(D)\text{MOS}_i$) obtained by human observers, namely

$$\min_{\mathbf{w},b} \|\text{ParaBoost}(\mathbf{s}_i) - (D)\text{MOS}_i\|_1, \quad i = 1, \ldots, m \tag{32}$$

where $\|\cdot\|_1$ denotes the l_1 norm.

To solve this problem, we demand that the maximum absolute difference in (32) is bounded by a certain level ε, and adopt the SVR [30] for its solution. We choose the radial basis function (RBF) as the kernel function in SVR. A linear kernel is also tested, yet its performance is not as good. One explanation is that quality score vector $\mathbf{s}_i$ and MOS_i (or $D\text{MOS}_i$) are not linearly correlated. For this reason, we only show results with the nonlinear RBF kernel for the rest of this paper.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

In the test stage, we define the quality score vector $\mathbf{s}_k$ of the k-th test image, where $k = 1, 2, \ldots, l$ and where l denotes the number of test images, and (31) is used to determine the quality score of the ParaBoost method, ParaBoost($\mathbf{s}_k$). In all experiments, we divide all the images into two sets (training and testing sets) and use the n-fold ($n = 5$) cross-validation, which is a widely used strategy in machine learning [31], [32], to select our training and testing sets. First, we equally divide all distorted images into five non-overlapping sets. One set is used for testing while the remaining four sets are used for training. For instance, in TID2008, there are 1700 distorted images totally. The size of training set is 1360 images, and the size of testing set is 340 images. We rotate this assignment five times so that each set is only used as the testing set once. Using this configuration, we can obtain the predicted objective quality scores for all the images in one of the five folds whenever we do the rotation once. The score results from the 5 folds are then combined to compute the overall correlation coefficients and the RMSE. This procedure can test if overfitting occurs.

In addition, note that the reference images in the training and testing sets are non-overlapping. For example, in TID2008, there are 25 reference images. The training set only comes from the distorted images associated with 20 of 25 reference images, and the testing set is selected from the ones belonging to the remaining five reference images. To be more specific, for training the model of AIQS #5, we choose the images belonging to both the first 20 reference images and also distortion type 12. Hence, 80 images are selected to train the model of AIQS #5.

Before applying the SVR [33] algorithm, we linearly scale the scores obtained from each IQS to the same range [0, 1] for normalization. The linear scaling process is conducted on both training and test data [34] via

$$y = \frac{x - \min(X)}{\max(X) - \min(X)} \tag{33}$$

where y is the scaled score, x is the raw score, and $\max(X)$, and $\min(X)$ specify the maximum and minimum values of the score X, respectively.

V. Scorer Selection in ParaBoost IQA System

To reach a balance between accurate quality evaluation and low computational complexity of the ParaBoost IQA system, it is desirable to develop a process that can add IQSs gradually and systematically. In this section, we propose two scorer selection methods using statistical testing [35] that can select one IQS at a time and add it to the ParaBoost system. Among these statistical testing methods, F-score (i.e., analysis of variance (ANOVA)) and Kruskal–Wallis (K–W) test have been used for feature selection [36]–[38]. Here, we use them in Method 1 and Method 2, respectively, to select IQSs.

The first method is based on the one-way ANOVA (1-way ANOVA), which is a parametric test method. The procedure is described below.

Method 1:
1-Way ANOVA

1) Divide the N scores obtained by each IQS into m groups, where each score group has $n_i (i = 1, \ldots, m)$ corresponding images and $N = \sum_{i=1}^{m} n_i$.
2) Compute the F values for each IQS through the following.
   a) Compute the mean of each group $\bar{s}_1, \bar{s}_2, \ldots, \bar{s}_m$
   $$\bar{s}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} s_{ij}, \quad i = 1, \ldots, m$$
   where s_{ij} represents the score of the j-th image of the i-th group.
   b) Compute the sum of squared deviations for each group $SS_1, SS_2, \ldots, SS_m$
   $$SS_i = \sum_{j=1}^{n_i} (s_{ij} - \bar{s}_i)^2, \quad i = 1, \ldots, m.$$
   c) Compute the within group sum of squares
   $$SS_{\text{within}} = \sum_{i=1}^{m} SS_i.$$
   d) Compute the within group variance
   $$\sigma^2_{\text{within}} = \frac{SS_{\text{within}}}{DF_{\text{within}}} = MS_{\text{within}}$$
   where $DF_{\text{within}} = N - m$.
   e) Compute the between group sum of squares
   $$SS_{\text{between}} = \sum_{i=1}^{m} n_i \bar{s}_i^2 - \frac{1}{N} \left(\sum_{i=1}^{m} n_i \bar{s}_i \right)^2.$$
   f) Compute the between group variance
   $$\sigma^2_{\text{between}} = \frac{SS_{\text{between}}}{DF_{\text{between}}} = MS_{\text{between}}$$
   where $DF_{\text{between}} = m - 1$.
   g) Then, the F statistic value is
   $$F = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}}. \tag{34}$$
3) Rank IQSs in the descending order of their F statistic values, and the ParaBoost system can select them one by one according to this order to achieve the best tradeoff between performance and complexity.

The second selection method is to use the K–W statistic, which is a nonparametric statistical test method and can be applied when the score data are not normally distributed. It is described below.

Method 2:
K–W Statistic

1) Divide the N scores obtained by each IQS into m groups and each score group has $n_i (i = 1, \ldots, m)$ corresponding images, where $N = \sum_{i=1}^{m} n_i$.
2) Rank all scores in the ascending order regardless of their group.

3) Compute the K–W statistic value H for each IQS using the following.
   a) Compute the average rank for score group i via
$$\bar{R}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} R_{ij}, \quad i = 1, \ldots, m$$
   where R_{ij} represents the rank of the j-th image of the i-th group.
   b) Compute the average rank for all images via
$$\bar{R} = \frac{1+2+\cdots+N}{N} = \frac{N+1}{2}.$$
   c) Then, the K–W statistic value can be determined via
$$H = \frac{12}{N(N+1)}\sum_{i=1}^{m} n_i(\bar{R}_i - \bar{R})^2. \quad (35)$$
4) Rank IQSs in the descending order by H values and then select them one by one for inclusion in the ParaBoost IQA system according to this order.

In the above two methods, N scores are divided into m groups, where the scores placed in the same group have similar corresponding MOSs. In general, the value of m is between 10 and 20 to make sure there are enough groups to run the statistical test. In addition, n_i is obtained after m is decided. The m groups can have different number of images (scores). That means all $n_i^{\prime}s$ $(i = 1, \ldots, m)$ can be different as long as the summation of $n_i^{\prime}s$ equals N. Also, if $F > F_{\text{crit}}$ (or $H > H_{\text{crit}}$), we reject the null hypothesis

$$H_0 : \{\text{No significant difference on the IQS among score groups}\}$$

with probability $P < \alpha$, where α is the significance level. Typically, α is set to 0.05 or 0.01. From the statistics point of view, the F (or H) value is a ratio of the variability between groups compared with the variability within the groups. If this ratio is large, then there is a higher probability to reject the null hypothesis, which also means that there is a significant difference among score groups. As a result, the IQS with a larger F (or H) value has a higher discriminating power and should be chosen first.

The F and H values and the corresponding rank of each scorer are listed in Table IX. From Tables IX and X, we see that there are some differences for the order of the selected IQSs between these two methods. For example, the seventh selected IQS is AIQS #4 by using Method 1, but BIQS #1 by using Method 2. This is the point when the performance of the ParaBoost system decided by Method 1 is always better than that decided by Method 2 under the same number of IQSs, as shown in Table X.

In addition, the scorer selection algorithms treat both AIQS and BIQS in the same way. In other words, we start the IQS selection from the empty set in order to achieve the best performance with minimal number of scorers. Performing in this way can make sure that the minimal number of IQSs is used (i.e., there may be no need to use all BIQSs if possible). And the results in Table XIII show that all BIQSs are still selected as the final IQSs, except BIQS #1 in some cases. Therefore, even if we do not impose any assumption for using BIQSs as the basis, the final selected IQSs still include all BIQSs for most of the cases.

TABLE IX
1-WAY ANOVA AND K–W STATISTIC FOR EACH IQS IN TID2008 DATABASE

IQS	1-way ANOVA		K-W statistic	
	F value	Rank	H value	Rank
BIQS #1	2.4957	17	417.2	7
BIQS #2	133.5516	4	1101.8	3
BIQS #3	176.0039	3	1009.4	4
BIQS #4	263.0011	2	1116.7	2
BIQS #5	270.1121	1	1209.5	1
BIQS #6	40.7945	5	488.2	5
AIQS #1	26.8595	8	303.8	8
AIQS #2	8.8806	13	147.6	12
AIQS #3	15.4905	10	224.4	11
AIQS #4	26.8884	7	267.9	9
AIQS #5	27.5269	6	474.6	6
AIQS #6	10.8617	12	141.4	13
AIQS #7	6.4441	14	90.9	15
AIQS #8	12.9817	11	119.7	14
AIQS #9	19.4745	9	259.5	10
AIQS #10	2.8555	16	48.9	16
AIQS #11	3.2488	15	31.3	17

Also, as shown in Table X, we only need 16 IQSs to achieve the best performance using Method 1. However, 17 IQSs are needed to achieve the same performance if we adopt Method 2. Thus, Method 1 is a better choice, which is probably due to the fact that most of the scorer outputs follow the normal distribution.

To verify this conjecture, we calculate the kurtosis of the score distribution for each IQS, and summarize their values in Table XI. Based on [39], the scores of each IQS are considered to be normally distributed if the kurtosis value is between 2 and 4. Although BIQS #1, AIQS #8, and AIQS #11 have kurtosis values slightly greater than 4, they are still close to the normal distribution. From Table XI, we see that 13 out of 17 IQSs are normally distributed, which explains why Method 1 works better for TID2008 database. Similarly, Method 2 is a better choice when most of the IQSs are not normally distributed.

VI. EXPERIMENTAL RESULTS

A. Image Quality Databases

We evaluate the performance of the proposed ParaBoost method with three commonly used image quality databases (namely, TID2008, LIVE, and CSIQ) as well as a new database known as TID2013. They are briefly described below.

The Tampere Image Database (TID2008) [5], [20] includes 25 reference images, 17 distortion types for each reference image, and four levels for each distortion type. The database contains 1700 distortion images, and the MOS provided in this database ranges from 0 to 9.

The LIVE Image Quality Database [4] has 29 reference images and 779 test images, consisting of five distortion types (JPEG2000, JPEG, white noise in the RGB components, Gaussian blur, and transmission errors in the JPEG2000 bit stream using a fast-fading Rayleigh channel model).

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE X
PERFORMANCE OF SELECTED IQSS BY METHODS 1 AND 2 IN TID2008 DATABASE

Method 1 (1-way ANOVA)				Method 2 (K-W statistic)			
Selected IQS	PCC	SROCC	RMSE	Selected IQS	PCC	SROCC	RMSE
BIQS #5	0.8246	0.8353	0.7591	BIQS #5	0.8246	0.8353	0.7591
BIQS #5,4	0.8685	0.8678	0.6651	BIQS #5,4	0.8685	0.8678	0.6651
BIQS #5,4,3	0.8728	0.8705	0.6549	BIQS #5,4,2	0.8749	0.8718	0.6499
BIQS #5,4,3,2	0.8788	0.8752	0.6404	BIQS #5,4,2,3	0.8788	0.8752	0.6404
BIQS #5,4,3,2,6	0.9022	0.8995	0.5787	BIQS #5,4,2,3,6	0.9022	0.8995	0.5787
BIQS #5,4,3,2,6 AIQS #5	0.9038	0.9012	0.5742	BIQS #5,4,2,3,6 AIQS #5	0.9038	0.9012	0.5742
BIQS #5,4,3,2,6 AIQS #5,4	0.9117	0.9120	0.5513	BIQS #5,4,2,3,6,1 AIQS #5	0.9109	0.9085	0.5538
BIQS #5,4,3,2,6 AIQS #5,4,1	0.9352	0.9316	0.4752	BIQS #5,4,2,3,6,1 AIQS #5,1	0.9320	0.9254	0.4865
BIQS #5,4,3,2,6 AIQS #5,4,1,9	0.9545	0.9544	0.4003	BIQS #5,4,2,3,6,1 AIQS #5,1,4	0.9412	0.9384	0.4535
BIQS #5,4,3,2,6 AIQS #5,4,1,9,3	0.9584	0.9580	0.3832	BIQS #5,4,2,3,6,1 AIQS #5,1,4,9	0.9555	0.9554	0.3958
BIQS #5,4,3,2,6 AIQS #5,4,1,9,3,8	0.9667	0.9658	0.3436	BIQS #5,4,2,3,6,1 AIQS #5,1,4,9,3	0.9593	0.9591	0.3791
BIQS #5,4,3,2,6 AIQS #5,4,1,9,3,8,6	0.9698	0.9690	0.3271	BIQS #5,4,2,3,6,1 AIQS #5,1,4,9,3,2	0.9642	0.9646	0.3560
BIQS #5,4,3,2,6 AIQS #5,4,1,9,3,8,6,2	0.9714	0.9721	0.3187	BIQS #5,4,2,3,6,1 AIQS #5,1,4,9,3,2,6	0.9683	0.9695	0.3354
BIQS #5,4,3,2,6 AIQS #5,4,1,9,3,8,6,2,7	0.9728	0.9730	0.3107	BIQS #5,4,2,3,6,1 AIQS #5,1,4,9,3,2,6,8	0.9713	0.9719	0.3193
BIQS #5,4,3,2,6 AIQS #5,4,1,9,3,8,6,2,7,11	0.9747	0.9752	0.2998	BIQS #5,4,2,3,6,1 AIQS #5,1,4,9,3,2,6,8,7	0.9726	0.9726	0.3119
BIQS #5,4,3,2,6 AIQS #5,4,1,9,3,8,6,2,7,11,10	0.9767	0.9772	0.2879	BIQS #5,4,2,3,6,1 AIQS #5,1,4,9,3,2,6,8,7,10	0.9748	0.9747	0.2993
BIQS #5,4,3,2,6,1 AIQS #5,4,1,9,3,8,6,2,7,11,10	0.9767	0.9772	0.2880	BIQS #5,4,2,3,6,1 AIQS #5,1,4,9,3,2,6,8,7,10,11	0.9767	0.9772	0.2880

TABLE XI
NORMAL DISTRIBUTION TEST FOR EACH IQS IN TID2008 DATABASE

IQS	Kurtosis	Normal distribution
BIQS #1	4.0066	Yes
BIQS #2	5.4900	No
BIQS #3	3.8265	Yes
BIQS #4	7.3002	No
BIQS #5	2.9064	Yes
BIQS #6	3.1276	Yes
AIQS #1	2.6999	Yes
AIQS #2	2.5784	Yes
AIQS #3	2.2457	Yes
AIQS #4	20.8384	No
AIQS #5	7.0692	No
AIQS #6	2.7488	Yes
AIQS #7	2.4628	Yes
AIQS #8	4.2552	Yes
AIQS #9	2.5750	Yes
AIQS #10	3.1408	Yes
AIQS #11	4.1181	Yes

The subjective quality scores provided in this database are DMOS, ranging from 0 to 100.

The Categorical Image Quality (CSIQ) Database [6] contains 30 reference images, and each image contains six distortion types (JPEG compression, JPEG2000 compression, global contrast decrements, additive Gaussian white noise, additive Gaussian pink noise, and Gaussian blurring) at 4 to 5 different levels, resulting in 866 distorted images. The score ratings (from 0 to 1) are reported in DMOS.

Besides the 17 distortion types in TID2008, the Tampere Image Database 2013 (TID2013) [7], [40] introduces seven new distortion types. They are: 1) change of color saturation (#18); 2) multiplicative Gaussian noise (#19); 3) comfort noise (#20); 4) lossy compression of noisy images (#21); 5) image color quantization with dither (#22); 6) chromatic aberrations (#23); and 7) sparse sampling and reconstruction (#24). Consequently, TID2013 has the richest diversity, consisting of 25 reference images, 24 distortion types for each reference, and five levels for each distortion type. The database contains 3000 distorted images with their subjective scores evaluated in MOS ranging from 0 to 9.

B. Performance Measures for IQA Methods

We use three indices to measure the performance of IQA methods [41], [42]. The first one is the Pearson correlation coefficient (PCC) between the objective and the subjective scores. It is used to evaluate prediction accuracy. The second one is the SROCC between the objective and the subjective scores. It is used to evaluate the prediction monotonicity. The third one is the RMSE between the objective and the subjective scores.

In order to compute PCC and SROCC, we use the following monotonic logistic function [41] and the procedure described in [41] to fit the objective scores to the subjective quality scores (MOS or DMOS):

$$f(x) = \frac{\beta_1 - \beta_2}{1 + \exp^{-\left(\frac{x-\beta_3}{|\beta_4|}\right)}} + \beta_2 \tag{36}$$

where x is the predicted objective score, $f(x)$ is the fitted objective score, and parameters β_j, $j = 1, 2, 3, 4$, are chosen

TABLE XII
PERFORMANCE OF EACH IQS IN THE OTHER THREE DATABASES

IQS	LIVE			CSIQ			TID2013		
	PCC	SROCC	RMSE	PCC	SROCC	RMSE	PCC	SROCC	RMSE
BIQS #1	0.8327	0.8406	15.1277	0.4340	0.4229	0.2365	-0.5674	-0.6757	1.2397
BIQS #2	0.8929	0.8973	12.3046	0.8146	0.8369	0.1523	0.7570	0.6952	0.8100
BIQS #3	0.9363	0.9385	9.5967	0.8609	0.7809	0.1336	0.7355	0.6622	0.8399
BIQS #4	0.8839	0.8944	12.7795	0.8143	0.7657	0.1524	0.7263	0.5492	0.8521
BIQS #5	0.9503	0.9551	8.5092	0.9071	0.9232	0.1105	0.8117	0.7533	0.7240
BIQS #6	0.8700	0.8726	13.4697	0.7777	0.7850	0.1650	0.6377	0.6229	0.9549
AIQS #1	0.7353	0.7494	18.5185	0.4461	0.2006	0.2350	-0.0229	-0.1257	1.2397
AIQS #2	0.5389	0.4170	23.0150	0.4679	0.3001	0.2320	-0.3387	-0.3870	1.2397
AIQS #3	0.9552	0.9525	8.0867	0.6858	0.5229	0.1911	-0.4936	-0.5373	1.2397
AIQS #4	0.5514	0.5078	22.7924	0.6062	0.5477	0.2088	0.3896	0.2535	1.1417
AIQS #5	0.5945	0.6809	21.9706	0.5985	0.5652	0.2103	0.4255	0.2015	1.1218
AIQS #6	0.1796	0.3194	26.8780	0.3788	0.3264	0.2430	-0.4893	-0.5105	1.2397
AIQS #7	0.4130	0.3628	24.8828	0.1880	0.1771	0.2578	-0.0655	-0.1594	1.2397
AIQS #8	0.5561	0.5202	22.7084	0.3356	0.2933	0.2473	0.0575	0.0645	1.2376
AIQS #9	0.8467	0.8493	14.5389	0.7811	0.7761	0.1639	0.5066	0.3424	1.0688
AIQS #10	0.3894	0.1076	25.1651	0.2259	0.3475	0.2557	-0.4517	-0.5079	1.2397
AIQS #11	0.3452	0.1947	25.6430	0.2923	0.1759	0.2511	-0.3955	-0.4894	1.2397

TABLE XIII
PERFORMANCE COMPARISONS FOR DIFFERENT COMBINATIONS OF IQSs

Database	IQS Selection Method	Selected IQSs	Total no. of IQSs	PCC	SROCC	RMSE
TID2008	All BIQSs	All BIQSs	6	0.9093	0.9068	0.5583
	All AIQSs	All AIQSs	11	0.9331	0.9334	0.4827
	All IQSs	All IQSs	17	0.9767	0.9772	0.2880
	Method 1	BIQS #5,4,3,2,6 AIQS #5,4,1,9,3,8,6,2,7,11,10	16	0.9767	0.9772	0.2879
	Method 2	BIQS #5,4,2,3,6,1 AIQS #5,1,4,9,3,2,6,8,7,10	16	0.9748	0.9747	0.2993
LIVE	All BIQSs	All BIQSs	6	0.9693	0.9693	6.7202
	All AIQSs	All AIQSs	11	0.9782	0.9789	5.6802
	All IQSs	All IQSs	17	0.9736	0.9757	6.2382
	Method 1	BIQS #5,3,2,6,4,1 AIQS #3,9,5	9	0.9820	0.9819	5.1556
	Method 2	BIQS #5,3,2,4,6,1 AIQS #3,9,4	9	0.9835	0.9837	4.9377
CSIQ	All BIQSs	All BIQSs	6	0.9590	0.9599	0.0744
	All AIQSs	All AIQSs	11	0.9579	0.9521	0.0753
	All IQSs	All IQSs	17	0.9740	0.9708	0.0595
	Method 1	BIQS #5,3,4,2,6 AIQS #3,5,4,9,2,1,6,10	13	0.9766	0.9733	0.0565
	Method 2	BIQS #5,2,3,4,6,1 AIQS #9,3,4,5,1,2,6	13	0.9766	0.9731	0.0564
TID2013	All BIQSs	All BIQSs	6	0.8565	0.8251	0.6398
	All AIQSs	All AIQSs	11	0.8920	0.8871	0.5604
	All IQSs	All IQSs	17	0.9567	0.9575	0.3610
	Method 1	BIQS #5,4,2,3,6,1 AIQS #9,5,4,8,7,1,11,2,3,10,6	17	0.9567	0.9575	0.3610
	Method 2	BIQS #5,2,4,3,6,1 AIQS #4,5,9,8,1,7,2,11,3,10,6	17	0.9567	0.9575	0.3610

to minimize the least-squares error between the subjective score and the fitted objective score. Initial estimates of the parameters are chosen based on the recommendation in [41].

C. Performance Comparison

To evaluate the performance of each individual IQS, we show three performance indices for both BIQSs and AIQSs against three databases (LIVE, CSIQ, and TID2013) in Table XII. Similar to what was shown in Table III, most BIQSs (except BIQS #1) have better correlation performance with MOS (DMOS) and AIQSs do not perform well when working alone. The only exceptions are AIQS #3 and AIQS #9. The former has excellent performance for the LIVE database, while the latter performs well for both LIVE and CSIQ.

With the two selection methods described in Section V, we show the smallest set of IQSs that achieves the best performance in Table XIII. As indicated in the table, the numbers of IQSs needed for the ParaBoost system in TID2008 and TID2013 are 16 and 17, and these two databases have over 17 and 24 distortion types, respectively. Due to the large variety of distortion types, they cannot be easily and correctly evaluated by using a small number of scorers. On the other hand, we can provide good quality assessment for LIVE and CSIQ with fewer IQSs (9 and 13, respectively) since there are only 5–6 distortion types in these two databases.

The ParaBoost performance of the selected IQSs is listed in Table XIII, which includes several benchmarking cases with a different combination of BIQSs or AIQSs. We observe that the fusion of all IQSs does not offer the best performance

TABLE XIV
PERFORMANCE COMPARISON AMONG 13 IQA MODELS IN FOUR DATABASES

IQA model	TID2008 (1700 images)			LIVE (779 images)			CSIQ (866 images)			TID2013 (3000 images)		
	PCC	SROCC	RMSE	PCC	SROCC	RMSE	PCC	SROCC	RMSE	PCC	SROCC	RMSE
PSNR	0.5355	0.5245	1.1333	0.8701	0.8756	13.4685	0.8001	0.8057	0.1576	0.6727	0.6394	0.9172
VSNR	0.6820	0.7046	0.9815	0.9235	0.9279	10.4816	0.8005	0.8108	0.1573	0.7123	0.6816	0.8699
VIF	0.8055	0.7496	0.7953	0.9597	0.9636	7.6737	0.9253	0.9194	0.0996	0.7711	0.6769	0.7894
SSIM	0.7715	0.7749	0.8537	0.9384	0.9479	9.4439	0.8594	0.8755	0.1342	0.7893	0.7417	0.7612
MS-SSIM	0.8389	0.8528	0.7303	0.9402	0.9521	9.3038	0.8666	0.8774	0.1310	0.8130	0.7698	0.7217
IW-SSIM	0.8488	0.8559	0.7094	0.9425	0.9567	9.1301	0.9025	0.9212	0.1131	0.8296	0.7779	0.6922
FSIM	0.8710	0.8805	0.6592	0.9540	0.9634	8.1938	0.9095	0.9242	0.1091	0.8560	0.8015	0.6408
FSIMc	0.8730	0.8840	0.6545	0.9545	0.9645	8.1501	0.9176	0.9309	0.1044	0.8744	0.8510	0.6016
MAD	0.8306	0.8340	0.7474	0.9672	0.9669	6.9419	0.9502	0.9466	0.0818	0.8221	0.7807	0.7058
CF-MMF (3 methods)	0.9307	0.9275	0.4907	0.9734	0.9732	6.2612	**0.9797**	**0.9755**	**0.0527**	0.9146	0.8973	0.5012
CD-MMF (3 methods)	**0.9517**	**0.9439**	**0.4120**	**0.9802**	**0.9805**	**5.4134**	0.9675	0.9668	0.0664	**0.9249**	**0.9143**	**0.4714**
ParaBoost (Method 1)	**0.9767**	**0.9772**	**0.2879**	**0.9820**	**0.9819**	**5.1556**	**0.9766**	**0.9733**	**0.0565**	**0.9567**	**0.9575**	**0.3610**
ParaBoost (Method 2)	**0.9748**	**0.9747**	**0.2993**	**0.9835**	**0.9837**	**4.9377**	**0.9766**	**0.9731**	**0.0564**	**0.9567**	**0.9575**	**0.3610**

while the complexity is the highest. Thus, it is essential to have a scorer selection mechanism. For example, we only need nine IQSs to give the highest PCC and SROCC for the LIVE database. By comparing Tables III, XII, and XIII, we see that the SROCC gains of the ParaBoost IQA system over the single best-performing IQS with respect to LIVE, CSIQ, TID2008, and TID2013 are 0.03, 0.05, 0.14, and 0.20, respectively. The rich diversity of the IQSs, the special training strategy, the ParaBoost structure, and the IQS selection scheme all contribute to the excellent performance of the proposed ParaBoost IQA system. The performance gain is larger if the distortion types in a given database are more diversified.

In Table XIV, we compare the performance of the proposed ParaBoost method with several state-of-the-art image quality metrics, such as VSNR [43], VIF [44], SSIM [3], MS-SSIM [45], IW-SSIM [8], FSIM [9], MAD [46], CF-MMF, and CD-MMF [17], [18]. The configuration of ParaBoost used for obtaining results in Table XIV is listed in Table XIII, including selected IQSs, and the total number of IQSs used. The top three IQA models are highlighted in bold. As shown in Table XIV, the two ParaBoost IQA methods rank the first and second in TID2008, LIVE, and TID2013. For CSIQ, they still rank the second and third. Furthermore, the proposed ParaBoost method has an impressive performance on both TID2008 and TID2013. For instance, in TID 2013, the SROCC gains are around 0.11 and 0.04 over the existing best formula-based approach (i.e., FSIMc) and learning-based method (i.e., CD-MMF), respectively.

Furthermore, to test the generality of the proposed ParaBoost method, we train the system based on one database and test it on the other three databases. The experiment results are shown in Table XV. In Table XV, the configurations (the number of used IQSs) of ParaBoost for all test databases are the same as the corresponding trained model (database). For instance, if the model is trained on TID2008, then the configurations (the number of used IQSs) of ParaBoost for all test databases (LIVE, CSIQ, and TID2013) are exactly the same as the one we used to train the model on TID2008. As shown in Table XV, the SROCC values are over 0.91 for most cases except when the system is trained on LIVE and tested on TID2013. This exception can be explained by

TABLE XV
CROSS-DATABASE SROCC PERFORMANCE OF ParaBoost SYSTEM

Test database / Model	TID2008	LIVE	CSIQ	TID2013
TID2008	-	0.9760	0.9560	0.9501
LIVE	0.9308	-	0.9601	0.8526
CSIQ	0.9528	0.9771	-	0.9169
TID2013	0.9704	0.9699	0.9630	-

TABLE XVI
COMPUTATION TIME REQUIRED FOR AN SD IMAGE (SIZE 720×480) AMONG 13 IQA MODELS

IQA Model	Time (secs/image)
PSNR	0.05
VSNR	0.40
VIF	1.47
SSIM	0.13
MS-SSIM	0.19
IW-SSIM	0.76
FSIM	0.52
FSIMc	0.52
MAD	54.69
CF-MMF (3 methods)	60.79
CD-MMF (3 methods)	56.60
ParaBoost (Method 1)	5.69
ParaBoost (Method 2)	5.94

the fact that there are 24 distortion types in TID2013, which cannot be completely covered by a training process performed on smaller distortion type sets (i.e., five image distortion types in LIVE database). We would like to point out that the SROCC values of the resulting ParaBoost IQA system is still greater than 0.91 for 11 out of the 12 cross-database evaluation cases. This shows the robustness of the proposed ParaBoost IQA methodology.

D. Computational Complexity

The computational complexity of 13 IQA models is compared in Table XVI. The measurement is in terms of the computation time required to evaluate a standard-definition (SD) image of size 720×480 by using a computer with Intel core i7 processor at 1.73 GHz. For our proposed IQA model (ParaBoost), it would take 5–6 seconds to complete the quality evaluation of an SD image. However, it would only take 0.05–1.5 seconds to finish the evaluation for the other formula-based models, except MAD. MAD needs about 55 seconds to complete the evaluation. Furthermore, compared with another learning-based model (MMF), our model (ParaBoost) has a large advantage on time saving (around 90%). This is because most of the IQSs in the proposed ParaBoost model utilize one simple feature instead of one complicated method. Although it is still more time-consuming to estimate the image quality than most of the formula-based IQA models, the quality prediction performance of ParaBoost is much better than the others. This paper concentrates on the proof of concept, and how to optimize the proposed method will be our next step.

VII. Conclusion and Future Work

A new ParaBoost IQA method has been proposed in this paper, which fuses resulting scores from multiple BIQS and AIQS. The design, training, and evaluation of each individual IQS were discussed. The ParaBoost architecture was carefully examined. Two statistical test-based methods were developed to select an optimal combination of IQSs to achieve the best performance with the minimum number of scorers. Instead of conventional weighting or voting schemes, a nonlinear SVR score fuser was adopted to combine the outputs from all selected IQSs. The experiments were conducted with respect to four well-known public databases (totally 6345 images). It was shown by experimental results that the proposed ParaBoost IQA method outperforms the existing state-of-the-art IQA models (including both formula-based and learning-based methods) with clear explanation for its success.

The extension of the ParaBoost approach to the problem of VQA is under our current investigation. One main challenge along this direction is the lack of large VQA databases. Furthermore, the design of powerful individual video quality scorers (VQSs) is still an open issue.

References

[1] B. Girod, "What's wrong with mean-squared error?" in *Digital Images and Human Vision*. Cambridge, MA, USA: MIT Press, 1993, pp. 207–220.

[2] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

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

[4] (2006). *LIVE Image Quality Assessment Database*. [Online]. Available: http://live.ece.utexas.edu/research/quality/subjective.htm

[5] (2008). *Tampere Image Database*. [Online]. Available: http://www.ponomarenko.info/tid2008.htm

[6] (2010). *Categorical Image Quality (CSIQ) Database*. [Online]. Available: http://vision.okstate.edu/csiq

[7] (2013). *Tampere Image Database 2013*. [Online]. Available: http://www.ponomarenko.info/tid2013.htm

[8] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[9] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[10] T.-J. Liu, W. Lin, and C.-C. J. Kuo, "Recent developments and future trends in visual quality assessment," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Submit Conf. (APSIPA ASC)*, Xi'an, China, Oct. 2011, pp. 1–10.

[11] T.-J. Liu, Y.-C. Lin, W. Lin, and C.-C. J. Kuo, "Visual quality assessment: Recent developments, coding applications and future trends," *APSIPA Trans. Signal Inf. Process.*, vol. 2, p. e4, Jul. 2013. [Online]. Available: http://journals.cambridge.org/article_S204877031300005X

[12] H. Luo, "A training-based no-reference image quality assessment algorithm," in *Proc. IEEE Int. Conf. Image Process.*, Singapore, Oct. 2004, pp. 2973–2976.

[13] S. Suresh, V. Babu, and N. Sundararajan, "Image quality measurement using sparse extreme learning machine classifier," in *Proc. 9th IEEE Int. Conf. Control, Autom., Robot. Vis.*, Singapore, Dec. 2006, pp. 1–6.

[14] S. Suresh, R. V. Babu, and H. J. Kim, "No-reference image quality assessment using modified extreme learning machine classifier," *Appl. Soft Comput.*, vol. 9, no. 2, pp. 541–552, 2009.

[15] M. Narwaria and W. Lin, "Objective image quality assessment based on support vector regression," *IEEE Trans. Neural Netw.*, vol. 21, no. 3, pp. 515–519, Mar. 2010.

[16] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[17] T.-J. Liu, W. Lin, and C.-C. J. Kuo, "A multi-metric fusion approach to visual quality assessment," in *Proc. 3rd IEEE Int. Workshop Quality Multimedia Exper. (QoMEX)*, Mechelen, Belgium, Sep. 2011, pp. 72–77.

[18] T.-J. Liu, W. Lin, and C.-C. J. Kuo, "Image quality assessment using multi-method fusion," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1793–1807, May 2013.

[19] L. Jin, S. Cho, T.-J. Liu, K. Egiazarian, and C.-C. J. Kuo, "Performance comparison of decision fusion strategies in BMMF-based image quality assessment," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Submit Conf. (APSIPA ASC)*, Hollywood, CA, USA, Dec. 2012, pp. 1–4.

[20] N. N. Ponomarenko, V. V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," *Adv. Modern Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.

[21] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2002, pp. I-477–I-480.

[22] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 3, no. 6, pp. 610–621, Nov. 1973.

[23] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[24] W. K. Pratt, *Digital Image Processing*, 4th ed. New York, NY, USA: Wiley, 2007.

[25] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2007.

[26] K.-H. Liu, T.-J. Liu, and H.-H. Liu, "A SIFT descriptor based method for global disparity vector estimation in multiview video coding," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Singapore, Jul. 2010, pp. 1214–1218.

[27] T.-J. Liu, K.-H. Liu, and H.-H. Liu, "Temporal information assisted video quality metric for multimedia," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Singapore, Jul. 2010, pp. 697–701.

[28] K.-H. Liu, S. Yan, and C.-C. J. Kuo, "Age estimation via grouping and decision fusion," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 11, pp. 2408–2423, Nov. 2015.

[29] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits Syst. Mag.*, vol. 6, no. 3, pp. 21–45, Sep. 2006.

[30] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

[31] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, 2001.

[32] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.

[33] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. ID 27.

[34] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, *A Practical Guide to Support Vector Classification*. Taipei, Taiwan: National Taiwan University, 2003.

[35] S. A. Glantz, *Primer of Biostatistics*, 6th ed. New York, NY, USA: McGraw-Hill, 2005.

[36] Y.-W. Chen and C.-J. Lin, "Combining SVMs with various feature selection strategies," in *Feature Extraction*. Berlin, Germany: Springer-Verlag, 2006, pp. 315–324.

[37] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, "Advancing feature selection research—ASU feature selection repository," Arizona State Univ., Tech. Rep., 2010.

[38] K.-H. Liu, S. Yan, and C.-C. J. Kuo, "Age group classification via structured fusion of uncertainty-driven shape features and selected surface features," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Steamboat Springs, CO, USA, Mar. 2014, pp. 445–452.

[39] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document Rec. ITU-R BT.500-13, International Telecommunication Union, Jan. 2012.

[40] N. Ponomarenko *et al.*, "Color image database TID2013: Peculiarities and preliminary results," in *Proc. 4th IEEE Eur. Workshop Vis. Inf. Process. (EUVIP)*, Paris, France, Jun. 2013, pp. 106–111.

[41] VQEG, *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase I*, document COM 9-80-E, Video Quality Experts Group, Mar. 2000. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI

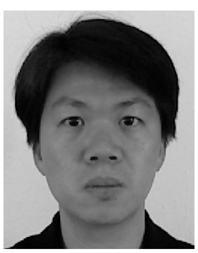
[42] VQEG, *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II*, document ITU-T SG09, Video Quality Experts Group, Aug. 2003. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII

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

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

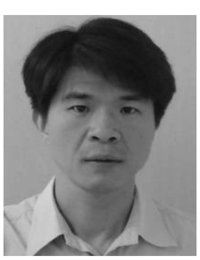
[45] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. Conf. Rec. 37th IEEE Asilomar Conf. Signals, Syst., Comput.*, Nov. 2003, pp. 1398–1402.

[46] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 011006-1–011006-21, 2010.



Tsung-Jung Liu (S'10–M'14) received the B.S. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1998, the M.S. degree in communication engineering from National Taiwan University, Taipei, Taiwan, in 2001, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2014.

He was a Postdoctoral Fellow with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, from October 2014 to January 2015. He is currently an Assistant Professor with the Graduate Institute of Communication Engineering, and the Department of Electrical Engineering, National Chung Hsing University, Taichung, Taiwan. His current research interests include machine learning, perceptual image/video processing, visual quality assessment, and big data analytics.

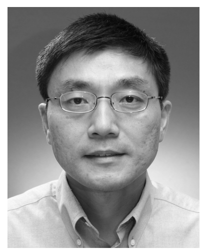


Kuan-Hsien Liu (S'10–M'14) received the B.S. degree in electrical engineering from National Central University, Zhongli, Taiwan, in 1999, the M.S. degree in communication engineering from National Taiwan University, Taipei, Taiwan, in 2001, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2014.

He is currently a Postdoctoral Fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei. His current research interests include digital image analysis and processing, computer vision, machine learning, and big data analytics.

Joe Yuchieh Lin received the B.S. and M.S. degrees in electrical engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2004 and 2006, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2015.

He is currently with Amazon.com, Inc., Seattle, WA, USA. His current research interests include video/image quality assessment, machine learning, and multimedia coding.



Weisi Lin (M'92–SM'98) received the Ph.D. degree from King's College London, London, U.K.

He served as the Laboratory Head of Visual Processing with the Institute for Infocomm Research, Singapore. He is currently an Associate Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. His current research interests include image processing, perceptual signal modeling, video compression, and multimedia communication, in which he has published over 130 journal papers and over 200 conference papers, holds seven patents, and has authored two books.

Prof. Lin is an Associate Editor (AE) of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE SIGNAL PROCESSING LETTERS, and the *Journal of Visual Communication and Image Representation*. He was an AE of the IEEE TRANSACTIONS ON MULTIMEDIA. He has served as a Guest Editor for eight special issues in international journals. He was a Technical Program Chair of the IEEE International Conference on Multimedia & Expo in 2013, the Pacific-Rim Conference on Multimedia in 2012, and the International Workshop on Quality of Multimedia Experience in 2014. He has been an Invited/Panelist/Keynote/Tutorial Speaker in more than ten international conferences. He was a Distinguished Lecturer of the Asia-Pacific Signal and Information Processing Association from 2012 to 2013.

C.-C. Jay Kuo (F'99) received the B.S. degree from National Taiwan University, Taipei, Taiwan, in 1980, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1985 and 1987, respectively, all in electrical engineering.

He is currently the Director of the Media Communications Laboratory and the Dean's Professor of Electrical Engineering with the University of Southern California, Los Angeles, CA, USA. He has co-authored about 230 journal papers, 870 conference papers, and 13 books. His current research interests include digital image/video analysis, multimedia data compression, computer vision, and machine learning.

Prof. Kuo is a fellow of the American Association for the Advancement of Science and the International Society for Optical Engineers.