Multiscale Image Quality Estimation

A. Murat Demirtas, *Student Member, IEEE,* Amy R. Reibman, *Fellow, IEEE,* Hamid Jafarkhani, *Fellow, IEEE*

Abstract

Multimedia communication is becoming pervasive because of the progress in wireless communications and multimedia coding. Estimating the quality of the visual content accurately is crucial in providing satisfactory service. State of the art visual quality assessment approaches are effective when the input image and the reference image have the same resolution. However, finding the quality of an image that has spatial resolution different than that of the reference image is still a challenging problem. To solve this problem, we develop a quality estimator (QE) which computes the quality of the input image without resampling the reference or the input images.

In this work, we begin by identifying the potential weaknesses of previous approaches used to estimate the quality of experience. Next, we design a QE to estimate the quality of a distorted image with a lower resolution compared to the reference image. We also propose a subjective test environment to explore the success of the proposed algorithm in comparison with other QEs. When the input and test images have different resolutions, the subjective tests demonstrate that in most cases the proposed method works better than other approaches. In addition, the proposed algorithm also performs well when the reference image and the test image have the same resolution.

Index Terms

image quality estimation, human visual system, paired comparison, spatial resolution, subjective tests

I. Introduction

In our daily life, we often view a visual content using a display which has different resolution specifications than the original content's resolution. To display the image, the content may be resized. For example, small displays decimate the image that is captured with a high resolution camera. A video is resampled if it is produced for HDTV but it is watched using a smart phone or an SDTV. On the other hand, the visual content which is created for a low resolution display needs to be interpolated to fill the screen. Hence, the effect of device's resolution to the perceived quality is very important when transmitting a visual content optimally over heterogeneous communications networks.

Quality estimators (QEs) are used to measure the perceived quality of the visual content. Having an accurate QE is helpful to define system requirements. This enables designers to have a consistent method to evaluate system performance. It also provides a benchmark, enabling a comparison among different image and video processing algorithms. Moreover, a QE can also be exploited to solve a rate-distortion optimization problem related to visual communications or storage [1].

Existing objective quality estimators are currently inadequate to accurately compare quality among two images or videos with different spatial or temporal resolutions. One approach is to use a no-reference (NR) QE [2]. An NR-QE does not use

A. Murat Demirtas and Hamid Jafarkhani are with Center for Pervasive Communications and Computing, University of Calfornia Irvine, Irvine, CA 92697-2700 USA e-mail: ademirta, hamidj@uci.edu

Amy R. Reibman is with AT&T Labs-Research, Florham Park, NJ. USA e-mail: amy@research.att.com

a reference image to estimate the quality. Instead, it may use the characteristics of possible artifacts like blocking or noise to estimate the quality, or it may use a model of the desired signal to identify deviations in the image from the model [3]. However, when a reference image is available, a no-reference approach to quality estimation will sacrifice valuable information. Some of the recent NR-QE studies can be found in [4], [5] and [6].

On the other hand, full-reference (FR) QEs exploit the reference image to find the quality. The simplest FR approach is mean square error (MSE). However as observed in [7] MSE-based approaches do not correlate well with subjective test results. In [8], [9], [10], [11], [12], recent studies about quality estimators are described. These QEs generally perform better than MSE because they also take into account the characteristics of the human visual system (HVS), either using low-level models or applying overarching principles. Currently, FR metrics are designed to calculate the QE of an input image (or video) if it has the same spatial (and temporal) resolution as the reference. Therefore, it is necessary to resize either the reference image or the input image prior to applying one of these QEs if the images have different resolutions. Both these options have drawbacks, as discussed in Section II.

Several studies have been performed to understand the effect of resolution on the quality. These studies focus on providing the best quality video transmission under the constraints of available bandwidth, and the resolution of the viewer's display. Reed and Lim [13] propose an algorithmic method to determine the best trade-off between spatial resolution, temporal resolution, and encoding quantization parameters. They optimize *objective* quality (measured by sum of absolute errors (SAE)) at the time of initial encoding by jointly adapting frame rate, spatial resolution, and quantization step size. Akyol et al. [2] present a system to choose the best settings for a scalable encoder for each temporal segment. The spatial resolution, temporal resolution, and quantization parameters are all chosen based on the content type in the given temporal segment. Quality is measured using a no-reference objective measure that quantifies flatness, blockiness, blurriness, and temporal jerkiness. Wang et al. [7] examine the *subjective* impact of jointly adjusting spatial resolution, temporal resolution, and quantization step-size. Smaller images were viewed with reduced resolution rather than being upsampled to the larger resolution. They show that people prefer a smaller image with smaller quantization errors compared to a larger image with larger quantization errors, for the same bit-rate. In addition, they also show that both peak signal to ratio (PSNR) and SAE do not correlate well with the subjective assessment of their videos.

Additional subjective studies of the trade-off among spatial resolution, temporal resolution, and quantization step-size include [14], [15], [16], [17]. Bae et al. [14] explore the preferred spatial resolution for a given level of quantization error, and demonstrate that (a) people prefer to see larger images with little visible quantization error rather than a smaller image which has no visible quantization errors, and (b) for more than some quantization errors, the amount of acceptable distortion increases as the spatial resolution decreases. Refs. [15] and [16] explore subjective video quality on mobile platforms when spatial resolution, temporal resolution, and quantization parameters are varied. Their results indicate that as the spatial resolution is decreased, the amount of decrease in subjective quality depends on the video content and on the quantization parameter. Cermak et al. [17] evaluated the test results of 2 VQEG projects. They used the mean opinion scores (MOS)s obtained for QIF(176×144), CIF(352×288), VGA(640×480), and HD(1920×1200) resolutions at many bit rates. Their results show the required bit rate to achieve a given level of video quality for a given screen resolution. These studies provide valuable

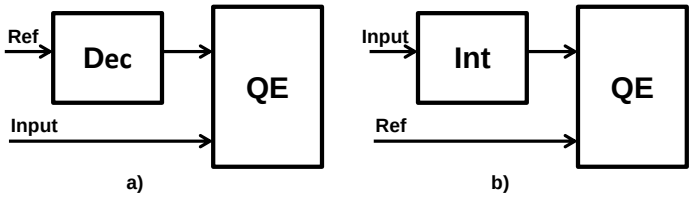

Fig. 1. Conventional Approaches a)Input image vs decimated reference image b)Interpolated input image vs reference image

information to understand the effect of spatial resolution, but they do not provide an objective metric that is used to quantify this effect.

In this work, we design a QE to estimate the quality of the corrupted displayed image with a lower resolution compared to the reference image. We also propose a subjective test environment where we can evaluate the success of the proposed algorithm in comparison with other QEs. When the reference and test images have different resolutions, the subjective tests demonstrate that in most cases the proposed method works better than other approaches. In addition, the proposed algorithm performs well when the reference image and the test image have the same resolution. Section II explores the potential weaknesses of previous approaches that are used to estimate the quality in our problem. In Section III, we describe our proposed QE approach, called Multiscale Image Quality Estimator (MIQE) and its colored version. In Section IV, we describe the subjective test environment and analyze the results of the subjective tests. Section V concludes the paper.

II. Limitations of the Previous Approaches

To date, QEs have been designed to estimate quality when the displayed image has the same number of pixels, or spatial resolution, as the reference image. However, viewers use a variety of devices, including TVs and smart phones, to watch movies and perform video conferencing. These devices have different spatial resolutions and are viewed from different distances. To adapt existing QEs to estimate quality when the displayed image has a different spatial resolution, there are two straightforward methods to force the number of pixels to be equal. These are illustrated in Figure 1 for the case where the image being considered has lower resolution than the reference image. (The extension to the opposite case is straightforward.) These two approaches are:

1) Compare the low-resolution image with the decimated reference image. (See Fig. 1(a).)
2) Compare the interpolated low-resolution image with the reference image. (See Fig. 1(b).)

Throughout this paper, we call these methods QE_{down} and QE_{up}, respectively. QE_{down} is used in Refs. [14] and [18] , and QE_{up} is used in Refs. [13] and [2] to estimate the quality. Despite their ease of use, QE_{down} and QE_{up} have significant limitations [19]. In this section, we present a motivating example that demonstrates these two approaches fail to accurately estimate subjective quality, then describe their drawbacks and limitations. We begin with some definitions.

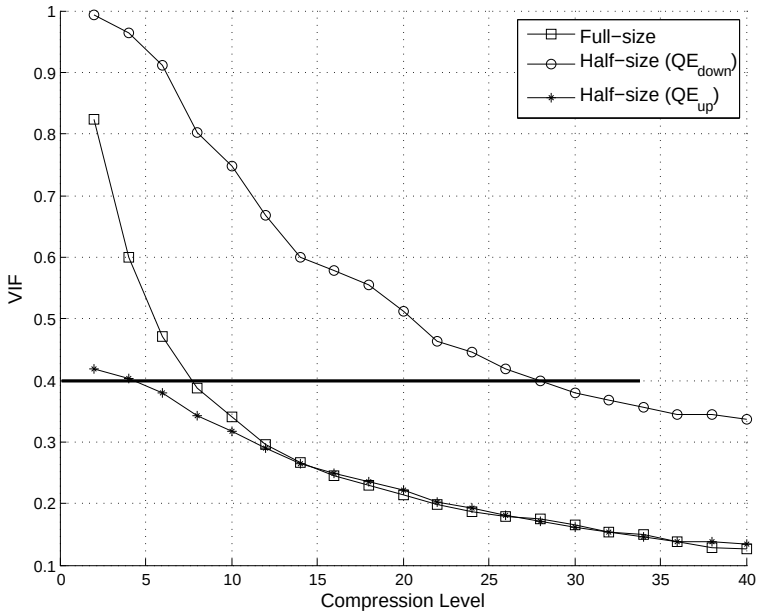

Fig. 2. Calculated VIF values of the test images for each compression level

A. Definitions

To mathematically compare two images that are viewed with different resolutions and/or different distances, we describe the images using an angular frequency representation. Specifically, we compute the angular frequency using [20]:

$$f(l) = \frac{\pi * d * n}{180 * h * 2 * 2^l} \tag{1}$$

In this expression, $f(l)$ denotes the angular frequency in cycles per degree (cyc/deg); d, h, and n represent the distance of the viewer, height of the screen, and the number of pixels in the vertical direction, respectively. In addition, l indicates the level of a subband decomposition. When there is no subband decomposition, we set $l = 1$.

Let the term *input image* describe the low-resolution degraded image, $\mathbf{i}$, that is input to either QE$_{down}$ or QE$_{up}$. Further, let $\mathbf{r}$ be the *reference image*. We reserve the term *test image* to denote those images that are shown to viewers during a subjective test. The decimated reference image, $\mathbf{d}_r$, is input to the QE-block in Fig. 1(a), while the interpolated input image, $\mathbf{u}_i$, is input to the QE-block in Fig. 1(b).

B. Motivating example using QE_{up} and QE_{down}

Our motivating example, from [19], uses the Visual Information Fidelity (VIF) [9] to compare image compression using JPEG 2000 with and without downsizing the image prior to compression. VIF has been demonstrated to have good performance in a variety of applications. Figure 2 shows quality estimated using VIF, as a function of compression level, for the 512×512 "log-seaside" image from the CSIQ database [21], when JPEG 2000 is applied to either a full-sized or half-sized image. The curve labeled *full-size* is the conventional VIF applied to the full-size compressed image. Two methods of computing VIF for the half-size compressed image are also shown. The curve labeled "Half-size (QE_{down})" is VIF computed using QE_{down}, and the curve labeled "Half-size (QE_{up})" is VIF computed using QE_{up}. To create a fair comparison, the compression level

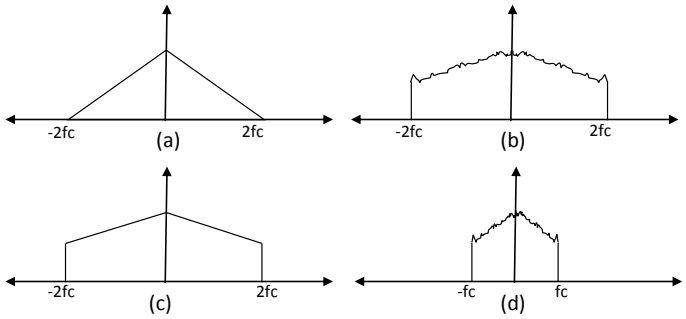

Fig. 3. 2D Projected Frequency Domain Representations of Image Models (a) Reference image (b) Input Image (c) Decimated Reference Image (d) Interpolated Input image. The downsampling rate is 2

for the half-sized images incorporates the impact of downsampling.

To explore whether VIF is accurate or not for this problem, we choose a VIF value of 0.4, corresponding to the dashed line in Figure 2. A visual inspection of these three images indicates that the half-size image computed using QE_{up} is substantially better than the full-size image, while the half-size image computed using QE_{down} is worse than both. Therefore, these three images have different visual quality although they have the same estimated quality. More details, including the three images, can be found in [19].

To understand the limitations of QE$_{down}$ and QE$_{up}$ better, we next describe the impact of these methods in the frequency domain.

C. Frequency domain comparison of QE$_{down}$ and QE$_{up}$

To estimate image quality objectively, $\mathbf{r}$ and $\mathbf{i}$ should be compared at the same visual angle. Our motivating example indicates that both QE_{up} and QE_{down} are not accurate. To understand their limitations better, we describe the effects of these methods pictorially, using simplified 1D frequency domain representations.

Figure 3a shows a sketch of the frequency representation of a *reference image*, $\mathbf{r}$, where the horizontal axis indicates frequency in cycles per degree and the vertical axis shows the magnitude of the frequency content of the signal. While we do not know how the *input image*, $\mathbf{i}$, is generated from the *reference image*, it is reasonable to assume that is obtained by low-pass filtering (LPF), downsampling (DS) , and some additional degradation. Figure 3b illustrates the frequency representation of such an *input image*. In this subsection, upsampling and downsampling operations rates are 2.

First, we consider the method of QE_{down}, where we compare a decimated reference image at downsampling rate 2, $\mathbf{d}_r$ with $\mathbf{i}$ as in Figure 1a. During QE_{down}, LPF and DS operations are performed consecutively. The resulting decimated reference image in the frequency domain is shown pictorially in Figure 3c, where the x-axis is in cycles/degree and y-axis illustrates the magnitude. The QE block of QE_{down} in Figure 1a compares the images represented in Figures 3c and 3b.

Similarly, the method QE_{up} estimates quality of an input image that is smaller than the reference image by comparing an interpolated $\mathbf{i}$ with the reference image $\mathbf{r}$. During QE_{up}, we upsample (US) $\mathbf{i}$ and perform LPF. The frequency domain representation of the interpolated *input image*, $\mathbf{u}_i$, is shown pictorially in Figure 3d. The QE block of QE_{up} in Figure 1b compares the images represented in Figures 3a and 3d.

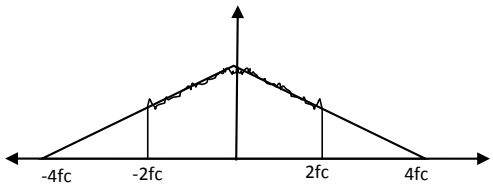

Fig. 4. Pictorial representation of the reference image and the input image that have the same visual angle in the frequency domain

Both QE_{down} and QE_{up} have several drawbacks. In QE_{down}, high-frequency content is lost from the reference image prior to comparison. On the other hand, in QE_{up}, the input image is further degraded before the comparison due to the LPF. In the former case, quality will be estimated as better than the actual quality, because the reference has been degraded. In the latter case, quality will be estimated as worse than the actual quality, because the input image has been further degraded.

A second drawback arises due to potential mismatch between the filters. Since it is not known what filter has been used to create the actual input image $\mathbf{i}$, it is not possible to ensure the same filter is used in QE_{down} to create $\mathbf{d}_r$. Difficulties arise not only due to mismatch between the magnitude responses of the LPFs, but also a spatial shift between the two images might be introduced if the two filters have different phase responses. This could significantly affect the accuracy of the estimated quality.

Finally, one factor that is completely ignored in both QE_{up} or QE_{down} is the viewing distance. Many existing QEs also do not account explicitly for viewing distance. However, the sensitivity of the HVS depends heavily on the frequency. The Contrast Sensitivity Function (CSF) [22] describes how humans perceive spatial frequency, and is effectively the spatial frequency response of the HVS. The CSF has more weight at lower frequencies, and recent subjective tests in [14] and [7] also show that our eyes are more resilient to distortions as the spatial resolution decreases. Upsampling contracts $\mathbf{i}$ in the frequency domain, causing the artifacts in $\mathbf{u}_i$ to become more visible. Therefore, QE_{up} will incorrectly estimate the quality of $\mathbf{i}$.

D. Design Requirements

To develop an accurate QE for our application, our design must satisfy several requirements. First, our QE must allow the input image $\mathbf{i}$ and reference image $\mathbf{r}$ to have different spatial resolutions. This requirement is not satisfied by pixel-based QE methods, like MSE and SAE. To fulfill this requirement, our QE is based on a frequency or wavelet representation. However, this is not sufficient, since as we have seen, even QEs like VIF which use a wavelet decomposition are forced to rely on either QE_{up} or QE_{down} when $\mathbf{i}$ has a different spatial resolution than $\mathbf{r}$.

Second, our design must compare $\mathbf{r}$ and $\mathbf{i}$ as if they are being viewed at the same visual angle. To accomplish this, we assume that $\mathbf{r}$ is placed at a greater distance. Using the angular frequency resolution defined in Eq. (1), if we increase the distance by a factor of two, the visual display resolution of $\mathbf{r}$ is also multiplied by two. The frequency domain representations of the $\mathbf{i}$ and the relocated $\mathbf{r}$ are shown in Figure 4. In this figure, the downsampling rate is 2. This requirement enables us to adjust the sensitivity of HVS according to the angular resolution. Hence, we model the effect of resolution change to the perception of the distortion accurately.

The final requirement is robustness to small shifts. It is not possible to know the processing that was applied to create the

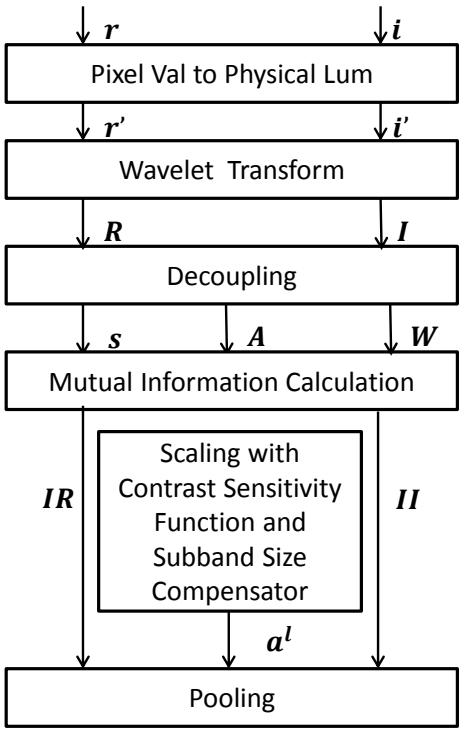

Fig. 5. Block diagram of the proposed QE

input image $\mathbf{i}$. In particular, filtering and downsampling often introduce small shifts into the resulting image. Moreover, using wavelet transform can also cause shift during decomposition. Therefore, for this application of estimating the quality of an image that has a different spatial resolution than the reference image, an accurate QE must be robust to small shifts. To achieve this, our strategy is to compute the *correlation* between image subbands instead of performing point by point comparison between these subbands. In the next section, we describe how we use wavelet decomposition and correlation to develop the proposed algorithm.

III. Multiscale Image Quality Estimator

The block diagram of the proposed QE, described in this section, is illustrated in Figure 5. As before, $\mathbf{r}$ represents the reference image and $\mathbf{i}$ represents the input image. The first block is *Pixel Value to Physical Luminance Conversion*. In this block, pixel values of $\mathbf{r}$ and $\mathbf{i}$ are converted to luminance values and these values are kept in $\mathbf{r}'$ and $\mathbf{i}'$ matrices. The second block is *Wavelet Decomposition*. $\mathbf{r}'$ and $\mathbf{i}'$ are decomposed into their subbands. These bands are collected in $\mathbf{R}$ and $\mathbf{I}$ matrices, respectively. *Decoupling* is the third block. It divides the subbands of $\mathbf{R}$ and $\mathbf{I}$ into smaller non-overlapping blocks. *Decoupling* consists of two steps. In the first, blocks of $\mathbf{R}$'s subbands are represented using Gaussian Scale Mixture (GSM). In the second step, the blocks of $\mathbf{I}$ are estimated from the blocks of $\mathbf{R}$ using a linear least square estimator. In the fourth block of the diagram, the *Mutual Information* is calculated between the blocks of $\mathbf{R}$ and $\mathbf{I}$. However, each subband has a different importance. Therefore, in the fifth block, we scale the similarity of each subband with the corresponding *sensitivity value and subband size*

TABLE I
WAVELET COEFFICIENTS

Indices	Analysis Low Pass Filter	Analysis High Pass Filter
0	0.6029490182363579	1.115087052456994
$\mp$ 1	0.2668641184428723	-0.5912717631142470
$\mp$ 2	-0.07822326652898785	-0.05754352622849957
$\mp$ 3	-0.01686411844287495	0.09127176311424948
$\mp$ 4	0.026748741080976	0

compensator. Finally, in the last block, *Pooling*, the value of the estimated quality is found. The final quality value is between 0 and 1, and this value increases as the quality increases. In what follows, we explain the details of each block.

1) Pixel Value to Physical Luminance Conversion: In the first block, we convert pixel values in $\mathbf{r}$ and $\mathbf{i}$ to luminance values $\mathbf{r}'$ and $\mathbf{i}'$. The image information is kept as pixel values in the memory. However, the display changes these values to adapt to HVS. These adapted values are called luminance values. Therefore, it is necessary to alter the pixel values to luminance values. We use the following equation to calculate the corresponding luminance value for each pixel [10]:

$$L(P) = (b + k * P)^{\gamma} \tag{2}$$

In this equation, *b*, *k* and, γ represent black-level offset, voltage scaling factor and the gamma of the display monitor, respectively. Typical parameter values are $b = 0$, $k = 0.02874$ and $\gamma = 2.2$. We apply this operation to both the reference image, $\mathbf{r}$, and the input image, $\mathbf{i}$, to obtain their luminance counterparts $\mathbf{r}'$ and $\mathbf{i}'$.

2) Wavelet Transform: In this step, the luminance image matrices $\mathbf{r}'$ and $\mathbf{i}'$ are decomposed into their subbands. There are two reasons for using subband decomposition. First, the space-frequency localization of the HVS is frequently modeled using wavelet-based approaches. Watson [23] modeled it using Cortex Transform. Daly also used a similar transform in Visible Difference Predictor (VDP) [24] model. Second, using a frequency-domain representation is required to enable a scale-independent QE. In this block, we use the $9/7$ bi-orthogonal wavelet transform [25]. The coefficients of the transform are in Table I.

The main disadvantage of using wavelet transforms is the fact that the ratio between the size of the input image and the size of the reference image can only be a power of 2. Nevertheless, the ratio of popular visual content standard pairs like CIF-QCIF, VGA-QVGA and HDTV-UHDTV satisfy this requirement.

The inputs to the wavelet decomposition, $\mathbf{r}'$ and $\mathbf{i}'$, contain the spatial luminance values. Output matrices of the wavelet decomposition block are $\mathbf{R}$ and $\mathbf{I}$. They are the wavelet transforms of $\mathbf{r}'$ and $\mathbf{i}'$, respectively. Matrices $\mathbf{R}$ and $\mathbf{I}$ consist of $3\times L + 1$ subbands where L represents the total number of levels. Each subband is represented with its level and orientation indices. Specifically, if $\mathbf{I}_{l,o}$ represents a subband of the $\mathbf{I}$, then l and o stand for the level and orientation, respectively. The subbands in the first level of the $\mathbf{I}$, ($\mathbf{I}_{1,2}$, $\mathbf{I}_{1,3}$ and $\mathbf{I}_{1,4}$), are $\mathbf{0}$ matrices.

3) Decoupling: Now, we have the wavelet subbands in $\mathbf{R}_{\mathbf{l,o}}$ and $\mathbf{I}_{\mathbf{l,o}}$ matrices. In this block, we find the relationship between each subband of $\mathbf{R}$ and $\mathbf{I}$, using two steps. In the first step, $\mathbf{R}_{\mathbf{l,o}}$ are represented as Gaussian Scale Mixtures (GSMs). In the second step, blocks of $\mathbf{I}_{\mathbf{l,o}}$ are estimated from the blocks of $\mathbf{R}_{\mathbf{l,o}}$. To estimate $\mathbf{I}_{\mathbf{l,o}}$ blocks, we assume that there is a linear relationship between the corresponding blocks of $\mathbf{R}_{\mathbf{l,o}}$ and $\mathbf{I}_{\mathbf{l,o}}$. We represent this linear relationship with attenuation

and additive noise. We also assume that, the additive noise is White Gaussian Noise and independent of the blocks of $\mathbf{R_{l,o}}$. Hence, we can use a Linear Least Square Estimator (LLSE) to find these estimation parameters. We elaborate the details of the decoupling process in the following paragraphs.

The first part of the decoupling process is based on a statistical evaluation of images. In this part, we assume that the reference image is accurately characterized by a model based on natural scene statistics (NSS) [26]. The hypothesis of NSS is that images of natural scenes occupy only a small subset of all possible images. Wainwright et al. [27] have used this information to represent images in terms of GSMs. To do that, they represent each subband of the image as a summation of Gaussian distributions [27]:

$$\vec{R}_{l,o,j} = S_{l,o,j}.\vec{U}_{l,o,j}, j = 1, \ldots, K \tag{3}$$

The index of the block is shown by j, and it can take values from 1 to the number of the blocks in the subband. In this expression, $\vec{R}_{l,o,j}$ represents the j^{th} block of subband (l,o). $\vec{U}_{l,o,j}$ is a Gaussian vector with mean zero and covariance $\mathbf{C_u}$. The covariance $\mathbf{C_u}$ can be computed by averaging the autocorrelation of $\vec{U}_{l,o,j}$ in the same subband. $S_{l,o,j}$ is a random number from a positive scalar random field. $S_{l,o,j}$ can be as follows:

$$S_{l,o,j} = \frac{\vec{R}_{l,o,j} C_u \vec{R}'_{l,o,j}}{M} \tag{4}$$

where M is the size of the block. Given $S_{l,o,j}$, $\vec{R}_{l,o,j}$ is normally distributed with $N(0, S^2_{l,o,j}\mathbf{C_u})$. Hence, after finding $S_{l,o,j}$, the blocks can be characterized as Gaussian distributions. The aim of the second part of the decoupling process is to represent the input blocks, $\vec{I}_{l,o,j}$, in terms of the reference blocks, $\vec{R}_{l,o,j}$. Here, the noise is assumed to be orthogonal to $\vec{R}_{l,o,j}$, so we use the LLSE to find the attenuation ($A_{l,o,j}$) and additive noise ($\vec{W}_{l,o,j}$) parameters using the following equations.

$$\vec{I}_{l,o,j} = A_{l,o,j}\vec{R}_{l,o,j} + \vec{W}_{l,o,j} \tag{5}$$

$$A^*_{l,o,j} = \min_{A_{l,o,j}} ||\vec{I}_{l,o,j} - A_{l,o,j}\vec{R}_{l,o,j}||_2 \tag{6}$$

This process is performed for each $\vec{R}_{l,o,j}$-$\vec{I}_{l,o,j}$ block pair. In Eq. (6), $A^*_{l,o,j}$ denotes the optimum attenuation that minimizes the noise. We can find the optimum attenuation by setting the derivative of Eq. (6) to $\mathbf{0}$. Next, we obtain the variance of the additive noise by subtracting the attenuated reference signal from noise. The corresponding attenuation and noise variance terms are expressed as follows:

$$A_{l,o,j} = cov(\vec{I}_{l,o,j}, \vec{R}_{l,o,j}) var(\vec{R}_{l,o,j})^{-1} \tag{7}$$

$$var(\vec{W}_{l,o,j}) = var(\vec{I}_{l,o,j}) - A_{l,o,j} cov(\vec{I}_{l,o,j}, \vec{R}_{l,o,j}) \tag{8}$$

In Eq. (7), $\vec{R}_{l,o,j}$ is a 16x1 vector. Hence, $var(\vec{R}_{l,o,j})$ is a scalar and we can find the inverse as long as $var(\vec{R}_{l,o,j})$ is not zero. To prevent having a zero in the denominator, we add a very small regularization constant to the $var(\vec{R}_{l,o,j})^{-1}$. These attenuation and noise variance values are obtained for each block, and are used in the next step to find the distortion in each

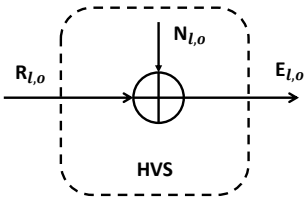

Fig. 6. Mathematical model of the reference image observation

block, employing mutual information as the distortion metric.

4) Mutual Information: Now, we have subband blocks of $\mathbf{I}$ and $\mathbf{R}$ in terms of GSMs. We use a similar model mentioned in the mutual information calculation of VIF [9]. Mutual information is a function of two random variables and is an indication of how much information we get from one random variable about the other one. As such, a lower mutual information is an indication of lack of similarity. In the extreme case of zero mutual information, the two random variables are independent. Therefore, mutual information of each block gives us an indication of local similarity.

$\mathbf{R}$ and $\mathbf{I}$ are the reference and input signals. We model $\mathbf{E}$ and $\mathbf{F}$ as being created by passing $\mathbf{R}$ and $\mathbf{I}$ through the HVS, as in [9], where the HVS is modeled as Additive White Gaussian Noise (AWGN) with unit variance. $A_{l,o,j}$ and $\vec{W}_{l,o,j}$ represent the attenuation, noise for each subband block, respectively. $s_{l,o,j}$ is a realization of Sl, o, j and it is computed as in [9], Eq(15). In this step, two items will be performed as follows.

First, the mutual information [28] between non-overlapping 4x4 blocks of $\mathbf{E_{l,o}}$ and $\mathbf{R_{l,o}}$ $(I(\mathbf{E_{l,o}}; \mathbf{R_{l,o}}))$ is found for all (l, o) pairs. Since we have used a correlation based block similarity metric during the mutual information computation, we have relieved the effect of shift variance which can occur as a result of wavelet decomposition or decimation filter mis-match. To measure the effect of shift variance to the quality estimator, we have exploited the third test scenario in Section III of [29]. According to this test scenario, identical quality scores should be produced despite a simple transformation of a degraded image, like cropping by a few pixels. We have computed the maximum quality variation of MIQE when the degraded images are cropped by between 0-9 pixels for 24 images. Mean and maximum of maximum quality variation values are $1.22 * 10^{-5}$ and $3.46 * 10^{-5}$, respectively. $\mathbf{E_{l,o}}$ is obtained by adding $\mathbf{R_{l,o}}$ with $\mathbf{N}$ which is an AWGN, as illustrated in Figure 6. Before computing the mutual information, the block indices are represented as a vector. Moreover, $\vec{R}_{l,o,j}$ has a Gaussian distribution if $\mathbf{s_{l,o,j}}$ is known. Using these facts, we can find the mutual information for a (l, o) pair as

$$I(\mathbf{E_{l,o}}; \mathbf{R_{l,o}}|s^K) = \frac{1}{2} \sum_{j=1}^{K} log_2 \left(\frac{s_{l,o,j}^2 \mathbf{C_u} + \sigma_n^2 \mathbf{I}}{\sigma_n^2 \mathbf{I}} \right) \tag{9}$$

Second, the mutual information between the blocks of $\mathbf{F_{l,o}}$ and $\mathbf{R_{l,o}}$ is found. The relationship between $\vec{F}_{l,o,j}$ and $\vec{R}_{l,o,j}$ is

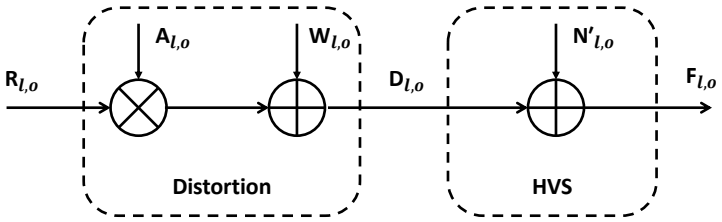

Fig. 7. Mathematical model of the input image observation

shown in Figure 7. Using this relationship we can write $\vec{F}_{l,o,j}$ in terms of $\vec{R}_{l,o,j}, A_{l,o,j}, \vec{W}_{l,o,j}$ and $\vec{N}'_{l,o,j}$ as follows:

$$\vec{F}_{l,o,j} = A_{l,o,j} * \vec{R}_{l,o,j} + \vec{W}_{l,o,j} + \vec{N}'_{l,o,j} \tag{10}$$

We use Eqs. (9) and (10) to find the mutual information between $\mathbf{F_{l,o}}$ and $\mathbf{R_{l,o}}$ as follows:

$$I(\mathbf{F_{l,o}}; \mathbf{R_{l,o}} | s^K) = \frac{1}{2} \sum_{j=1}^{K} log_2 \left(\frac{s^2_{l,o,j} A^2_{l,o,j} \mathbf{C_u} + \sigma^2_{q_{l,o,j}} \mathbf{I}}{\sigma^2_{q_{l,o,j}} \mathbf{I}} \right) \tag{11}$$

where, $\sigma^2_{q_{l,o,j}} = \sigma^2_{w_{l,o,j}} + \sigma^2_n$. We use Eqs. (9) and (11) to find the mutual information for each subband. However, we also need to scale these values ($I(\mathbf{E_{l,o}}; \mathbf{R_{l,o}})$, $I(\mathbf{F_{l,o}}; \mathbf{R_{l,o}})$) according to the size and the importance of the subband.

5) Scaling with Contrast Sensitivity Function and Subband Size Compensator: To estimate the quality accurately, we next scale the value of each (l, o) pair. To calculate the magnitude of these scaling coefficients, we take into account two factors: the visual importance of each subband determined by the characteristics of HVS, and the size difference of each subband.

First, HVS has a different weight for each subband [22]. The corresponding weight changes according to the resolution and height of the screen, the distance between the viewer and the screen, and the level of the subband. To find the effect of HVS, we first calculate the angular frequency by Eq. (1). As described in [21], people are also more sensitive to vertical and horizontal wavelet subbands than the diagonal wavelet subbands. Therefore, we adjust angular frequency values for different orientations as follows:

$$f(o) = \begin{cases} f, & \text{if } o \, \epsilon \, horizontal, vertical \\ f/0.7, & \text{otherwise} \end{cases}$$

Next, the CSF is calculated using the following formula from [20]:

$$CSF(f) = (0.69 + 0.31 * f).e^{-0.28*f} \tag{12}$$

This CSF function is used to compute the quality of monochromatic images. We can take into account the effect of color if we use the chromatic and achromatic CSF functions as described in [30]. In calculating the QE, if we use both chromatic and achromatic CSF functions instead of the one described in Eq. (12) the corresponding QE is called MIQE-Color (MIQEC).

The second factor is the size of the subband. As the subband level increases by one, the size of the the band decreases by four. Rouse et al. [31] show that incorporating this improves the accuracy of quality estimation. As a result, each subband

CSF value is also scaled with 2^{2l} to compensate the size of each level l. Hence, the scaling coefficient for each subband is calculated as follows:

$$g_l = CSF(f(2l)) * 2^{2l} \tag{13}$$

6) Pooling: In this step, we calculate the quality using the scaling coefficients and mutual information for each subband. Specifically, we compute the estimated quality by

$$MIQE = \frac{\sum_{l,o \in subbands} g_l * I_{I,l,o}}{\sum_{l,o \in subbands} g_l * I_{R,l,o}} \tag{14}$$

In this expression, g_l is the scaling coefficient for level l given in Eq. (13). $I_{I,l,o}$ and $I_{R,l,o}$ represent the scaled mutual information of the input and the reference signals for the pair (l, o).

To check the validity of the proposed algorithm, it is necessary to compare the estimated quality values with the preference of viewers. In the following section we describe the details of our subjective tests to obtain viewer preferences.

IV. Subjective Tests

Subjective tests are performed to obtain the viewers' preferences for images that have different spatial resolutions. The test results are used both to ensure that the proposed algorithm is accurate and to compare our algorithm with existing approaches. Our subjective tests are performed by employing a paired comparison test using two distorted test images, one with full resolution and the other with half resolution. In the following subsections, we describe how we create the test set, perform subjective tests, and analyze results.

A. Test Set Creation

Original images are selected from the National Park Service Digital Image Archive and WikiMedia Commons. There are in total 24 images in 5 categories. These categories are human, landscape, plant, animal, and urban. These images represent a wide variety of scenes. They include homogeneous regions, edges, and details. The resolution of the reference images is 800×800. The images that are downloaded from the databases have greater resolution than 800×800; hence, they are decimated and cropped. We decimate the images such that minimum edge resolution of the decimated image is greater than or equal to 800. Next, we determine the border coordinates to crop the image. We choose these coordinates to keep the maximum information in the reference image.

The test images shown to viewers consist of two groups: high resolution and low resolution images. High resolution test images are obtained by distorting the reference images. Low resolution test images are created by decimating the reference images and then corrupting the decimated images.

To obtain a subjective estimate of image quality, it is conventional to ask viewers for their opinion score and then compute the mean opinion scores (MOSs). This method is effective when the reference image and the test image have equal resolutions; however, finding the qualities of images that have different spatial resolutions is more challenging [32]. Therefore, instead of obtaining the MOS score for each image, we use paired comparison to find the relative preference of test images with different resolutions.

To obtain the test images, we corrupt a full-size reference image $\mathbf{r}$ and a decimated reference image $\mathbf{d}_r$ using one of the following four distortions: compression with JPEG, compression with JPEG 2000, blurring, or noise. For the two compression algorithms, we use three distortion levels, representing good quality, medium quality and bad quality images. However, both blur and noise have only two distortions levels: good and bad. A $2D$ Gaussian low pass filter is used for blurring, where the distortion level of the blur is controlled by the standard deviation of the filter. In addition, we use AWGN as noise distortion. To explore the impact of filter type, we use two low-pass filter types to decimate the $\mathbf{r}$. The first is the Non-Normative LPF, which is a Sine-windowed Sinc-function. We implement the filter employing Eq. 4 from Segall and Sullivan [33]. The parameters are $D = 2.5$ and $N = 3$ and phase offset$= 0$, and the implementation is performed using floating point. The second LPF is the Raised-Cosine function which is formulated as follows.

$$f(x) = \frac{1}{2}(1 + cos(\frac{\pi * (w - w_c * (1 - \alpha))}{2 * \alpha * w_c})) \tag{15}$$

In this expression, w_c is the cut-off frequency and α is the roll-off factor. Non-Normative filter only causes blurring. It especially decreases the magnitude of the signal between frequencies 0.4π and 0.5π. On the other hand, Raised Cosine causes both aliasing and blurring. When it is necessary to distinguish the two decimated reference images, we use $\mathbf{d}_{r1}$ and $\mathbf{d}_{r2}$, to represent $\mathbf{d}_r$s created using the Non-Normative and Raised-Cosine LPFs, respectively. All distortion types are used to distort decimated reference images $\mathbf{d}_{r1}$, while only JPEG 2000 is used to distort decimated reference images $\mathbf{d}_{r2}$. We use JP2K-RC as an abbreviation for the latter case. To minimize the number of comparisons in our subjective test, we use all images in the reference image set to obtain $\mathbf{d}_{r1}$ and half of the images in the set to obtain $\mathbf{d}_{r2}$.

The next challenge is to determine appropriate levels of distortion for each image pair in the paired comparison experiment. Ideally, we should choose the distortion level of the low resolution test image, relative to that of the high resolution test image, such that we obtain the most information about how the QEs perform. If we choose the image pairs such that all QEs rank the images in the same order, our subjective test will not be able to distinguish the relative performance among the QEs. If instead, we pick an image pair such that one QE classifies the images correctly and the remaining QEs classify them incorrectly, we immediately determine the best QE for that image pair. However, this is difficult because we do not know the viewers' preferences a-priori.

In consequence, we use a method similar to that described in [34] to select the images. According to this approach, for a given high resolution test image, we select a corresponding low resolution test image which divides a set of QEs into two groups, with approximately half of the QEs classifying an image pair one way and approximately half classifying the pair the other way. We perform QE$_{down}$ approach while estimating the quality of the low resolution test image using the Non-Normative filter [33].

We use this strategy to create a pair of images for each reference image, each distortion, and each distortion level (good, medium, or bad), where one image in the pair has high resolution and the other low. In our subjective test, in addition to simply having viewers rate this pair, we create additional pairs using these images. For JPEG and JPEG 2000, we use three additional pairs per reference image by pairing the bad-quality high-resolution image with the medium-quality low-resolution image, and the medium-quality high-resolution image with both good and bad low-resolution images. For blur and noise, we create

one additional pair per reference image by pairing the bad-quality high resolution image with the good-quality low resolution image. These additional pairs provide additional information about viewers' preferences while balancing the expense of the subjective test.

B. Subjective test implementation

Until now, we have described how we create the test images and form the possible pairs. Next, we describe the subjective tests themselves. The tests contain a total of 504 pairs of images, with 60 people viewing 252 pairs each. Hence, each pair is seen by 30 viewers, and each test session takes approximately an hour. The viewers selected for the test are graduate students and undergraduate students. They have clear vision and they are non-experts.

During tests, the viewers saw images that have different spatial resolutions side by side. They were asked to choose which image they prefer more. They had 10 seconds to see each image pair, but they could choose a preference before or after 10 seconds. The distance between the viewers and the screen was 4 times the height of the low resolution image. Tests were performed using 17" Dell N7110 Inspiron with a display resolution of 1600×900.

C. Analysis of Subjective Test Results

In this section, we compare the performance of our proposed QE to six others using the results of the subjective test. We begin by examining whether the QEs can rank the images in each pair according to the viewers' preferences using QE_{down} setup. Then, we study how using the QE_{up} approach will affect the ranking. We also explore the sensitivity to the choice of decimation filter applied in the QE_{down}, with emphasis on the case when it differs from the filter used to create low resolution test images. Next, we apply a well-known statistical model to transform viewer preferences into an estimate of subjective quality among all the images with the same reference image and distortion type. Lastly, we explore the performance of the proposed algorithm when the images have the same resolutions. We compare our proposed MIQE and MIQEC to seven FR and two NR state-of-the-art QE methods in a QE_{down} set-up. Compared state-of-the-art FR QE methods are Structural Similarity Index (SSIM) [8], VIF [9], Visual Signal-to-Noise Ratio (VSNR) [10], PSNR, Multiscale-SSIM (MSSIM) [11], Information Content Weighted SSIM (IWSSIM) [12] and Complex Wavelet SSIM (CWSSIM) [35]. The two NR QE methods are BIQI [4] and BRISQUE [6].

In our first comparison, we begin by identifying the image that was preferred by most viewers, for each pair shown to viewers. We also use each QE to rank the two images in each pair. Dividing the total number of correct rankings by the total number of pairs gives us the fraction of correct rankings. Table II shows the percentage of correct rankings for each QE according to different distortions and different decimation filters.

In "Different Distortions" heading of Table II, we evaluate the effect of different distortions to the correct ranking. The Non-Normative filter is used while performing QE_{down}, and test images are obtained by distorting the images in $\mathbf{d}_{r1}$. According to the table, MIQEC algorithm performs better than other QEs for all distortions except JP2K. If the distortion is JP2K, MIQE performs better than other QEs. Both MIQE and MIQEC have higher ranking fractions for blur and noise compared to other distortions. The results also show that the proposed QE is robust to small shifts. The proposed QE uses a wavelet

TABLE II
FRACTION OF CORRECT RANKINGS

	Different Distortion				Different Filter	
	Blur	JP2K	JPEG	Noise	JP2K	JP2K-RC
SSIM	0.597	0.601	0.569	0.354	0.569	0.556
VIF	0.556	0.531	0.597	0.396	0.500	0.417
VSNR	0.722	0.677	0.667	0.660	0.653	0.653
MIQE	0.819	**0.774**	0.729	0.910	**0.736**	**0.792**
PSNR	0.708	0.747	0.688	0.910	0.708	0.708
MSSIM	0.597	0.524	0.569	0.340	0.500	0.403
IWSSIM	0.542	0.531	0.569	0.340	0.500	0.403
MIQEC	**0.917**	0.760	**0.736**	**0.924**	0.736	0.708
CWSSIM	0.736	0.760	0.639	0.507	0.514	0.583
BIQI	0.667	0.573	0.59	0.563	0.653	0.681
BRISQUE	0.625	0.510	0.576	0.507	0.542	0.514

LPF during subband decomposition instead of employing the Non-Normative filter to estimate quality. Nevertheless, MIQEs's correct ranking fraction is higher than other QEs.

In "Different Filters" heading of Table II, the effect of different filters to the correct ranking is analyzed. The fifth and the sixth columns of Table II illustrate the fraction of correct rankings when the decimation filter is Non-Normative and Raised Cosine LPF, respectively. In each case, we use the same LPF during the creation of test images and the decimation operation in QE$_{down}$. Here, the comparison is performed for the reference images which are common in $\mathbf{d}_{r1}$ and $\mathbf{d}_{r2}$. Note that, the number of images in $\mathbf{d}_{r1}$ and $\mathbf{d}_{r2}$ are different. Therefore, the QE values for JP2K in "Different Distortions" and "Different Filters" are slightly different. The ranking fractions show that MIQE performs slightly better when the Raised Cosine LPF is employed instead of the Non-Normative to create test images. Furthermore, the ranking fractions of other QEs either decrease or remain the same when the Raised Cosine LPF is employed during QE$_{down}$.

We also study the effect of a mis-match between the decimation filters that are utilized in the creation of the test images and QE_{down}. We employ the 9/7 biorthogonal wavelet, raised cosine and non-normative filters to compute the fraction of correct ranking scores in this scenario. If the test images are created using the image set d_{r1}, we apply the wavelet and the raised cosine filters in the decimation block of the QE_{down}. Otherwise, we employ the Non-Normative and Wavelet filters to compute the quality. The filter mis-match may cause the estimated quality to be under-estimated if the QE is not shift invariant. On the other hand, the Dec block causes QE_{down} to over-estimate quality because it causes detail loss in the reference image. The two effects may, coincidentally, cancel each other, producing an accurate quality estimation. No one filter in QE_{down} is likely to perform best for all possible content, distortions, and image-creation filter. When we search for the best decimation filter in QE_{down} over a variety of scenarios, in a few cases, QE_{down} outperforms MIQE and MIQEC. However, MIQE and MIQEC do not rely on a specific filter, and there is no one quality estimation approach that consistently outperforms the others, i.e., the approach providing the highest correct ranking scores varies depending on the applied filter, content, and distortion. On average, MIQEC has the highest ranking scores. Tables III and IV show the percentage of correct rankings for each QE using Wavelet and Raised Cosine decimation filters in QE_{down}, respectively.

As described in Section II, we can also use QE_{up} to estimate the quality using existing approaches. We use bilinear and 9/7 biorthogonal wavelet filters while interpolating the test images in QE_{up}. We explore the effect of QE_{up} using different distortions and decimation filters. For low resolution test images, a QE_{up} approach results in smaller QE values compared to

TABLE III
FRACTION OF CORRECT RANKINGS WHEN THE LOW PASS FILTER IS A WAVELET FILTER

	Different Distortion				Different Filter	
	Blur	JP2K	JPEG	Noise	JP2K	JP2K-RC
SSIM	0.819	**0.788**	0.722	0.590	**0.750**	**0.792**
VIF	0.639	0.545	0.618	0.576	0.528	0.569
VSNR	0.597	0.677	0.729	0.882	0.625	0.750
MIQE	0.819	0.774	0.729	0.910	0.736	0.792
PSNR	0.819	0.719	0.694	**0.965**	0.681	0.764
MSSIM	0.875	0.615	0.715	0.396	0.597	0.528
IWSSIM	0.875	0.642	**0.771**	0.521	0.625	0.528
MIQEC	**0.917**	0.760	0.736	0.924	0.736	0.728
CWSSIM	0.778	0.490	0.556	0.410	0.514	0.486
BIQI	0.667	0.573	0.59	0.563	0.653	0.681
BRISQUE	0.625	0.510	0.576	0.507	0.542	0.514

TABLE IV
FRACTION OF CORRECT RANKINGS WHEN THE LOW PASS FILTER IS A RAISED COSINE FILTER

	Different Distortion				Different Filter	
	Blur	JP2K	JPEG	Noise	JP2K	JP2K-RC
SSIM	0.819	0.684	0.681	0.465	0.667	0.694
VIF	0.514	0.538	0.604	0.465	0.514	0.597
VSNR	0.514	0.747	**0.743**	0.951	0.694	**0.806**
MIQE	0.819	**0.774**	0.729	0.910	**0.736**	0.792
PSNR	0.750	0.760	0.708	**0.951**	0.722	0.792
MSSIM	0.514	0.531	0.611	0.354	0.500	0.500
IWSSIM	0.625	0.545	0.667	0.382	0.514	0.514
MIQEC	**0.917**	0.760	0.736	0.924	0.736	0.728
CWSSIM	0.778	0.503	0.590	0.424	0.514	0.486
BIQI	0.667	0.573	0.59	0.563	0.653	0.681
BRISQUE	0.625	0.510	0.576	0.507	0.542	0.514

QE_{down}. Ranking classification results show that, MIQEC has the highest ranking scores. Moreover, the fraction of correct ranking scores obtained using the wavelet filter are higher than that of the bilinear filter.

The ranking classification only considers whether one image is preferred to the other, either objectively or subjectively, but does not consider the magnitude of the preference. Therefore, in our second comparison we compute the distance similarity between the subjective results and the QE scores. Before describing how we compute distance similarity, we first describe how we obtain distances both subjectively and objectively.

Assigning a continuous quality value to each image is a well-defined problem in statistics literature [36],[37]. The relationship between the actual qualities of images I_i and I_j, and the viewers' preference is expressed as follows:

$$P(Y = \pi_{ij}) = Q_i - Q_j + Z \tag{16}$$

where Q_i and Q_j denote the qualities of I_i and I_j, respectively, Y is a random variable which represents the tendency to choose I_i over I_j, and π_{ij} denotes the ratio of preferring I_i to all comparisons between I_i and I_j, and Z is a random variable that models the deviation from the actual quality difference. The distribution of Z changes according to the model used. Thurstone-Mosteller (*TM*) [36] assumes that Z has a Gaussian distribution, whereas Bradley-Terry (*BT*) [37] models Z with a logistic distribution. Handley [38] shows that the *BT* model provides two important advantages. First, *BT* offers a mathematically developed model where quantities like confidence can be calculated analytically. Second, unlike *TM*, we can use *BT* when we do not have all the paired comparison results [38]. Because of these reasons, we choose *BT* to convert paired

comparisons to continuous values.

We use Eq. (20) to find the relationship between the preference probability and quality as follows:

$$\pi_{mn} = \int_{Q_n - Q_m}^{\infty} \frac{e^{-\frac{(x-\mu)}{s}}}{s(1+e^{-\frac{(x-\mu)}{s}})^2} dx = \frac{e^{\frac{\mu}{s}}}{e^{\frac{\mu}{s}} + (\frac{\pi_n}{\pi_m})^{\frac{1}{s}}} \tag{17}$$

In Eq. (17), $Q_m - Q_n = log(\frac{\pi_m}{\pi_n})$, where π_k's are positive real numbers and $\sum_{k=1}^{N} \pi_k = 1$. μ denotes the mean of the logistic distribution and s is a parameter proportional to the standard deviation of the distribution. If $\mu = 0$ and $s = 1$, we can write $\pi_{mn} = \frac{\pi_m}{\pi_m + \pi_n}$. We find π_m values using maximum likelihood. ***BT*** assumes that all comparisons are independent. Hence, we can calculate the likelihood function for all comparisons by.

$$L(\pi_{mn}) = \prod_{m<n} \binom{n_{mn}}{\alpha_{mn}} \pi_{mn}^{\alpha_{mn}} (1 - \pi_{mn})^{n_{mn} - \alpha_{mn}} \tag{18}$$

In this expression, n_{mn} denotes the total number of viewers and α_{mn} represents the number of viewers that prefer I_m. To maximize the likelihood, we take the derivative of Eq. (18) with respect to π_m and equate it to zero:

$$\sum_{m \neq n} \frac{n_{mn}}{\pi_m + \pi_n} - \sum_{m<n} \frac{\alpha_{mn}}{\pi_m} = 0 \tag{19}$$

The ***BT*** model provides relative scores only for a specific image and distortion in our paired comparisons. Therefore, we cannot analyze these scores across images or across distortions. Further, only their differences are well-defined, not their actual values. Therefore, we compute the differences of the ***BT*** scores, (ΔBT), for all images with the same reference image and same distortion, regardless of spatial resolution.

Next, to determine the distances between these images objectively, we recognize that the QEs have different ranges and often have a nonlinear relationship with subjective quality. Therefore, it is important to normalize the QE scores. In this paper, we choose to normalize the QE scores using the observed probability distribution of QE scores across the set of 504 images we considered. Specifically, we use the observed cumulative distribution function of the QE scores to map the QE scores into values between 0 and 1. During this mapping, we decrease the Kullback-Liebler Divergence [28] among different QEs. We then define an objective distance between two images by the difference, ΔQE, between their normalized QE scores.

Now, we have sets of N images for each reference image and distortion combination, and we have meaningful ***BT*** scores defined within each set. We compute the distances between each pair of images, both objectively and subjectively to create a collection of $T = \binom{N}{2}$ pairs. Thus, there are $\binom{T}{2}$ possible combinations of two pairs, which may or may not contain one image in common. To assess the accuracy of the QEs, we compute the similarity of the distances of these combinations.

While there are many ways to compute the similarity among distances, in this paper we focus on the following strategy, which is illustrated by the example in Figure 8. We have three test images, TI_1, TI_2 and TI_3, each created from the same reference image using the same distortion. The ***BT*** scores of the images are BT^1, BT^2 and BT^3, and we assume that as the index of the image increases the quality improves. Therefore, $\Delta BT^1 < \Delta BT^2 < \Delta BT^3$, and $\Delta BT^1 - \Delta BT^2 < 0$ and $\Delta BT^2 - \Delta BT^3 < 0$.

Next, consider two QEs, QE_A and QE_B, for estimating quality, where the QE_A and QE_B values for image TI_i are QE_A^i,

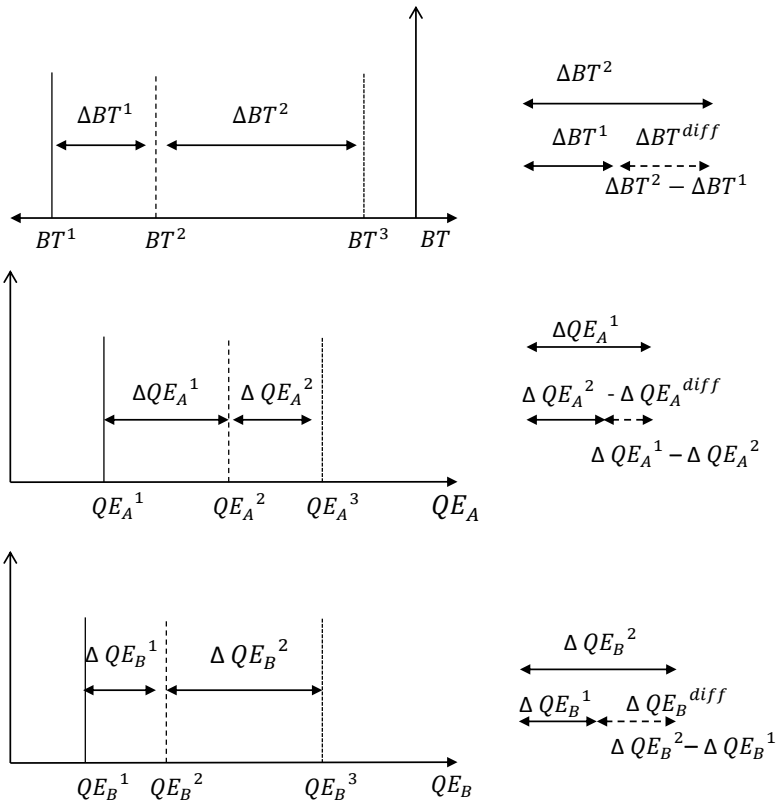

Fig. 8. Illustration of distance dissimilarity

and QE_B^i, respectively. As illustrated in Figure 8, the distribution of QE_B scores are more similar to the distribution of BT scores than the QE_A scores are. This is because $|\Delta QE_B^1| < |\Delta QE_B^2|$ and $|\Delta BT^1| < |\Delta BT^2|$, while $|\Delta QE_A^1| > |\Delta QE_A^2|$. A correct ordering of distances only happens when all of the following conditions hold: $sign(\Delta BT_1) = sign(\Delta QE_1)$, $sign(\Delta BT_2) = sign(\Delta QE_2)$, and $sign(\Delta BT_1 - \Delta BT_2) = sign(\Delta QE_1 - \Delta QE_2)$.

We employ this approach to compute the distance similarities for each distortion type. We also calculate the distance similarities for JP2K when we use different decimation filters during QE computation. We obtain the fraction of correct distance similarities by dividing the total number of correct distance similarities by the total number of comparisons. Table V presents these fractions for each distortion type and for different decimation filters when the distortion is JP2K. According to the table, MIQE has the highest distance similarity fraction. When the decimation filters are different, the distance similarity values of JP2K-RC is relatively higher than JP2K for all QEs except for SSIM. Hence, it can be inferred that the filter type affects the correct distance similarity fraction.

In addition to these metrics, we explore the relationship between QE score differences and BT score differences graphically. Figures 9-12 illustrate the relationship between QEs and BT scores for all distortions. We also compute Spearman Rank Correlation Coefficients (SRCCs) for each QE. These SRCC values are shown in Table VI.

During comparison, we also compute the standard deviation and 95% Confidence Interval CI of the metric scores for each distortion. The results are tabulated in Tables VII-X. In these tables, we compute 95% CI using the following expression:

TABLE V
FRACTION OF CORRECT DISTANCE SIMILARITY FOR DIFFERENT DISTORTIONS

	Different Distortion				Different Filter	
	Blur	JP2K	JPEG	Noise	JP2K	JP2K-RC
SSIM	0.117	0.369	0.350	0.206	0.344	0.320
VIF	0.264	0.334	0.423	0.161	0.294	0.354
VSNR	0.161	0.366	0.463	0.194	0.392	0.417
MIQE	**0.539**	**0.465**	**0.539**	**0.533**	**0.471**	**0.534**
PSNR	0.167	0.426	0.460	0.161	0.406	0.421
MSSIM	0.147	0.334	0.389	0.192	0.287	0.330
IWSSIM	0.272	0.358	0.400	0.208	0.314	0.343
MIQEC	0.519	0.457	0.469	0.506	0.474	0.342
CWSSIM	0.406	0.328	0.411	0.228	0.324	0.192
BIQI	0.025	0.048	0.019	0.008	0.029	0.103
BRISQUE	0.011	0.052	0.022	0.031	0.058	0.042

TABLE VI
SPEARMAN CORRELATION COEFFICIENT BETWEEN BRADLEY TERRY SCORE DIFFERENCE AND QUALITY ESTIMATOR DIFFERENCE FOR EACH DISTORTION TYPE

	SSIM	VIF	VSNR	MIQE	PSNR	MSSIM	IWSSIM	MIQEC	CWSSIM	BIQI	BRISQUE
Blur	0.721	0.711	0.710	0.691	0.656	0.746	0.724	**0.782**	0.430	0.434	0.705
JP2K	0.302	0.376	0.399	**0.503**	0.432	0.313	0.305	0.446	0.181	0.274	0.244
JPEG	0.326	0.303	0.464	**0.637**	0.411	0.263	0.280	0.498	0.439	0.297	0.309
Noise	0.255	0.193	0.482	**0.651**	0.387	0.183	0.185	0.647	0.201	0.394	0.255

$$CI = z_{\alpha/2}.\sigma/\sqrt{K} \tag{20}$$

where $z_{\alpha/2}$ = 1.96 (α =0.95), σ is the standard deviation and K is the number of images.

Finally, MIQE and MIQEC have been designed to compare images with different spatial resolutions. However, it also has a reasonable performance if the images have the same resolution. We employ LIVE [39], TID [40] and CSIQ [21] databases to evaluate the performance of MIQE and MIQEC when the test and reference images have the same resolution. We calculate Pearson, Spearman and Kendall correlation coefficients using the corresponding regression approach for each database. The results are provided in Table XI. Highlighted values show the highest correlation coefficients. The QE with the highest correlation values changes for different databases and correlation types. However, the correlation coefficients of MIQE and MIQEC are very close to the highlighted values in each case. PSNR performs relatively well in Tables II-VI and poorly in Table XI where the test and reference images have the same resolution. This is the first subjective test for different resolutions in which these QEs have been compared.

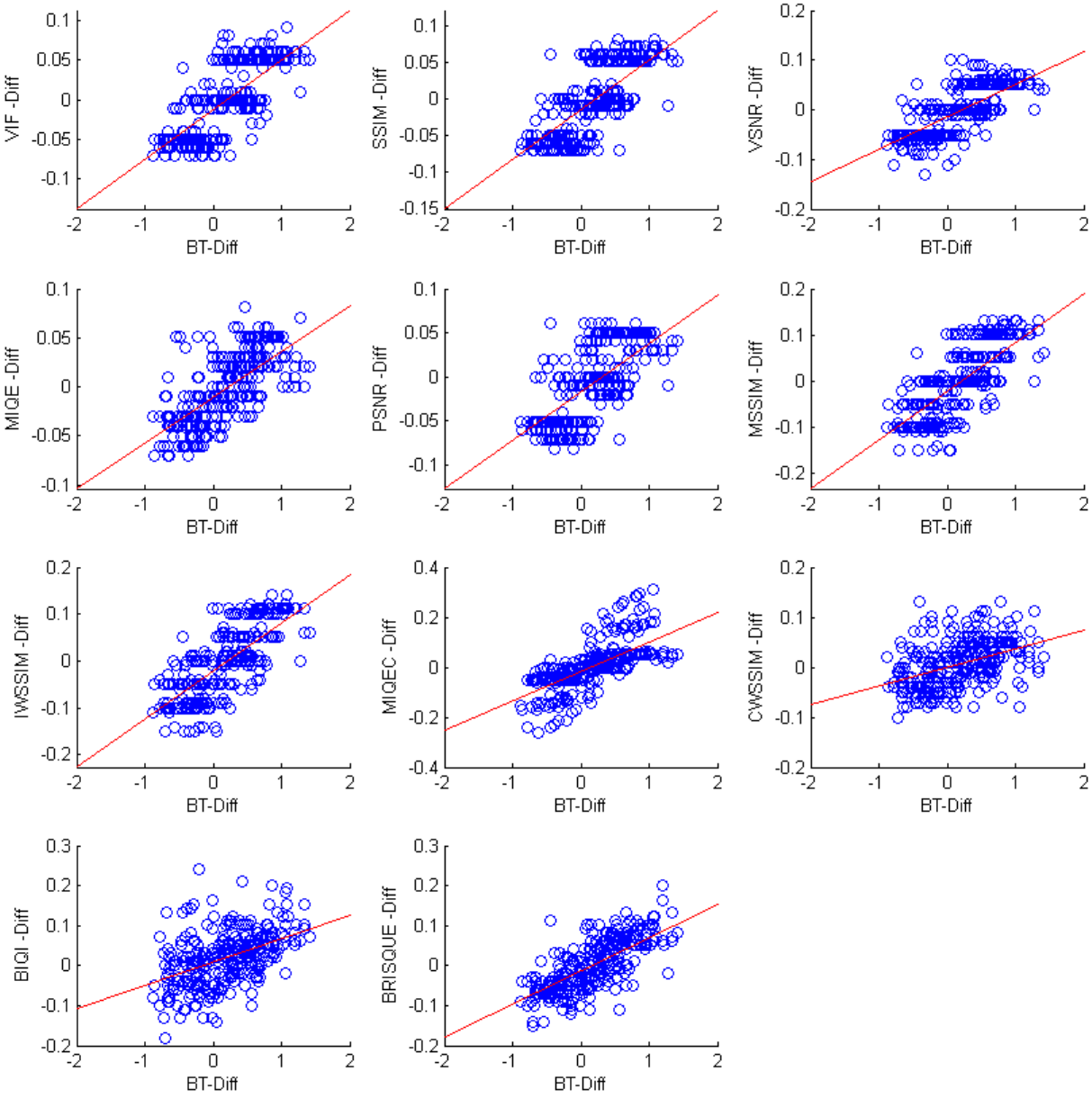

Fig. 9. Illustration of relationship between QE scores' difference and BT scores' difference when the distortion is Blur.

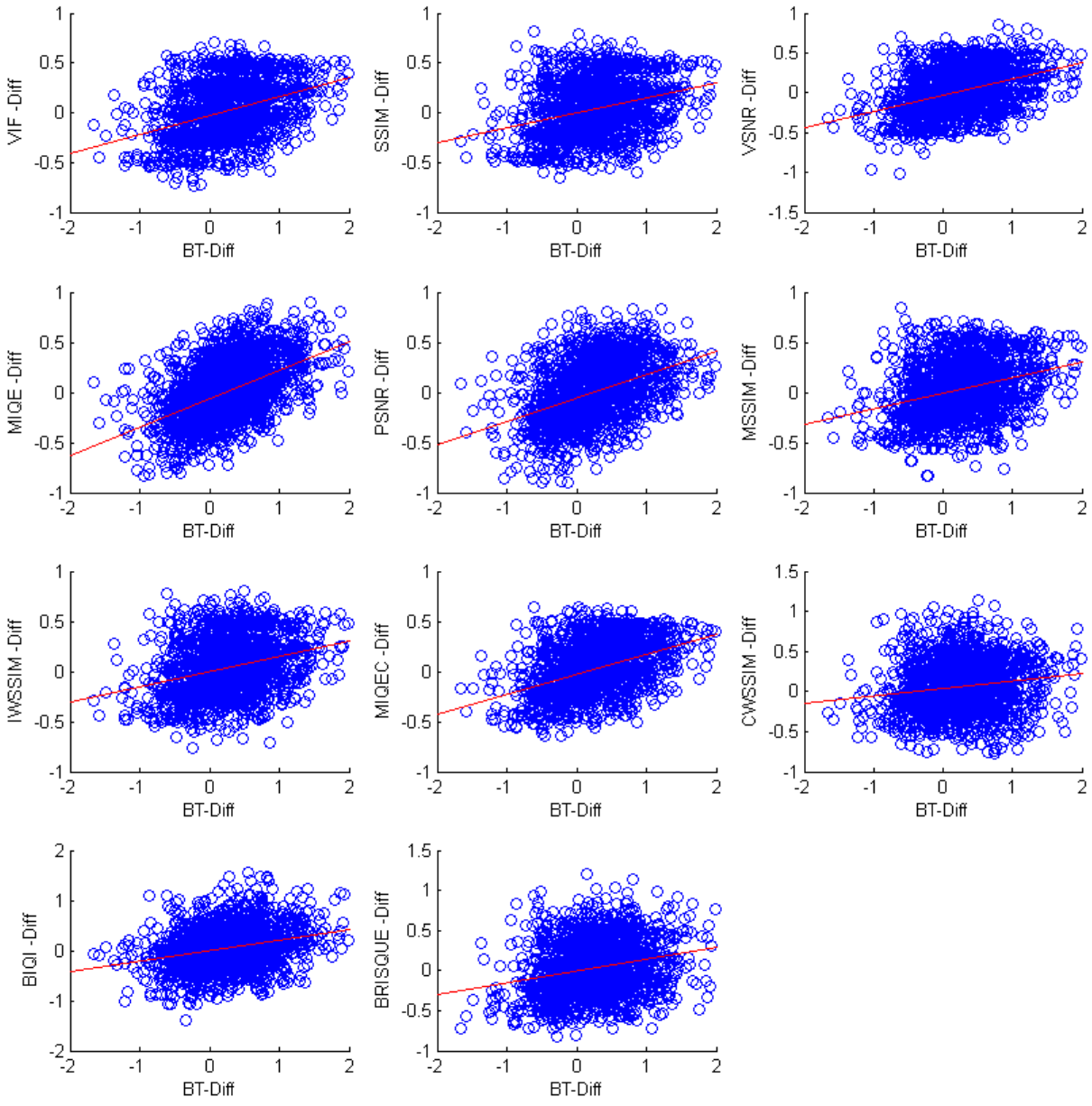

Fig. 10. Illustration of relationship between QE scores' difference and BT scores' difference when the distortion is JP2K.

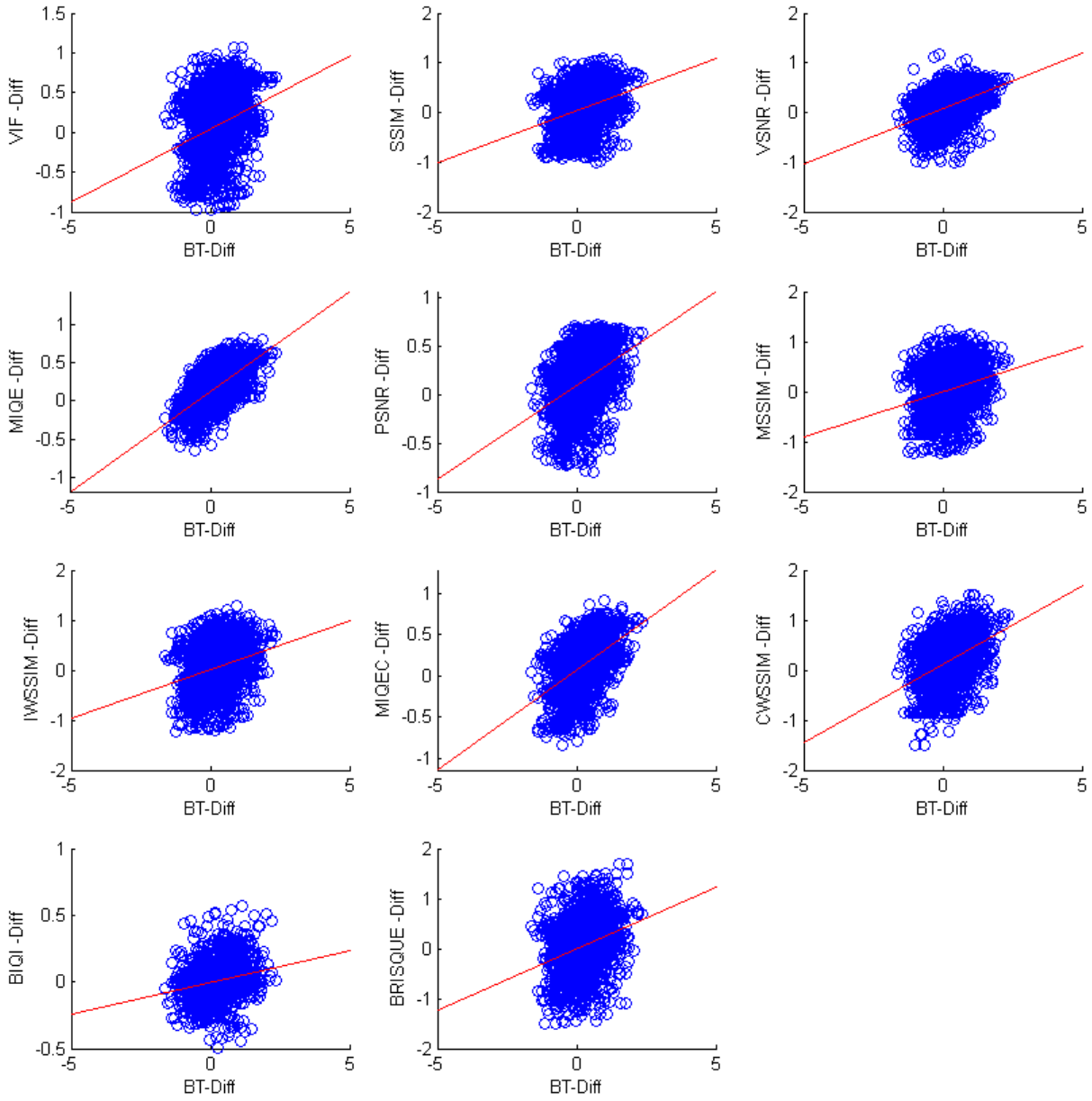

Fig. 11. Illustration of relationship between QE scores' difference and BT scores' difference when the distortion is JPEG.

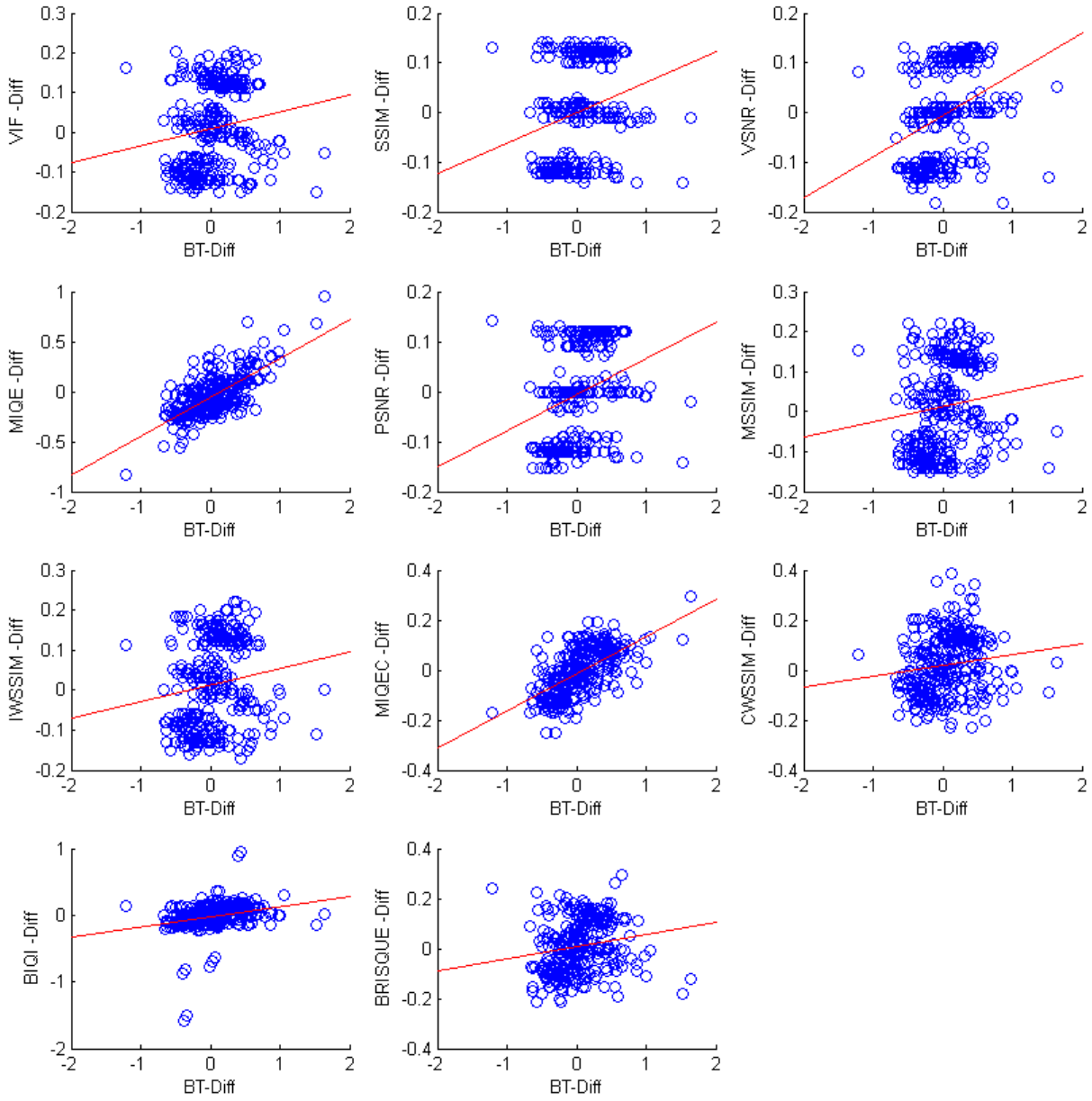

Fig. 12. Illustration of relationship between QE scores' difference and BT scores' difference when the distortion is Noise.

TABLE VII
STANDARD DEVIATION OF CORRECT RANKING SCORES FOR EACH QUALITY ESTIMATOR

	SSIM	VIF	VSNR	MIQE	PSNR	MSSIM	IWSSIM	MIQEC	CWSSIM	BIQI	BRISQUE
Blur	0.207	0.273	0.235	0.190	0.242	0.235	0.186	0.139	0.225	0.225	0.220
JP2K	0.248	0.181	0.235	0.150	0.185	0.241	0.248	0.174	0.283	0.246	0.207
JPEG	0.170	0.177	0.155	0.096	0.128	0.177	0.177	0.147	0.198	0.177	0.18
Noise	0.169	0.133	0.139	0.139	0.220	0.115	0.115	0.163	0.243	0.254	0.242
JP2K-RC	0.233	0.175	0.226	0.149	0.193	0.221	0.221	0.215	0.262	0.181	0.251
JP2K-NN	0.273	0.221	0.214	0.184	0.234	0.273	0.273	0.166	0.289	0.241	0.258

TABLE VIII
95% CONFIDENCE INTERVAL OF CORRECT RANKING SCORES FOR EACH QUALITY ESTIMATOR

	SSIM	VIF	VSNR	MIQE	PSNR	MSSIM	IWSSIM	MIQEC	CWSSIM	BIQI	BRISQUE
Blur	0.083	0.109	0.094	0.076	0.097	0.094	0.074	0.056	0.090	0.109	0.088
JP2K	0.099	0.072	0.094	0.060	0.074	0.096	0.099	0.070	0.113	0.072	0.083
JPEG	0.068	0.071	0.062	0.038	0.051	0.071	0.071	0.059	0.079	0.071	0.074
Noise	0.068	0.053	0.056	0.056	0.088	0.046	0.046	0.065	0.097	0.053	0.097
JP2K-RC	0.132	0.099	0.128	0.084	0.109	0.125	0.125	0.122	0.148	0.102	0.142
JP2K-NN	0.154	0.125	0.121	0.104	0.132	0.154	0.154	0.094	0.163	0.136	0.146

TABLE IX
STANDARD DEVIATION OF CORRECT DISTANT SIMILARITY SCORES FOR EACH QUALITY ESTIMATOR

	SSIM	VIF	VSNR	MIQE	PSNR	MSSIM	IWSSIM	MIQEC	CWSSIM	BIQI	BRISQUE
Blur	0.286	0.198	0.265	0.288	0.215	0.220	0.232	0.390	0.277	0.055	0.025
JP2K	0.181	0.209	0.209	0.197	0.207	0.203	0.209	0.199	0.182	0.091	0.067
JPEG	0.134	0.139	0.147	0.108	0.155	0.142	0.116	0.133	0.127	0.023	0.027
Noise	0.212	0.305	0.264	0.325	0.244	0.250	0.294	0.287	0.304	0.023	0.056
JP2K-RC	0.233	0.175	0.226	0.149	0.193	0.221	0.221	0.215	0.262	0.181	0.251
JP2K-NN	0.180	0.200	0.237	0.237	0.232	0.186	0.210	0.219	0.199	0.050	0.069

TABLE X
95% CONFIDENCE INTERVAL OF CORRECT DISTANT SIMILARITY SCORES FOR EACH QUALITY ESTIMATOR

	SSIM	VIF	VSNR	MIQE	PSNR	MSSIM	IWSSIM	MIQEC	CWSSIM	BIQI	BRISQUE
Blur	0.114	0.079	0.106	0.115	0.086	0.088	0.093	0.156	0.111	0.022	0.010
JP2K	0.072	0.084	0.084	0.079	0.083	0.081	0.084	0.080	0.073	0.036	0.027
JPEG	0.056	0.054	0.059	0.043	0.062	0.057	0.046	0.053	0.051	0.009	0.011
Noise	0.085	0.122	0.106	0.130	0.098	0.100	0.118	0.115	0.122	0.009	0.022
JP2K-RC	0.071	0.040	0.073	0.061	0.078	0.051	0.064	0.116	0.074	0.007	0.026
JP2K-NN	0.072	0.080	0.095	0.095	0.093	0.074	0.084	0.088	0.080	0.020	0.028

TABLE XI
CORRELATION COEFFICIENTS

	LIVE			TID			CSIQ			Average		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
SSIM	0.936	0.939	0.794	0.740	0.775	0.577	0.859	0.876	0.691	0.845	0.863	0.687
VIF	**0.972**	**0.972**	**0.857**	0.777	0.749	0.586	0.925	0.919	0.754	0.891	0.880	0.732
VSNR	0.952	0.939	0.785	0.232	0.705	0.535	0.801	0.811	0.625	0.662	0.818	0.648
MIQE	0.962	0.964	0.838	**0.840**	0.807	0.623	0.916	0.911	0.738	**0.906**	0.894	0.733
PSNR	0.847	0.893	0.727	0.279	0.553	0.403	0.800	0.806	0.608	0.642	0.751	0.579
MSSIM	0.947	0.951	0.818	0.790	0.854	0.657	0.898	0.913	0.739	0.878	0.906	0.738
IWSSIM	0.952	0.960	0.838	0.809	**0.856**	**0.664**	0.903	0.921	0.753	0.888	**0.913**	**0.751**
MIQEC	0.960	0.961	0.828	0.829	0.788	0.608	**0.926**	**0.930**	**0.765**	0.905	0.893	0.734
CWSSIM	0.862	0.894	0.721	0.334	0.528	0.378	0.640	0.674	0.491	0.612	0.699	0.53
BIQI	0.832	0.897	0.720	0.414	0.351	0.244	0.695	0.619	0.442	0.647	0.622	0.469
BRISQUE	0.928	0.941	0.789	0.408	0.322	0.228	0.739	0.556	0.423	0.692	0.606	0.48

V. Conclusions

In this paper, we have explored the quality estimation of images with resolutions different from that of the reference image. We have shown that conventional approaches have significant limitations to estimate this quality. We have proposed several ideas to overcome these limitations. We have developed an algorithm using these ideas to solve the problem. We have performed subjective tests to verify whether the proposed approach improves the results.

We analyze the subjective test results using different test cases. When the reference and test images have different resolutions, the subjective tests demonstrate that in most cases the proposed method works better than other approaches. In addition, the proposed algorithm performs well when the reference image and the test image have the same resolution. Similar ideas can be applied to video.

As a future work, we will focus on developing a non-dyadic QE to compute the quality of different resolution images. We will also prepare subjective tests to evaluate the performance of QE algorithms for different downsampling rates.

Acknowledgment

The authors would like to thank those who volunteered to take the subjective tests.

References

[1] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[2] E. Akyol, A. M. Tekalp, and M. R. Civanlar, "Content-aware scalability-type selection for rate adaptation of scalable video," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, 2007.

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

[4] 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, 2010.

[5] L. He, D. Tao, X. Li, and X. Gao, "Sparse representation for blind image quality assessment." in *CVPR*, 16 – 21 June 2012 2012, pp. 1146–1153.

[6] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[7] D. Wang, F. Speranza, A. Vincent, T. Martin, and P. Blanchfield, "Toward optimal rate control: A study of the impact of spatial resolution, frame rate, and quantization on subjective video quality and bit rate," in *SPIE Visual Communications and Image Processing*, vol. 5150, 2003, pp. 198–209.

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

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

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

[11] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *37th IEEE Asilomar Conference on Signals, Systems and Computers*, vol. 2, 2003, pp. 1398 – 1402.

[12] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.

[13] E. Reed and J. Lim, "Optimal multidimensional bit-rate control for video communication," *IEEE Trans. Image Proc.*, vol. 11, no. 8, pp. 873 – 885, Aug. 2002.

[14] S. H. Bae, T. N. Pappas, and B.-H. Juang, "Subjective evaluation of spatial resolution and quantization noise tradeoffs," *IEEE Trans. Image Proc.*, vol. 18, pp. 495–508, Mar. 2009.

[15] Y. Xue, Y.-F. Ou, Z. Ma, and Y. Wang, "Perceptual video quality assessment on a mobile platform considering both spatial resolution and quantization artifacts," in *Packet Video Workshop*, December 2010.

[16] Y.-F. Ou, Y. Xue, Z. Ma, and Y. Wang, "A perceptual video quality model for mobile platform considering impact of spatial, temporal, and amplitude resolutions," in *IEEE IVMSP Workshop: Perception and Visual Signal Analysis*, June 2011.

[17] G. Cermak, M. Pinson, and S. Wolf, "The relationship among video quality, screen resolution, and bit rate," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 258–262, 2011.

[18] C. Chan, S. Wee, and J. Apostolopoulos, "Multiple distortion measures for packetized scalable media," *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1671 –1686, Dec. 2008.

[19] A. M. Demirtas, H. Jafarkhani, and A. R. Reibman, "Quality estimation for images and video with different spatial resolutions," in *Human Vision and Electronic Imaging XVII*, vol. 8291, Feb. 2012.

[20] K.N.Ngan, K. Leong, and H. Singh, "Adaptive cosine transform coding of images in perceptual domain," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1743–1750, 1989.

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

[22] D. H. Kelly, "Motion and vision II: Stabilized spatio-temporal threshold surface," *J. Opt. Soc. Am.*, vol. 69, no. 10, pp. 1340–1349, 1979.

[23] A. B. Watson, "The cortex transform: Rapid computation of simulated neural images," *Comput. Vision Graph. Image Process.*, vol. 39, no. 3, pp. 311–327, Sep. 1987.

[24] S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital images and human vision*, A. B. Watson, Ed. Cambridge, MA, USA: MIT Press, 1993, pp. 179–206.

[25] A. Cohen, I. Daubechies, and J. C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 45, no. 5, pp. 485–560, June 1992.

[26] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *Journal of Mathematical Imaging and Vision*, vol. 18, pp. 17–33, 2003.

[27] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," *Applied and Computational Harmonic Analysis*, vol. 11, pp. 89–123, 2001.

[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.

[29] F. M. Ciaramello and A. R. Reibman, "Systematic stress testing of image quality estimators," in *ICIP*, 2011, pp. 3101–3104.

[30] D. H. Kelly, "Spatiotemporal variation of chromatic and achromatic contrast thresholds," *J. Opt. Soc. Am.*, vol. 73, no. 6, pp. 742–750, 1983.

[31] D. M. Rouse, R. Pépion, S. S. Hemami, and P. L. Callet, "Image utility assessment and a relationship with image quality assessment," in *Human Vision and Electronic Imaging*, 2009.

[32] J.-S. Lee, F. D. Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: Application to scalable video coding," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 882–893, 2011.

[33] C. A. Segall and G. J. Sullivan, "Spatial scalability within the H.264/AVC scalable video coding extension," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 17, no. 9, pp. 1121–1135, Sep 2007.

[34] A. R. Reibman, "A strategy to jointly test image quality estimators subjectively," in *Proc. International Conference on Image Processing*, 30 Sept–3 Oct. 2012.

[35] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: a new image quality index," *IEEE Trans. Image Proc.*, vol. 18, no. 11, pp. 2385–2401, 2009.

[36] L. L. Thurstone, "A law of comparative judgment," *Psychological Review*, vol. 34, pp. 273–286, 1927.

[37] R. A. Bradley and M. E. Terry, "The rank analysis of incomplete block designs — I. The method of paired comparisons," *Biometrika*, vol. 39, pp. 324–345, 1952.

[38] J. C. Handley, "Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment," in *Image Proc, Image Qual., and Image Capture Sys. Conf. (PICS'01)*, 2001, pp. 108–112.

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

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