Image Quality Assessment: From Error Visibility to Structural Similarity

Zhou Wang, *Member, IEEE,* Alan C. Bovik, *Fellow, IEEE*
Hamid R. Sheikh, *Student Member, IEEE,* and Eero P. Simoncelli, *Senior Member, IEEE*

Abstract— **Objective methods for assessing perceptual image quality have traditionally attempted to quantify the visibility of errors between a distorted image and a reference image using a variety of known properties of the human visual system. Under the assumption that human visual perception is highly adapted for extracting structural information from a scene, we introduce an alternative framework for quality assessment based on the degradation of structural information. As a specific example of this concept, we develop a Structural Similarity Index and demonstrate its promise through a set of intuitive examples, as well as comparison to both subjective ratings and state-of-the-art objective methods on a database of images compressed with JPEG and JPEG2000.**[1]

Keywords— **Error sensitivity, human visual system (HVS), image coding, image quality assessment, JPEG, JPEG2000, perceptual quality, structural information, structural similarity (SSIM).**

I. Introduction

Digital images are subject to a wide variety of distortions during acquisition, processing, compression, storage, transmission and reproduction, any of which may result in a degradation of visual quality. For applications in which images are ultimately to be viewed by human beings, the only "correct" method of quantifying visual image quality is through subjective evaluation. In practice, however, subjective evaluation is usually too inconvenient, time-consuming and expensive. The goal of research in *objective* image quality assessment is to develop quantitative measures that can automatically predict perceived image quality.

An objective image quality metric can play a variety of roles in image processing applications. First, it can be used to dynamically *monitor* and adjust image quality. For example, a network digital video server can examine the quality of video being transmitted in order to control and allocate streaming resources. Second, it can be used to *optimize* algorithms and parameter settings of image processing systems. For instance, in a visual communication system, a quality metric can assist in the optimal design of prefiltering and bit assignment algorithms at the encoder and of optimal reconstruction, error concealment and postfiltering algorithms at the decoder. Third, it can be used to *benchmark* image processing systems and algorithms.

Objective image quality metrics can be classified according to the availability of an original (distortion-free) image, with which the distorted image is to be compared. Most existing approaches are known as *full-reference*, meaning that a complete reference image is assumed to be known. In many practical applications, however, the reference image is not available, and a *no-reference* or "blind" quality assessment approach is desirable. In a third type of method, the reference image is only partially available, in the form of a set of extracted features made available as side information to help evaluate the quality of the distorted image. This is referred to as *reduced-reference* quality assessment. This paper focuses on full-reference image quality assessment.

The simplest and most widely used full-reference quality metric is the mean squared error (MSE), computed by averaging the squared intensity differences of distorted and reference image pixels, along with the related quantity of peak signal-to-noise ratio (PSNR). These are appealing because they are simple to calculate, have clear physical meanings, and are mathematically convenient in the context of optimization. But they are not very well matched to perceived visual quality (e.g., [1]–[9]). In the last three decades, a great deal of effort has gone into the development of quality assessment methods that take advantage of known characteristics of the human visual system (HVS). The majority of the proposed perceptual quality assessment models have followed a strategy of modifying the MSE measure so that errors are penalized in accordance with their visibility. Section II summarizes this type of error-sensitivity approach and discusses its difficulties and limitations. In Section III, we describe a new paradigm for quality assessment, based on the hypothesis that the HVS is highly adapted for extracting structural information. As a specific example, we develop a measure of structural similarity that compares local patterns of pixel intensities that have been normalized for luminance and contrast. In Section IV, we compare the test results of different quality assessment models against a large set of subjective ratings gathered for a database of 344 images compressed with JPEG and JPEG2000.

The work of Z. Wang and E. P. Simoncelli was supported by the Howard Hughes Medical Institute. The work of A. C. Bovik and H. R. Sheikh was supported by the National Science Foundation and the Texas Advanced Research Program. Z. Wang and E. P. Simoncelli are with the Howard Hughes Medical Institute, the Center for Neural Science and the Courant Institute for Mathematical Sciences, New York University, New York, NY 10012 USA (email: zhouwang@ieee.org; eero.simoncelli@nyu.edu). A. C. Bovik and H. R. Sheikh are with the Laboratory for Image and Video Engineering (LIVE), Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (email: bovik@ece.utexas.edu; hamid.sheikh@ieee.org).

[1] A MatLab implementation of the proposed algorithm is available online at `http://www.cns.nyu.edu/~lcv/ssim/`.

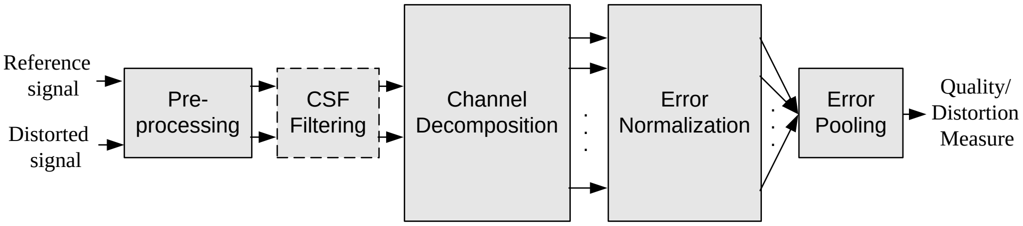


Fig. 1. A prototypical quality assessment system based on error sensitivity. Note that the CSF feature can be implemented either as a separate stage (as shown) or within "Error Normalization".

II. Image Quality Assessment Based on Error Sensitivity

An image signal whose quality is being evaluated can be thought of as a sum of an undistorted reference signal and an error signal. A widely adopted assumption is that the loss of perceptual quality is directly related to the visibility of the error signal. The simplest implementation of this concept is the MSE, which objectively quantifies the strength of the error signal. But two distorted images with the same MSE may have very different types of errors, some of which are much more visible than others. Most perceptual image quality assessment approaches proposed in the literature attempt to weight different aspects of the error signal according to their visibility, as determined by psychophysical measurements in humans or physiological measurements in animals. This approach was pioneered by Mannos and Sakrison [10], and has been extended by many other researchers over the years. Reviews on image and video quality assessment algorithms can be found in [4], [11]–[13].

A. Framework

Fig. 1 illustrates a generic image quality assessment framework based on error sensitivity. Most perceptual quality assessment models can be described with a similar diagram, although they differ in detail. The stages of the diagram are as follows:

Pre-processing. This stage typically performs a variety of basic operations to eliminate known distortions from the images being compared. First, the distorted and reference signals are properly scaled and aligned. Second, the signal might be transformed into a color space (e.g., [14]) that is more appropriate for the HVS. Third, quality assessment metrics may need to convert the digital pixel values stored in the computer memory into luminance values of pixels on the display device through pointwise nonlinear transformations. Fourth, a low-pass filter simulating the point spread function of the eye optics may be applied. Finally, the reference and the distorted images may be modified using a nonlinear point operation to simulate light adaptation.

CSF Filtering. The contrast sensitivity function (CSF) describes the sensitivity of the HVS to different spatial and temporal frequencies that are present in the visual stimulus. Some image quality metrics include a stage that weights the signal according to this function (typically implemented using a linear filter that approximates the frequency response of the CSF). However, many recent metrics choose to implement CSF as a base-sensitivity normalization factor after channel decomposition.

Channel Decomposition. The images are typically separated into subbands (commonly called "channels" in the psychophysics literature) that are selective for spatial and temporal frequency as well as orientation. While some quality assessment methods implement sophisticated channel decompositions that are believed to be closely related to the neural responses in the primary visual cortex [2], [15]–[19], many metrics use simpler transforms such as the discrete cosine transform (DCT) [20], [21] or separable wavelet transforms [22]–[24]. Channel decompositions tuned to various temporal frequencies have also been reported for video quality assessment [5], [25].

Error Normalization. The error (difference) between the decomposed reference and distorted signals in each channel is calculated and normalized according to a certain masking model, which takes into account the fact that the presence of one image component will decrease the visibility of another image component that is proximate in spatial or temporal location, spatial frequency, or orientation. The normalization mechanism weights the error signal in a channel by a space-varying visibility threshold [26]. The visibility threshold at each point is calculated based on the energy of the reference and/or distorted coefficients in a neighborhood (which may include coefficients from within a spatial neighborhood of the same channel as well as other channels) and the base-sensitivity for that channel. The normalization process is intended to convert the error into units of just noticeable difference (JND). Some methods also consider the effect of contrast response saturation (e.g., [2]).

Error Pooling. The final stage of all quality metrics must combine the normalized error signals over the spatial extent of the image, and across the different channels, into a single value. For most quality assessment methods, pooling takes the form of a Minkowski norm:

$$E\left(\{e_{l,k}\}\right) = \left(\sum_l \sum_k |e_{l,k}|^\beta\right)^{1/\beta} \tag{1}$$

where $e_{l,k}$ is the normalized error of the k-th coefficient in the l-th channel, and β is a constant exponent typically chosen to lie between 1 and 4. Minkowski pooling may be performed over space (index k) and then over frequency (index l), or vice-versa, with some non-linearity between them, or possibly with different exponents β. A spatial map indicating the relative importance of different regions may also be used to provide spatially variant weighting [25], [27], [28].

B. Limitations

The underlying principle of the error-sensitivity approach is that perceptual quality is best estimated by quantifying the visibility of errors. This is essentially accomplished by simulating the functional properties of early stages of the HVS, as characterized by both psychophysical and physiological experiments. Although this bottom-up approach to the problem has found nearly universal acceptance, it is important to recognize its limitations. In particular, the HVS is a complex and highly non-linear system, but most models of early vision are based on linear or quasi-linear operators that have been characterized using restricted and simplistic stimuli. Thus, error-sensitivity approaches must rely on a number of strong assumptions and generalizations. These have been noted by many previous authors, and we provide only a brief summary here.

The Quality Definition Problem. The most fundamental problem with the traditional approach is the definition of image quality. In particular, it is not clear that error visibility should be equated with loss of quality, as some distortions may be clearly visible but not so objectionable. An obvious example would be multiplication of the image intensities by a global scale factor. The study in [29] also suggested that the correlation between image fidelity and image quality is only moderate.

The Suprathreshold Problem. The psychophysical experiments that underlie many error sensitivity models are specifically designed to estimate the threshold at which a stimulus is just barely visible. These measured threshold values are then used to define visual error sensitivity measures, such as the CSF and various masking effects. However, very few psychophysical studies indicate whether such near-threshold models can be generalized to characterize perceptual distortions significantly larger than threshold levels, as is the case in a majority of image processing situations. In the suprathreshold range, can the relative visual distortions between different channels be normalized using the visibility thresholds? Recent efforts have been made to incorporate suprathreshold psychophysics for analyzing image distortions (e.g., [30]–[34]).

The Natural Image Complexity Problem. Most psychophysical experiments are conducted using relatively simple patterns, such as spots, bars, or sinusoidal gratings. For example, the CSF is typically obtained from threshold experiments using global sinusoidal images. The masking phenomena are usually characterized using a superposition of two (or perhaps a few) different patterns. But all such patterns are much simpler than real world images, which can be thought of as a superposition of a much larger number of simple patterns. Can the models for the interactions between a few simple patterns generalize to evaluate interactions between tens or hundreds of patterns? Is this limited number of simple-stimulus experiments sufficient to build a model that can predict the visual quality of complex-structured natural images? Although the answers to these questions are currently not known, the recently established Modelfest dataset [35] includes both simple and complex patterns, and should facilitate future studies.

The Decorrelation Problem. When one chooses to use a Minkowski metric for spatially pooling errors, one is implicitly assuming that errors at different locations are statistically independent. This would be true if the processing prior to the pooling eliminated dependencies in the input signals. Empirically, however, this is not the case for linear channel decomposition methods such as the wavelet transform. It has been shown that a strong dependency exists between intra- and inter-channel wavelet coefficients of natural images [36], [37]. In fact, state-of-the-art wavelet image compression techniques achieve their success by exploiting this strong dependency [38]–[41]. Psychophysically, various visual masking models have been used to account for the interactions between coefficients [2], [42]. Statistically, it has been shown that a well-designed nonlinear gain control model, in which parameters are optimized to reduce dependencies rather than for fitting data from masking experiments, can greatly reduce the dependencies of the transform coefficients [43], [44]. In [45], [46], it is shown that optimal design of transformation and masking models can reduce both statistical and perceptual dependencies. It remains to be seen how much these models can improve the performance of the current quality assessment algorithms.

The Cognitive Interaction Problem. It is widely known that cognitive understanding and interactive visual processing (e.g., eye movements) influence the perceived quality of images. For example, a human observer will give different quality scores to the same image if s/he is provided with different instructions [4], [30]. Prior information regarding the image content, or attention and fixation, may also affect the evaluation of the image quality [4], [47]. But most image quality metrics do not consider these effects, as they are difficult to quantify and not well understood.

III. Structural Similarity Based Image Quality Assessment

Natural image signals are highly structured: Their pixels exhibit strong dependencies, especially when they are spatially proximate, and these dependencies carry important information about the structure of the objects in the visual scene. The Minkowski error metric is based on pointwise signal differences, which are independent of the underlying signal structure. Although most quality measures based on error sensitivity decompose image signals using linear transformations, these do not remove the strong dependencies, as discussed in the previous section. The motivation of our new approach is to find a more direct way to compare the structures of the reference and the distorted signals.

A. New Philosophy

In [6] and [9], a new framework for the design of image quality measures was proposed, based on the assumption that the human visual system is highly adapted to extract structural information from the viewing field. It follows

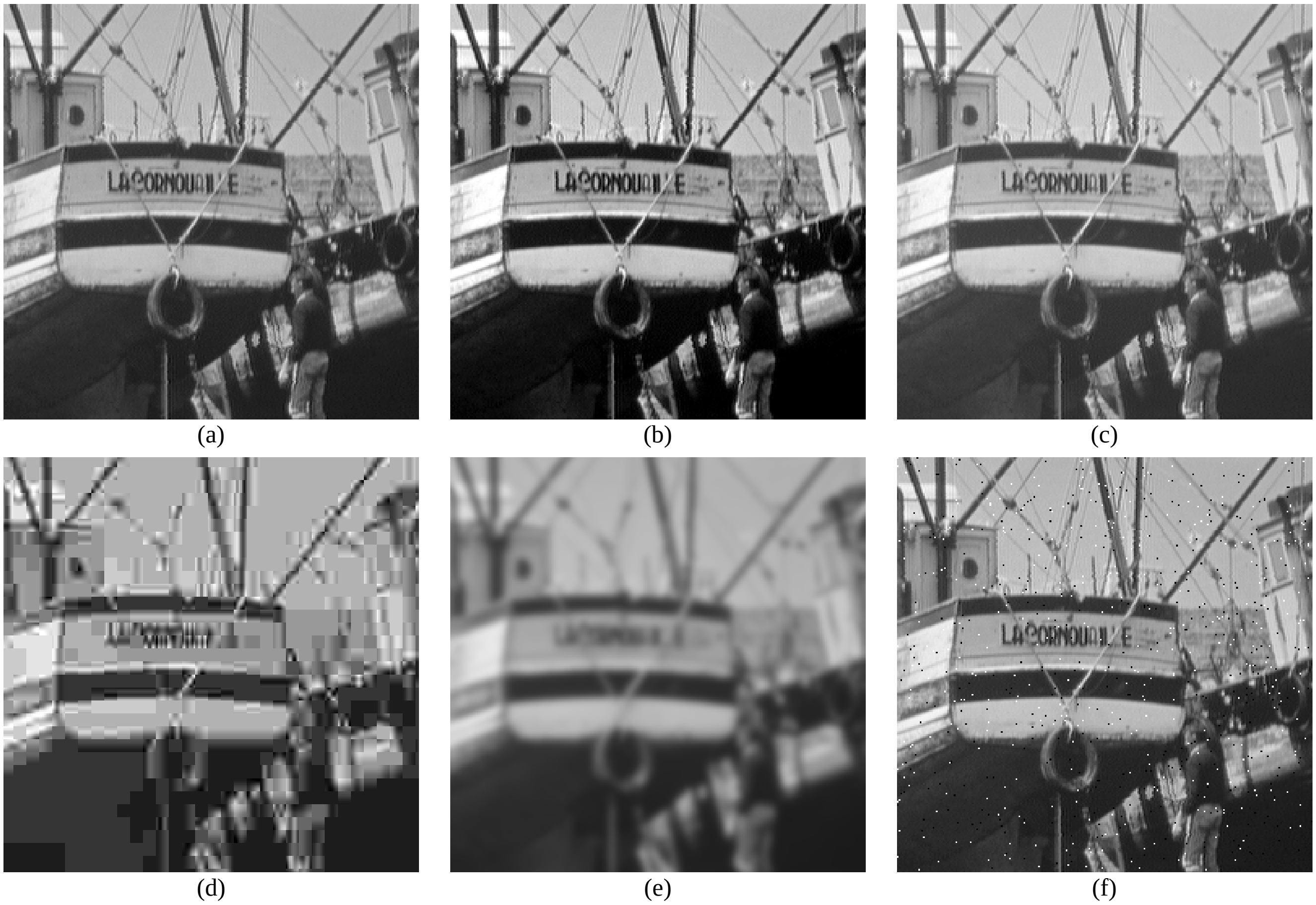


Fig. 2. Comparison of “Boat” images with different types of distortions, all with MSE = 210. (a) Original image (8bits/pixel; cropped from 512×512 to 256×256 for visibility); (b) Contrast stretched image, MSSIM = 0.9168; (c) Mean-shifted image, MSSIM = 0.9900; (d) JPEG compressed image, MSSIM = 0.6949; (e) Blurred image, MSSIM = 0.7052; (f) Salt-pepper impulsive noise contaminated image, MSSIM = 0.7748.

that a measure of structural information change can provide a good approximation to perceived image distortion.

This new philosophy can be best understood through comparison with the error sensitivity philosophy. First, the error sensitivity approach estimates *perceived errors* to quantify image degradations, while the new philosophy considers image degradations as *perceived changes in structural information.* A motivating example is shown in Fig. 2, where the original “Boat” image is altered with different distortions, each adjusted to yield nearly identical MSE relative to the original image. Despite this, the images can be seen to have drastically different perceptual quality. With the error sensitivity philosophy, it is difficult to explain why the contrast-stretched image has very high quality in consideration of the fact that its visual difference from the reference image is easily discerned. But it is easily understood with the new philosophy since nearly all the structural information of the reference image is preserved, in the sense that the original information can be nearly fully recovered via a simple pointwise inverse linear luminance transform (except perhaps for the very bright and dark regions where saturation occurs). On the other hand, some structural information from the original image is permanently lost in the JPEG compressed and the blurred images, and therefore they should be given lower quality scores than the contrast-stretched and mean-shifted images.

Second, the error-sensitivity paradigm is a *bottom-up* approach, simulating the function of relevant early-stage components in the HVS. The new paradigm is a *top-down* approach, mimicking the hypothesized functionality of the overall HVS. This, on the one hand, avoids the suprathreshold problem mentioned in the previous section because it does not rely on threshold psychophysics to quantify the perceived distortions. On the other hand, the cognitive interaction problem is also reduced to a certain extent because probing the structures of the objects being observed is thought of as the purpose of the entire process of visual observation, including high level and interactive processes.

Third, the problems of natural image complexity and decorrelation are also avoided to some extent because the new philosophy does not attempt to predict image quality by accumulating the errors associated with psychophysically understood simple patterns. Instead, the new philosophy proposes to evaluate the structural changes between two complex-structured signals directly.

B. The Structural SIMilarity (SSIM) Index

We construct a specific example of a structural similarity quality measure from the perspective of image formation. A previous instantiation of this approach was made in [6]–[8] and promising results on simple tests were achieved. In this paper, we generalize this algorithm, and provide a more extensive set of validation results.

The luminance of the surface of an object being observed is the product of the illumination and the reflectance, but the structures of the objects in the scene are independent of the illumination. Consequently, to explore the structural information in an image, we wish to separate the influence of the illumination. We define the structural information in an image as those attributes that represent the structure of objects in the scene, independent of the average luminance and contrast. Since luminance and contrast can vary across a scene, we use the *local* luminance and contrast for our definition.

The system diagram of the proposed quality assessment system is shown in Fig. 3. Suppose $\mathbf{x}$ and $\mathbf{y}$ are two non-negative image signals, which have been aligned with each other (e.g., spatial patches extracted from each image). If we consider one of the signals to have perfect quality, then the similarity measure can serve as a quantitative measurement of the quality of the second signal. The system separates the task of similarity measurement into three comparisons: luminance, contrast and structure. First, the luminance of each signal is compared. Assuming discrete signals, this is estimated as the mean intensity:

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \, . \qquad (2)$$

The luminance comparison function $l(\mathbf{x},\mathbf{y})$ is then a function of μ_x and μ_y.

Second, we remove the mean intensity from the signal. In discrete form, the resulting signal $\mathbf{x}-\mu_x$ corresponds to the projection of vector $\mathbf{x}$ onto the hyperplane defined by

$$\sum_{i=1}^{N} x_i = 0 \, . \qquad (3)$$

We use the standard deviation (the square root of variance) as an estimate of the signal contrast. An unbiased estimate in discrete form is given by

$$\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N} (x_i - \mu_x)^2 \right)^{1/2} \, . \qquad (4)$$

The contrast comparison $c(\mathbf{x},\mathbf{y})$ is then the comparison of σ_x and σ_y.

Third, the signal is normalized (divided) by its own standard deviation, so that the two signals being compared have unit standard deviation. The structure comparison $s(\mathbf{x},\mathbf{y})$ is conducted on these normalized signals $(\mathbf{x}-\mu_x)/\sigma_x$ and $(\mathbf{y}-\mu_y)/\sigma_y$.

Finally, the three components are combined to yield an overall similarity measure:

$$S(\mathbf{x},\mathbf{y}) = f(l(\mathbf{x},\mathbf{y}), c(\mathbf{x},\mathbf{y}), s(\mathbf{x},\mathbf{y})) \, . \qquad (5)$$

An important point is that the three components are relatively independent. For example, the change of luminance and/or contrast will not affect the structures of images.

In order to complete the definition of the similarity measure in Eq. (5), we need to define the three functions $l(\mathbf{x},\mathbf{y})$, $c(\mathbf{x},\mathbf{y})$, $s(\mathbf{x},\mathbf{y})$, as well as the combination function $f(\cdot)$. We also would like the similarity measure to satisfy the following conditions:

1. Symmetry: $S(\mathbf{x},\mathbf{y}) = S(\mathbf{y},\mathbf{x})$;
2. Boundedness: $S(\mathbf{x},\mathbf{y}) \leq 1$;
3. Unique maximum: $S(\mathbf{x},\mathbf{y}) = 1$ if and only if $\mathbf{x} = \mathbf{y}$ (in discrete representations, $x_i = y_i$ for all $i = 1, 2, \cdots, N$);

For luminance comparison, we define

$$l(\mathbf{x},\mathbf{y}) = \frac{2\,\mu_x\,\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \, . \qquad (6)$$

where the constant C_1 is included to avoid instability when $\mu_x^2 + \mu_y^2$ is very close to zero. Specifically, we choose

$$C_1 = (K_1\,L)^2 \, , \qquad (7)$$

where L is the dynamic range of the pixel values (255 for 8-bit grayscale images), and $K_1 \ll 1$ is a small constant. Similar considerations also apply to contrast comparison and structure comparison described later. Eq. (6) is easily seen to obey the three properties listed above.

Equation (6) is also qualitatively consistent with Weber's law, which has been widely used to model light adaptation (also called luminance masking) in the HVS. According to Weber's law, the magnitude of a just-noticeable luminance change ΔI is approximately proportional to the background luminance I for a wide range of luminance values. In other words, the HVS is sensitive to the *relative* luminance change, and not the absolute luminance change. Letting R represent the size of luminance change relative to background luminance, we rewrite the luminance of the distorted signal as $\mu_y = (1+R)\mu_x$. Substituting this into Eq. (6) gives

$$l(\mathbf{x},\mathbf{y}) = \frac{2(1+R)}{1 + (1+R)^2 + C_1/\mu_x^2} \, . \qquad (8)$$

If we assume C_1 is small enough (relative to μ_x^2) to be ignored, then $l(\mathbf{x},\mathbf{y})$ is a function only of R, qualitatively consistent with Weber's law.

The contrast comparison function takes a similar form:

$$c(\mathbf{x},\mathbf{y}) = \frac{2\,\sigma_x\,\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \, , \qquad (9)$$

where $C_2 = (K_2\,L)^2$, and $K_2 \ll 1$. This definition again satisfies the three properties listed above. An important feature of this function is that with the same amount of

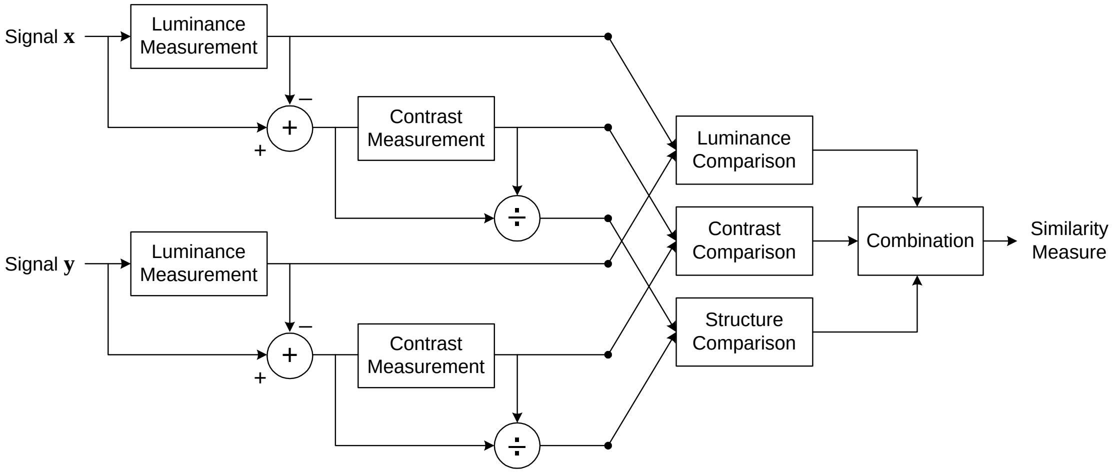


Fig. 3. Diagram of the structural similarity (SSIM) measurement system.

contrast change $\Delta\sigma = \sigma_y - \sigma_x$, this measure is less sensitive to the case of high base contrast σ_x than low base contrast. This is consistent with the contrast masking feature of the HVS.

Structure comparison is conducted after luminance subtraction and variance normalization. Specifically, we associate the two unit vectors $(\mathbf{x} - \mu_x)/\sigma_x$ and $(\mathbf{y} - \mu_y)/\sigma_y$, each lying in the hyperplane defined by Eq. (3), with the structure of the two images. The correlation (inner product) between these is a simple and effective measure to quantify the structural similarity. Notice that the correlation between $(\mathbf{x} - \mu_x)/\sigma_x$ and $(\mathbf{y} - \mu_y)/\sigma_y$ is equivalent to the correlation coefficient between $\mathbf{x}$ and $\mathbf{y}$. Thus, we define the structure comparison function as follows:

$$s(\mathbf{x},\mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\,\sigma_y + C_3}. \tag{10}$$

As in the luminance and contrast measures, we have introduced a small constant in both denominator and numerator. In discrete form, σ_{xy} can be estimated as:

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)\,. \tag{11}$$

Geometrically, the correlation coefficient corresponds to the cosine of the angle between the vectors $\mathbf{x} - \mu_x$ and $\mathbf{y} - \mu_y$. Note also that $s(\mathbf{x},\mathbf{y})$ can take on negative values.

Finally, we combine the three comparisons of Eqs. (6), (9) and (10) and name the resulting similarity measure the Structural SIMilarity (SSIM) index between signals $\mathbf{x}$ and $\mathbf{y}$:

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^{\alpha}\cdot[c(\mathbf{x},\mathbf{y})]^{\beta}\cdot[s(\mathbf{x},\mathbf{y})]^{\gamma}\,, \tag{12}$$

where $\alpha > 0$, $\beta > 0$ and $\gamma > 0$ are parameters used to adjust the relative importance of the three components. It is easy to verify that this definition satisfies the three conditions given above. In order to simplify the expression, we set $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$ in this paper. This results in a specific form of the SSIM index:

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\,\mu_x\,\mu_y + C_1)\,(2\,\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)\,(\sigma_x^2 + \sigma_y^2 + C_2)}\,. \tag{13}$$

The "universal quality index" (UQI) defined in [6], [7] corresponds to the special case that $C_1 = C_2 = 0$, which produces unstable results when either $(\mu_x^2 + \mu_y^2)$ or $(\sigma_x^2 + \sigma_y^2)$ is very close to zero.

The relationship between the SSIM index and more traditional quality metrics may be illustrated geometrically in a vector space of image components. These image components can be either pixel intensities or other extracted features such as transformed linear coefficients. Fig. 4 shows equal-distortion contours drawn around three different example reference vectors, each of which represents the local content of one reference image. For the purpose of illustration, we show only a two-dimensional space, but in general the dimensionality should match the number of image components being compared. Each contour represents a set of images with equal distortions relative to the enclosed reference image. Fig. 4(a) shows the result for a simple Minkowski metric. Each contour has the same size and shape (a circle here, as we are assuming an exponent of 2). That is, perceptual distance corresponds to Euclidean distance. Fig. 4(b) shows a Minkowski metric in which different image components are weighted differently. This could be, for example, weighting according to the CSF, as is common in many models. Here the contours are ellipses, but still are all the same size. These are shown aligned with the axes, but in general could be tilted to any fixed orientation.

Many recent models incorporate contrast masking behaviors, which has the effect of rescaling the equal-distortion contours according to the signal magnitude, as shown in Fig. 4(c). This may be viewed as a type of *adaptive* distortion metric: it depends not just on the difference between the signals, but also on the signals themselves. Fig. 4(d) shows a combination of contrast masking (magnitude weighting) followed by component weighting.

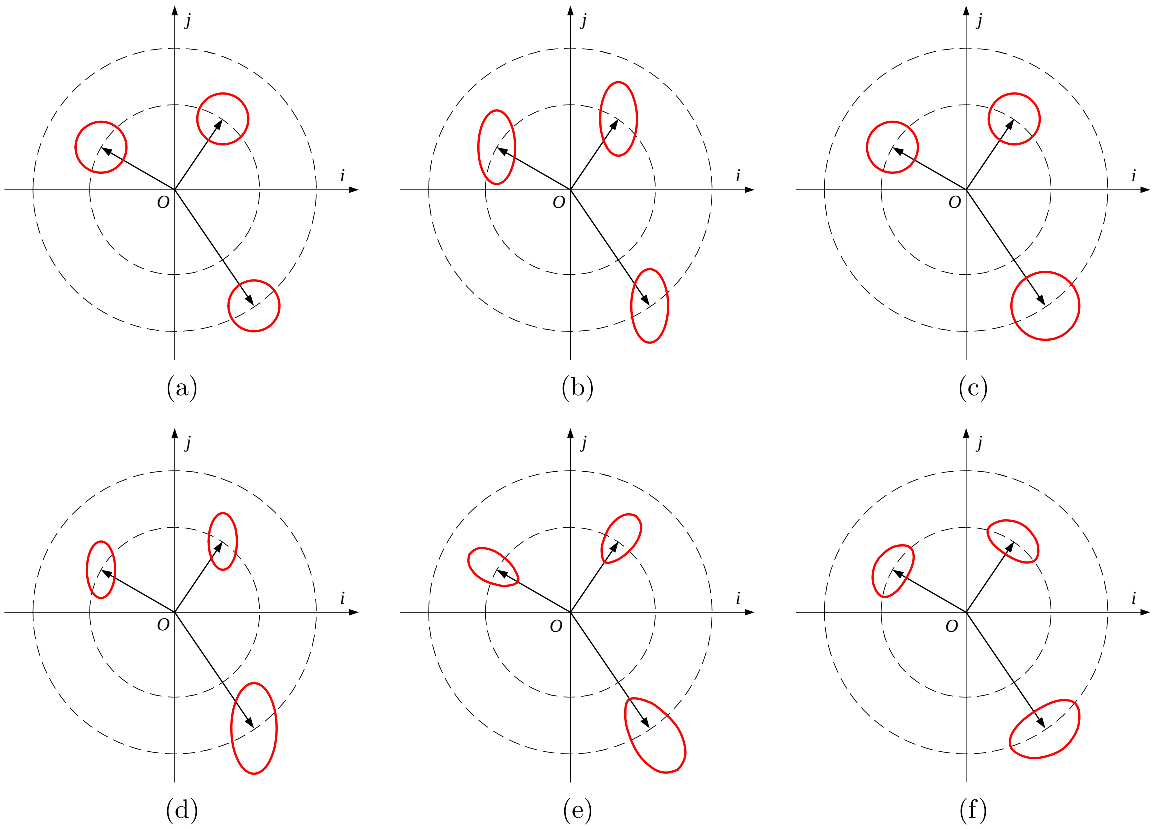


Fig. 4. Three example equal-distance contours for different quality metrics. (a) Minkowski error measurement systems; (b) component-weighted Minkowski error measurement systems; (c) magnitude-weighted Minkowski error measurement systems; (d) magnitude and component-weighted Minkowski error measurement systems; (e) the proposed system (a combination of Eqs. (9) and (10)) with more emphasis on $s(\mathbf{x}, \mathbf{y})$; (f) the proposed system (a combination of Eqs. (9) and (10)) with more emphasis on $c(\mathbf{x}, \mathbf{y})$. Each image is represented as a vector, whose entries are image components. Note: this is an illustration in 2-D space. In practice, the number of dimensions should be equal to the number of image components used for comparison (e.g, the number of pixels or transform coefficients).

Our proposed method, on the other hand, separately computes a comparison of two independent quantities: the vector lengths, and their angles. Thus, the contours will be aligned with the axes of a polar coordinate system. Figs. 4(e) and 4(f) show two examples of this, computed with different exponents. Again, this may be viewed as an *adaptive* distortion metric, but unlike previous models, both the size and the shape of the contours are adapted to the underlying signal. Some recent models that use divisive normalization to describe masking effects also exhibit signal-dependent contour orientations (e.g., [45], [46], [48]), although precise alignment with the axes of a polar coordinate system as in Figs. 4(e) and 4(f) is not observed in these methods.

C. Image Quality Assessment using SSIM index

For image quality assessment, it is useful to apply the SSIM index locally rather than globally. First, image statistical features are usually highly spatially non-stationary. Second, image distortions, which may or may not depend on the local image statistics, may also be space-variant. Third, at typical viewing distances, only a local area in the image can be perceived with high resolution by the human observer at one time instance (because of the foveation feature of the HVS [49], [50]). And finally, localized quality measurement can provide a spatially varying quality map of the image, which delivers more information about the quality degradation of the image and may be useful in some applications.

In [6], [7], the local statistics μ_x, σ_x and σ_{xy} are computed within a local 8×8 square window, which moves pixel-by-pixel over the entire image. At each step, the local statistics and SSIM index are calculated within the local window. One problem with this method is that the resulting SSIM index map often exhibits undesirable "blocking" artifacts. In this paper, we use an 11×11 circular-symmetric Gaussian weighting function $\mathbf{w} = \{ w_i \,|\, i = 1, 2, \cdots, N \}$, with standard deviation of 1.5 samples, normalized to unit sum ($\sum_{i=1}^{N} w_i = 1$). The estimates of local statistics μ_x, σ_x and σ_{xy} are then modified accordingly as

$$\mu_x = \sum_{i=1}^{N} w_i \, x_i \, . \tag{14}$$

$$\sigma_x = \left(\sum_{i=1}^{N} w_i \, (x_i - \mu_x)^2 \right)^{1/2} . \tag{15}$$

$$\sigma_{xy} = \sum_{i=1}^{N} w_i \, (x_i - \mu_x)(y_i - \mu_y) \, . \tag{16}$$

With such a windowing approach, the quality maps exhibit a locally isotropic property. Throughout this paper, the SSIM measure uses the following parameter settings: K_1= 0.01; K_2= 0.03. These values are somewhat arbitrary, but we find that in our current experiments, the performance of the SSIM index algorithm is fairly insensitive to variations of these values.

In practice, one usually requires a single overall quality measure of the entire image. We use a mean SSIM (MSSIM) index to evaluate the overall image quality:

$$\mathrm{MSSIM}(\mathbf{X},\mathbf{Y}) = \frac{1}{M}\sum_{j=1}^{M}\mathrm{SSIM}(\mathbf{x}_j,\mathbf{y}_j)\,, \tag{17}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are the reference and the distorted images, respectively; $\mathbf{x}_j$ and $\mathbf{y}_j$ are the image contents at the j-th local window; and M is the number of local windows in the image. Depending on the application, it is also possible to compute a weighted average of the different samples in the SSIM index map. For example, region-of-interest image processing systems may give different weights to different segmented regions in the image. As another example, it has been observed that different image textures attract human fixations with varying degrees (e.g., [51], [52]). A smoothly varying foveated weighting model (e.g., [50]) can be employed to define the weights. In this paper, however, we use uniform weighting. A MatLab implementation of the SSIM index algorithm is available online at [53].

IV. Experimental Results

Many image quality assessment algorithms have been shown to behave consistently when applied to distorted images created from the same original image, using the same type of distortions (e.g., JPEG compression). However, the effectiveness of these models degrades significantly when applied to a set of images originating from different reference images, and/or including a variety of different types of distortions. Thus, cross-image and cross-distortion tests are critical in evaluating the effectiveness of an image quality metric. It is impossible to show a thorough set of such examples, but the images in Fig. 2 provide an encouraging starting point for testing the cross-distortion capability of the quality assessment algorithms. The MSE and MSSIM measurement results are given in the figure caption. Obviously, MSE performs very poorly in this case. The MSSIM values exhibit much better consistency with the qualitative visual appearance.

A. Best-case/worst-case Validation

We also have developed a more efficient methodology for examining the relationship between our objective measure and perceived quality. Starting from a distorted image, we ascend/descend the gradient of MSSIM while constraining the MSE to remain equal to that of the initial distorted image. Specifically, we iterate the following two linear-algebraic steps:

$$(1)\quad \mathbf{Y} \;\rightarrow\; \mathbf{Y} \pm \lambda\, P(\mathbf{X},\mathbf{Y})\,\vec{\nabla}_{\mathbf{Y}}\mathrm{MSSIM}(\mathbf{X},\mathbf{Y})$$

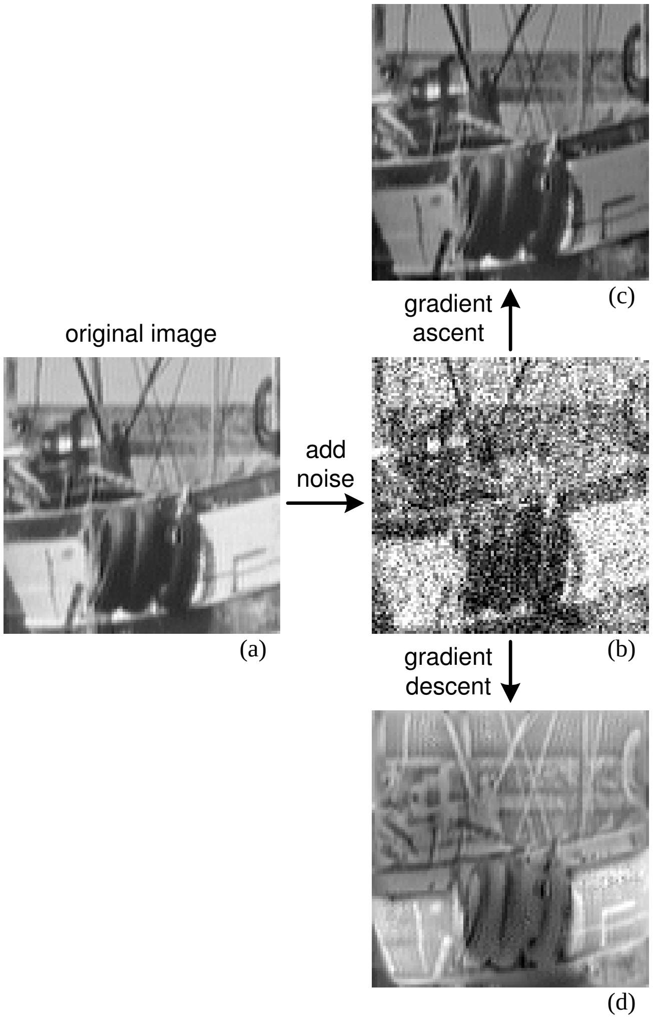


Fig. 5. Best- and worst-case MSSIM images, with identical MSE. These are computed by gradient ascent/descent iterative search on MSSIM measure, under the constraint of MSE = 2500. (a) Original image (100×100, 8bits/pixel, cropped from the "Boat" image); (b) Initial image, contaminated with Gaussian white noise (MSSIM = 0.3021); (c) Maximum MSSIM image (MSSIM = 0.9337); (d) Minimum MSSIM image (MSSIM = −0.5411).

$$(2)\quad \mathbf{Y} \;\rightarrow\; \mathbf{X} + \sigma\,\hat{E}(\mathbf{X},\mathbf{Y})$$

where σ is the square root of the constrained MSE, λ controls the step size, $\hat{E}(\mathbf{X},\mathbf{Y})$ is a unit vector defined by

$$\hat{E}(\mathbf{X},\mathbf{Y}) = \frac{\mathbf{Y}-\mathbf{X}}{||\mathbf{Y}-\mathbf{X}||},$$

and $P(\mathbf{X},\mathbf{Y})$ is a projection operator:

$$P(\mathbf{X},\mathbf{Y}) = \mathbf{I} - \hat{E}(\mathbf{X},\mathbf{Y})\,\hat{E}^T(\mathbf{X},\mathbf{Y}),$$

with $\mathbf{I}$ the identity operator. MSSIM is differentiable and this procedure converges to a local maximum/minimum of the objective measure. Visual inspection of these best- and worst-case images, along with the initial distorted image, provides a visual indication of the types of distortion deemed least/most important by the objective measure.

Therefore, it is an expedient and direct method for revealing perceptual implications of the quality measure. An example is shown in Fig. 5, where the initial image is contaminated with Gaussian white noise. It can be seen that the local structures of the original image are very well preserved in the maximal MSSIM image. On the other hand, the image structures are changed dramatically in the worst-case MSSIM image, in some cases reversing contrast.

B. Test on JPEG and JPEG2000 Image Database

We compare the cross-distortion and cross-image performances of different quality assessment models on an image database composed of JPEG and JPEG2000 compressed images. Twenty-nine high-resolution 24 bits/pixel RGB color images (typically 768×512 or similar size) were compressed at a range of quality levels using either JPEG or JPEG2000, producing a total of 175 JPEG images and 169 JPEG2000 images. The bit rates were in the range of 0.150 to 3.336 and 0.028 to 3.150 bits/pixel, respectively, and were chosen non-uniformly such that the resulting distribution of subjective quality scores was approximately uniform over the entire range. Subjects viewed the images from comfortable seating distances (this distance was only moderately controlled, to allow the data to reflect natural viewing conditions), and were asked to provide their perception of quality on a continuous linear scale that was divided into five equal regions marked with adjectives “Bad”, “Poor”, “Fair”, “Good” and “Excellent”. Each JPEG and JPEG2000 compressed image was viewed by 13 $\sim$ 20 subjects and 25 subjects, respectively. The subjects were mostly male college students.

Raw scores for each subject were normalized by the mean and variance of scores for that subject (i.e., raw values were converted to Z-scores [54]) and then the entire data set was rescaled to fill the range from 1 to 100. Mean opinion scores (MOSs) were then computed for each image, after removing outliers (most subjects had no outliers). The average standard deviations (for each image) of the subjective scores for JPEG, JPEG2000 and all images were 6.00, 7.33 and 6.65, respectively. The image database, together with the subjective score and standard deviation for each image, has been made available on the Internet at [55].

The luminance component of each JPEG and JPEG2000 compressed image is averaged over local 2×2 window and downsampled by a factor of 2 before the MSSIM value is calculated. Our experiments with the current dataset show that the use of the other color components does not significantly change the performance of the model, though this should not be considered generally true for color image quality assessment. Unlike many other perceptual image quality assessment approaches, no specific training procedure is employed before applying the proposed algorithm to the database, because the proposed method is intended for general-purpose image quality assessment (as opposed to image compression alone).

Figs. 6 and 7 show some example images from the database at different quality levels, together with their SSIM index maps and absolute error maps. Note that at low bit rate, the coarse quantization in JPEG and JPEG2000 algorithms often results in smooth representations of fine-detail regions in the image (e.g., the tiles in Fig.6(d) and the trees in Fig.7(d)). Compared with other types of regions, these regions may not be worse in terms of pointwise difference measures such as the absolute error. However, since the structural information of the image details are nearly completely lost, they exhibit poorer visual quality. Comparing Fig. 6(g) with Fig. 6(j), and Fig. 7(g) with 6(j)), we observe that the SSIM index is better in capturing such poor quality regions. Also notice that for images with intensive strong edge structures such as Fig. 7(c), it is difficult to reduce the pointwise errors in the compressed image, even at relatively high bit rate, as exemplified by Fig. 7(l). However, the compressed image supplies acceptable perceived quality as shown in Fig. 7(f). In fact, although the visual quality of Fig. 7(f) is better than Fig. 7(e), its absolute error map Fig. 7(l) appears to be worse than Fig. 7(k), as is confirmed by their PSNR values. The SSIM index maps, Figs. 7(h) and 7(i), deliver better consistency with perceived quality measurement.

The quality assessment models used for comparison include PSNR, the well-known Sarnoff model [56][2], UQI [7] and MSSIM. The scatter plot of MOS versus model prediction for each model is shown in Fig. 8. If PSNR is considered as a benchmark method to evaluate the effectiveness of the other image quality metrics, the Sarnoff model performs quite well in this test. This is in contrast with previous published test results (e.g., [57], [58]), where the performance of most models (including the Sarnoff model) were reported to be statistically equivalent to root mean squared error [57] and PSNR [58]. The UQI method performs much better than MSE for the simple cross-distortion test in [7], [8], but does not deliver satisfactory results in Fig. 8. We think the major reason is that at nearly flat regions, the denominator of the contrast comparison formula is close to zero, which makes the algorithm unstable. By inserting the small constants C_1 and C_2, MSSIM completely avoids this problem and the scatter slot demonstrates that it supplies remarkably good prediction of the subjective scores.

In order to provide quantitative measures on the performance of the objective quality assessment models, we follow the performance evaluation procedures employed in the video quality experts group (VQEG) Phase I FR-TV test [58], where four evaluation metrics were used. First, logistic functions are used in a fitting procedure to provide a non-linear mapping between the objective/subjective scores. The fitted curves are shown in Fig. 8. In [58], Metric 1 is the correlation coefficient between objective/subjective scores after variance-weighted regression analysis. Metric 2 is the correlation coefficient between objective/subjective scores after non-linear regression analysis. These two metrics combined, provide an evaluation of *prediction accuracy*. The third metric is the Spearman rank-order correlation co-

[2] The JNDmetrix software available online from the Sarnoff Corporation, at `http://www.sarnoff.com/products_services/video_vision/jndmetrix/`.

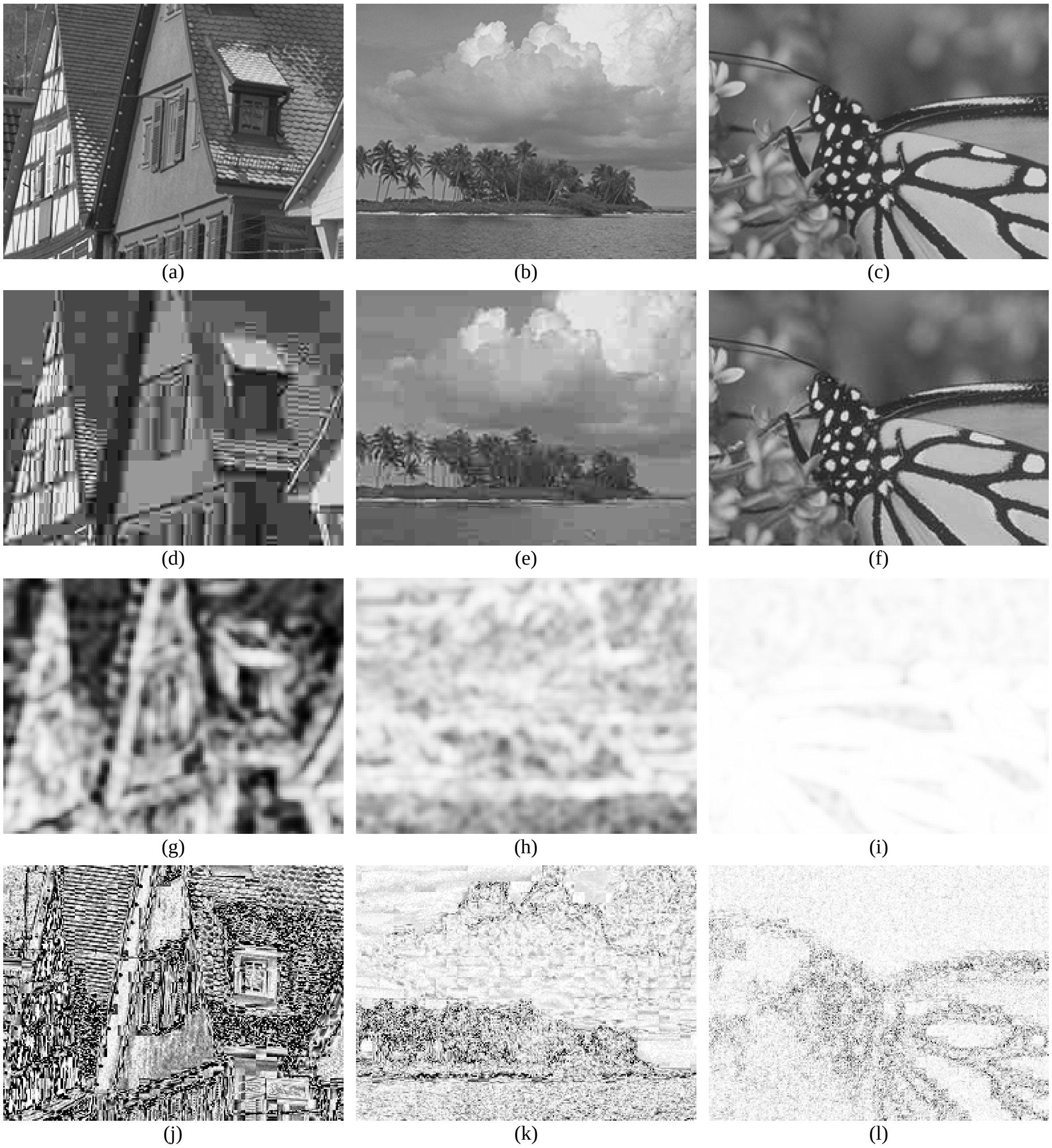


Fig. 6. Sample JPEG images compressed to different quality levels (original size: 768×512; cropped to 256×192 for visibility). (a), (b) and (c) are the original "Buildings", "Ocean" and "Monarch" images, respectively. (d) Compressed to 0.2673 bits/pixel, PSNR = 21.98dB, MSSIM = 0.7118; (e) Compressed to 0.2980 bits/pixel, PSNR = 30.87dB, MSSIM = 0.8886; (f) Compressed to 0.7755 bits/pixel, PSNR = 36.78dB, MSSIM = 0.9898. (g), (h) and (i) show SSIM maps of the compressed images, where brightness indicates the magnitude of the local SSIM index (squared for visibility). (j), (k) and (l) show absolute error maps of the compressed images (contrast-inverted for easier comparison to the SSIM maps).

efficient between the objective/subjective scores. It is considered as a measure of *prediction monotonicity*. Finally, metric 4 is the outlier ratio (percentage of the number of predictions outside the range of ±2 times of the standard deviations) of the predictions after the non-linear mapping, which is a measure of *prediction consistency*. More details on these metrics can be found in [58]. In addition to these, we also calculated the mean absolute prediction error (MAE), and root mean square prediction error (RMS) after non-linear regression, and weighted mean absolute prediction error (WMAE) and weighted root mean square prediction error (WRMS) after variance-weighted regression. The evaluation results for all the models being compared are given in Table I. For every one of these criteria, MSSIM performs better than all of the other models being compared.

V. Discussion

In this paper, we have summarized the traditional approach to image quality assessment based on error-sensitivity, and have enumerated its limitations. We have proposed the use of structural similarity as an alternative

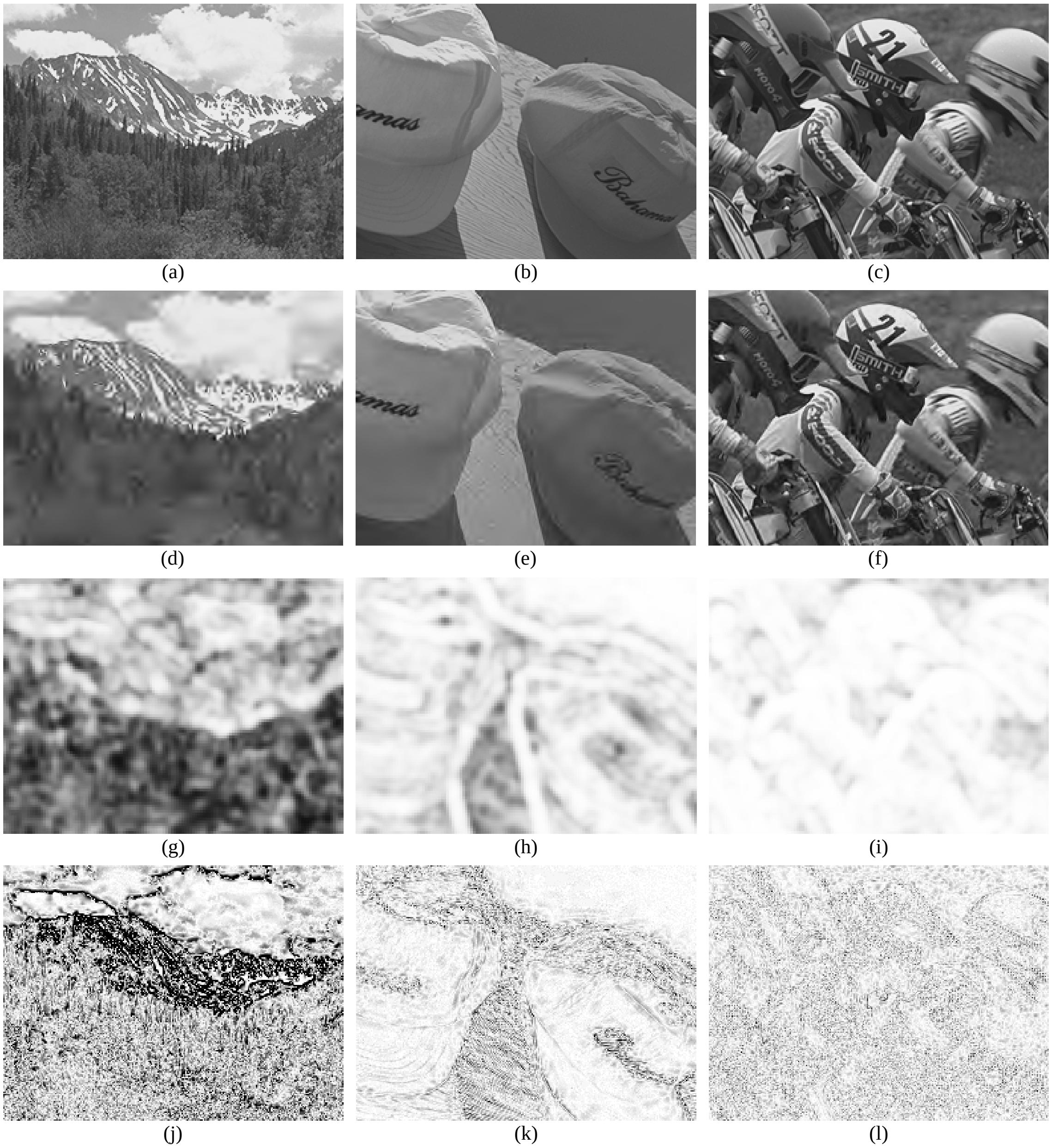
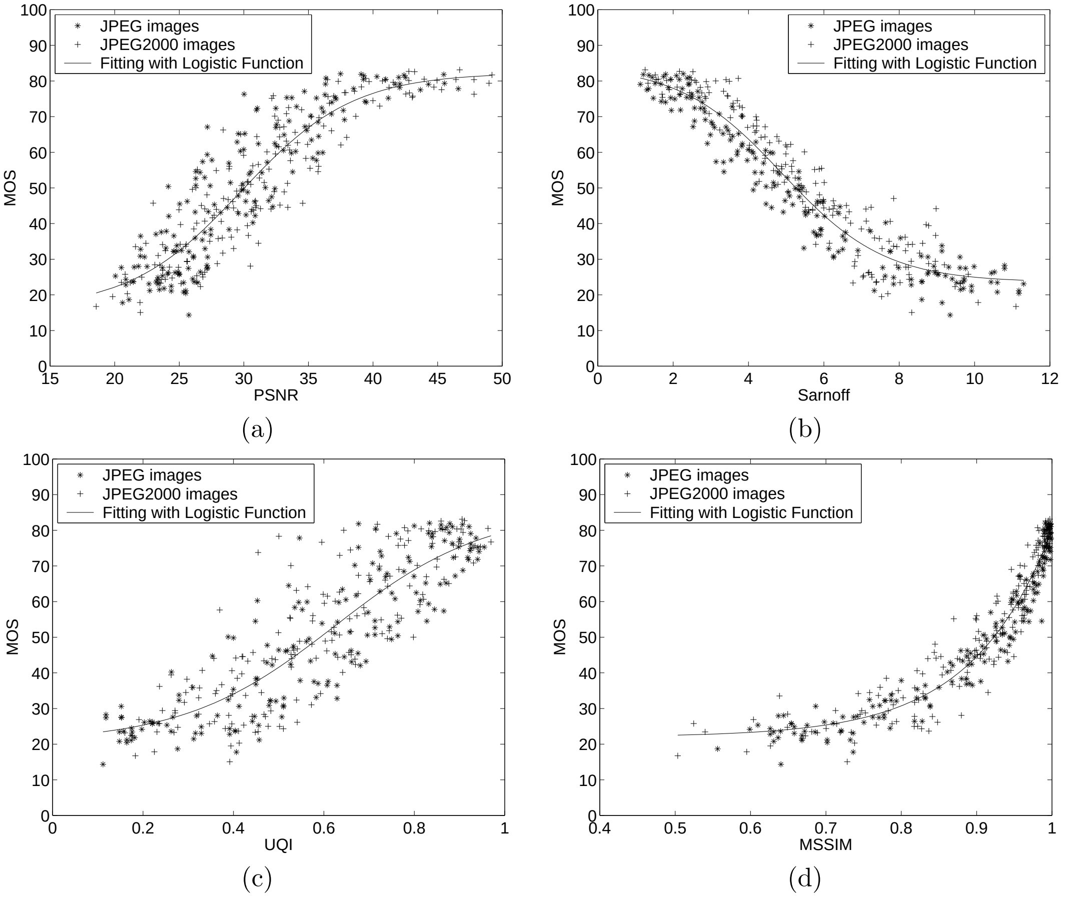


Fig. 7. Sample JPEG2000 images compressed to different quality levels (original size: 768×512; cropped to 256×192 for visibility). (a), (b) and (c) are the original "Stream", "Caps" and "Bikes" images, respectively. (d) Compressed to 0.1896 bits/pixel, PSNR = 23.46dB, MSSIM = 0.7339; (e) Compressed to 0.1982 bits/pixel, PSNR = 34.56dB, MSSIM = 0.9409; (f) Compressed to 1.1454 bits/pixel, PSNR = 33.47dB, MSSIM = 0.9747. (g), (h) and (i) show SSIM maps of the compressed images, where brightness indicates the magnitude of the local SSIM index (squared for visibility). (j), (k) and (l) show absolute error maps of the compressed images (contrast-inverted for easier comparison to the SSIM maps).

motivating principle for the design of image quality measures. To demonstrate our structural similarity concept, we developed an SSIM index and showed that it compares favorably with other methods in accounting for our experimental measurements of subjective quality of 344 JPEG and JPEG2000 compressed images.

Although the proposed SSIM index method is motivated from substantially different design principles, we see it as complementary to the traditional approach. Careful analysis shows that both the SSIM index and several recently developed divisive-normalization based masking models exhibit input-dependent behavior in measuring signal distortions [45], [46], [48]. It seems possible that the two approaches may eventually converge to similar solutions.

There are a number of issues that are worth investigation with regard to the specific SSIM index of Eq. (12). First, the optimization of the SSIM index for various image processing algorithms needs to be studied. For example, it may be employed for rate-distortion optimizations in the design of image compression algorithms. This is not an easy task since Eq. (12) is mathematically more cumbersome than MSE. Second, the application scope of the SSIM



(a)

(b)

(c)

(d)

Fig. 8. Scatter plots of subjective mean opinion score (MOS) versus model prediction. Each sample point represents one test image. (a) PSNR; (b) Sarnoff model [56]; (c) UQI [7] (equivalent to MSSIM with square window and $K_1 = K_2 = 0$); (d) MSSIM (Gaussian window, $K_1 = 0.01, K_2 = 0.03$).

TABLE I

Performance comparison of image quality assessment models. CC: correlation coefficient; MAE: mean absolute error; RMS: root mean squared error; OR: outlier ratio; WMAE: weighted mean absolute error; WRMS: weighted root mean squared error; SROCC: Spearman rank-order correlation coefficient

	Non-linear Regression				Variance-weighted Regression				Rank-order
Model	CC	MAE	RMS	OR	CC	WMAE	WRMS	OR	SROCC
PSNR	0.905	6.53	8.45	0.157	0.903	6.18	8.26	0.140	0.901
Sarnoff	0.956	4.66	5.81	0.064	0.956	4.42	5.62	0.061	0.947
UQI	0.866	7.76	9.90	0.189	0.861	7.64	9.79	0.195	0.863
MSSIM	0.967	3.95	5.06	0.041	0.967	3.79	4.87	0.041	0.963

index may not be restricted to image processing. In fact, because it is a symmetric measure, it can be thought of as a similarity measure for comparing any two signals. The signals can be either discrete or continuous, and can live in a space of arbitrary dimensionality.

We consider the proposed SSIM indexing approach as a particular implementation of the philosophy of structural similarity, from an image formation point of view. Under the same philosophy, other approaches may emerge that could be significantly different from the proposed SSIM indexing algorithm. Creative investigation of the concepts of structural information and structural distortion are likely to drive the success of these innovations.

VI. Acknowledgement

The authors would like to thank Dr. Jesus Malo and Dr. L. Lu for insightful comments, Dr. Jeffrey Lubin and Dr. Douglas Dixon for providing the Sarnoff JNDmetrix software, Dr. Philip Corriveau and Dr. John Libert for supplying the MatLab routines used in VQEG Phase I FR-TV test for the regression analysis of subjective/objective data comparison, and Visual Delights, Inc. for allowing the authors to use their images for subjective experiments.

References

[1] B. Girod, "What's wrong with mean-squared error," in *Digital Images and Human Vision* (A. B. Watson, ed.), pp. 207–220, the MIT press, 1993.

[2] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. SPIE*, vol. 2179, pp. 127–141, 1994.

[3] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Communications*, vol. 43, pp. 2959–2965, Dec. 1995.

[4] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Processing*, vol. 70, pp. 177–200, Nov. 1998.

[5] S. Winkler, "A perceptual distortion metric for digital color video," *Proc. SPIE*, vol. 3644, pp. 175–184, 1999.

[6] Z. Wang, *Rate scalable foveated image and video communications.* PhD thesis, Dept. of ECE, The University of Texas at Austin, Dec. 2001.

[7] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, pp. 81–84, Mar. 2002.

[8] Z. Wang, "Demo images and free software for 'a universal image quality index'," `http://anchovy.ece.utexas.edu/~zwang/research/quality_index/demo.html`.

[9] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 4, (Orlando), pp. 3313–3316, May 2002.

[10] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Information Theory*, vol. 4, pp. 525–536, 1974.

[11] T. N. Pappas and R. J. Safranek, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Proc.* (A. Bovik, ed.), Academic Press, 2000.

[12] Z. Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Applications* (B. Furht and O. Marques, eds.), CRC Press, 2003.

[13] S. Winkler, "Issues in vision modeling for perceptual video quality assessment," *Signal Processing*, vol. 78, pp. 231–252, 1999.

[14] A. B. Poirson and B. A. Wandell, "Appearance of colored patterns: pattern-color separability," *Journal of Optical Society of America A: Optics and Image Science*, vol. 10, no. 12, pp. 2458–2470, 1993.

[15] A. B. Watson, "The cortex transform: rapid computation of simulated neural images," *Computer Vision, Graphics, and Image Processing*, vol. 39, pp. 311–327, 1987.

[16] S. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity," in *Digital images and human vision* (A. B. Watson, ed.), pp. 179–206, Cambridge, Massachusetts: The MIT Press, 1993.

[17] J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital images and human vision* (A. B. Watson, ed.), pp. 163–178, Cambridge, Massachusetts: The MIT Press, 1993.

[18] D. J. Heeger and P. C. Teo, "A model of perceptual image fidelity," in *Proc. IEEE Int. Conf. Image Proc.*, pp. 343–345, 1995.

[19] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. Information Theory*, vol. 38, pp. 587–607, 1992.

[20] A. B. Watson, "DCT quantization matrices visually optimized for individual images," in *Proc. SPIE*, vol. 1913, pp. 202–216, 1993.

[21] A. B. Watson, J. Hu, and J. F. III. McGowan, "DVQ: A digital video quality metric based on human vision," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 20–29, 2001.

[22] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Processing*, vol. 6, pp. 1164–1175, Aug. 1997.

[23] A. P. Bradley, "A wavelet visible difference predictor," *IEEE Trans. Image Processing*, vol. 5, pp. 717–730, May 1999.

[24] Y. K. Lai and C.-C. J. Kuo, "A Haar wavelet approach to compressed image quality measurement," *Journal of Visual Communication and Image Representation*, vol. 11, pp. 17–40, Mar. 2000.

[25] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatio-temporal model of the human visual system," in *Proc. SPIE*, vol. 2668, pp. 450–461, 1996.

[26] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," *Journal of Optical Society of America*, vol. 14, no. 9, pp. 2379–2391, 1997.

[27] W. Xu and G. Hauske, "Picture quality evaluation based on error segmentation," *Proc. SPIE*, vol. 2308, pp. 1454–1465, 1994.

[28] W. Osberger, N. Bergmann, and A. Maeder, "An automatic image quality assessment technique incorporating high level perceptual factors," in *Proc. IEEE Int. Conf. Image Proc.*, pp. 414–418, 1998.

[29] D. A. Silverstein and J. E. Farrell, "The relationship between image fidelity and image quality," in *Proc. IEEE Int. Conf. Image Proc.*, pp. 881–884, 1996.

[30] D. R. Fuhrmann, J. A. Baro, and J. R. Cox Jr., "Experimental evaluation of psychophysical distortion metrics for JPEG-encoded images," *Journal of Electronic Imaging*, vol. 4, pp. 397–406, Oct. 1995.

[31] A. B. Watson and L. Kreslake, "Measurement of visual impairment scales for digital video," in *Human Vision, Visual Processing, and Digital Display, Proc. SPIE*, vol. 4299, 2001.

[32] J. G. Ramos and S. S. Hemami, "Suprathreshold wavelet coefficient quantization in complex stimuli: psychophysical evaluation and analysis," *Journal of the Optical Society of America A*, vol. 18, pp. 2385–2397, 2001.

[33] D. M. Chandler and S. S. Hemami, "Additivity models for suprathreshold distortion in quantized wavelet-coded images," in *Human Vision and Electronic Imaging VII, Proc. SPIE*, vol. 4662, Jan. 2002.

[34] J. Xing, "An image processing model of contrast perception and discrimination of the human visual system," in *SID Conference*, (Boston), May 2002.

[35] A. B. Watson, "Visual detection of spatial contrast patterns: Evaluation of five simple models," *Optics Express*, vol. 6, pp. 12–33, Jan. 2000.

[36] E. P. Simoncelli, "Statistical models for images: Compression, restoration and synthesis," in *Proc 31st Asilomar Conf on Signals, Systems and Computers*, (Pacific Grove, CA), pp. 673–678, IEEE Computer Society, November 1997.

[37] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Trans. Image Processing*, vol. 10, pp. 1647–1658, Nov. 2001.

[38] J. M. Shapiro, "Embedded image coding using zerotrees of wavelets coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.

[39] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 6, pp. 243–250, June 1996.

[40] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans. Image Processing*, vol. 8, pp. 1688–1701, December 1999.

[41] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards, and Practice.* Kluwer Academic Publishers, 2001.

[42] J. M. Foley and G. M. Boynton, "A new model of human luminance pattern vision mechanisms: Analysis of the effects of pattern orientation, spatial phase, and temporal frequency," in *Computational Vision Based on Neurobiology, Proc. SPIE* (T. A. Lawton, ed.), vol. 2054, 1994.

[43] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," *Nature: Neuroscience*, vol. 4, pp. 819–825, Aug. 2001.

[44] M. J. Wainwright, O. Schwartz, and E. P. Simoncelli, "Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons," in *Probabilistic Models of the Brain: Perception and Neural Function* (R. Rao, B. Olshausen, and M. Lewicki, eds.), MIT Press, 2002.

[45] J. Malo, R. Navarro, I. Epifanio, F. Ferri, and J. M. Artigas, "Non-linear invertible representation for joint statistical and perceptual feature decorrelation," *Lecture Notes on Computer Science*, vol. 1876, pp. 658–667, 2000.

[46] I. Epifanio, J. Gutiérrez, and J. Malo, "Linear transform for simultaneous diagonalization of covariance and perceptual metric matrix in image coding," *Pattern Recognition*, vol. 36, pp. 1799–1811, Aug. 2003.

[47] W. F. Good, G. S. Maitz, and D. Gur, "Joint photographic experts group (JPEG) compatible data compression of mammo-

grams," *Journal of Digital Imaging*, vol. 7, no. 3, pp. 123–132, 1994.

[48] A. Pons, J. Malo, J. M. Artigas, and P. Capilla, "Image quality metric based on multidimensional contrast perception models," *Displays*, vol. 20, pp. 93–110, 1999.

[49] W. S. Geisler and M. S. Banks, "Visual performance," in *Handbook of Optics* (M. Bass, ed.), McGraw-Hill, 1995.

[50] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Trans. Image Processing*, vol. 10, pp. 1397–1410, Oct. 2001.

[51] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, pp. 970–982, Sept. 2000.

[52] U. Rajashekar, L. K. Cormack, and A. C. Bovik, "Image features that draw fixations," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 3, pp. 313–316, Sept. 2003.

[53] Z. Wang, "The SSIM index for image quality assessment," `http://www.cns.nyu.edu/~lcv/ssim/`.

[54] A. M. van Dijk, J. B. Martens, and A. B. Watson, "Quality assessment of coded images using numerical category scaling," in *Proc. SPIE*, vol. 2451, 1995.

[55] H. R. Sheikh, Z. Wang, A. C. Bovik, and L. K. Cormack, "Image and video quality assessment research at LIVE," `http://live.ece.utexas.edu/research/quality/`.

[56] J. Lubin, "A visual discrimination mode for image system design and evaluation," in *Visual Models for Target Detection and Recognition* (E. Peli, ed.), pp. 245–283, Singapore: World Scientific Publishers, 1995.

[57] J.-B. Martens and L. Meesters, "Image dissimilarity," *Signal Processing*, vol. 70, pp. 155–176, Nov. 1998.

[58] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," Mar. 2000. `http://www.vqeg.org/`.

Zhou Wang (S'97-A'01-M'02) received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 1993, the M.S. degree from South China University of Technology, Guangzhou, China, in 1995, and the Ph.D. degree from The University of Texas at Austin in 2001.

He is currently a Research Associate at Howard Hughes Medical Institute and Laboratory for Computational Vision at New York University. Previously, he was a Research Engineer at AutoQuant Imaging, Inc., Watervliet, NY. From 1998 to 2001, he was a Research Assistant at the Laboratory for Image and Video Engineering at The University of Texas at Austin. In the summers of 2000 and 2001, he was with Multimedia Technologies, IBM T. J. Watson Research Center, Yorktown Heights, NY. He worked as a Research Assistant in periods during 1996 to 1998 at the Department of Computer Science, City University of Hong Kong, China. His current research interests include digital image and video coding, processing and quality assessment, and computational vision.

Alan Conrad Bovik (S'81-M'81-SM'89-F'96) is currently the Cullen Trust for Higher Education Endowed Professor in the Department of Electrical and Computer Engineering at the University of Texas at Austin, where he is the Director of the Laboratory for Image and Video Engineering (LIVE) in the Center for Perceptual Systems. During the Spring of 1992, he held a visiting position in the Division of Applied Sciences, Harvard University, Cambridge, Massachusetts. His current research interests include digital video, image processing, and computational aspects of biological visual perception. He has published nearly 400 technical articles in these areas and holds two U.S. patents. He is also the editor/author of the Handbook of Image and Video Processing (New York: Academic, 2000). He is a registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial and academic institutions.

Dr. Bovik was named Distinguished Lecturer of the IEEE Signal Processing Society in 2000, received the IEEE Signal Processing Society Meritorious Service Award in 1998, the IEEE Third Millennium Medal in 2000, the University of Texas Engineering Foundation Halliburton Award in 1991 and is a two-time Honorable Mention winner of the international Pattern Recognition Society Award for Outstanding Contribution (1988 and 1993). He was named a Dean's Fellow in the College of Engineering in the Year 2001. He is a Fellow of the IEEE and has been involved in numerous professional society activities, including: Board of Governors, IEEE Signal Processing Society, 1996-1998; Editor-in-Chief, IEEE Transactions on Image Processing, 1996-2002; Editorial Board, The Proceedings of the IEEE, 1998-present; and Founding General Chairman, First IEEE International Conference on Image Processing, held in Austin, Texas, in November, 1994.

Hamid Rahim Sheikh (S'00) received his B.Sc. degree in Electrical Engineering from the University of Engineering and Technology, Lahore, Pakistan, and his M.S. degree in Engineering from the University of Texas at Austin in May 2001, where he is currently pursuing a Ph.D. degree.

His research interests include using natural scene statistical models and human visual system models for image and video quality assessment.

Eero P. Simoncelli (S'92-M'93-SM'04) received the B.A. degree in Physics in 1984 from Harvard University, Cambridge, MA, a certificate of advanced study in mathematics in 1986 from Cambridge University, Cambridge, England, and the M.S. and Ph.D. degrees in 1988 and 1993, both in Electrical Engineering from the Massachusetts Institute of Technology, Cambridge.

He was an assistant professor in the Computer and Information Science department at the University of Pennsylvania from 1993 until 1996. He moved to New York University in September of 1996, where he is currently an Associate Professor in Neural Science and Mathematics. In August 2000, he became an Associate Investigator of the Howard Hughes Medical Institute, under their new program in Computational Biology. His research interests span a wide range of topics in the representation and analysis of visual images, in both machine and biological vision systems.