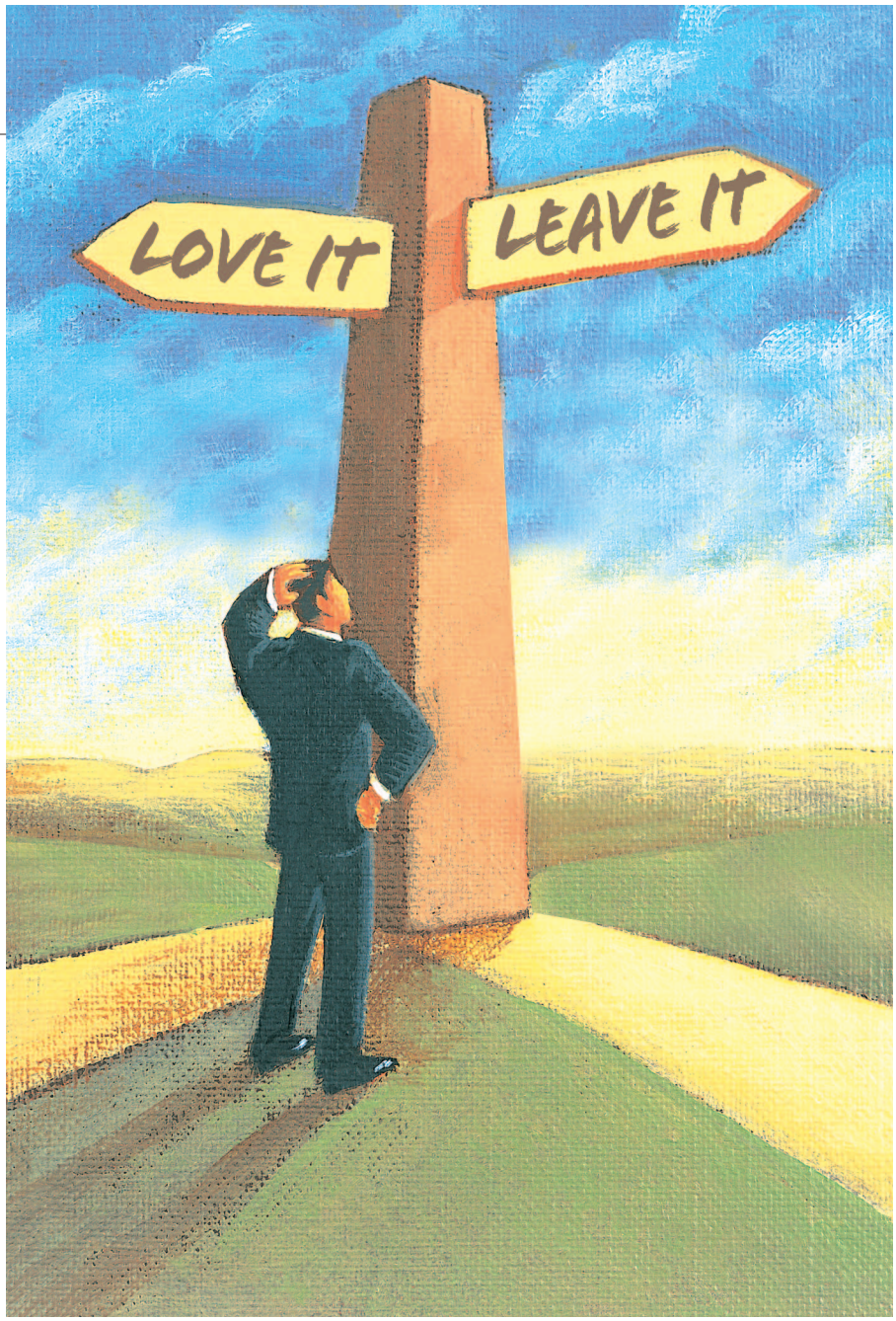
Zhou Wang and Alan C. Bovik



Mean Squared Error: Love It or Leave It?

A new look at signal fidelity measures

For more than 50 years, the mean-squared error (MSE) has been the dominant quantitative performance metric in the field of signal processing. It remains the standard criterion for the assessment of signal quality and fidelity; it is the method of choice for comparing competing signal processing methods and systems, and, perhaps most importantly, it is the nearly ubiquitous preference of design engineers seeking to optimize signal processing algorithms. This is true despite the fact that in many of these applications, the MSE exhibits weak performance and has been widely criticized for serious shortcomings, especially when dealing with perceptually important signals such as speech and images. Yet the MSE has exhibited remarkable staying power, and prevailing attitudes towards the MSE seem to range from "it's easy to use and not so bad" to "everyone else uses it."

So what is the secret of the MSE—why is it still so popular? And is this popularity misplaced? What is wrong with the MSE when it does not work well? Just how wrong is the MSE in these cases? If not the MSE, what else can be used? These are the questions we'll be concerned with in this article. Our backgrounds are primarily in the field of image processing, where the MSE has a particularly bad reputation, but where, ironically, it is used nearly as much as in other areas of signal processing. Our discussion will often deal with the role of the MSE (and alternative methods) for processing visual signals. Owing to the poor performance of the MSE as a visual metric, interesting alternatives are arising in the image processing field. Our goal is to stimulate fruitful thought and discussion regarding the role of the MSE in processing other types of signals. More specifically, we hope to inspire signal processing engineers to rethink whether the MSE is truly the criterion of choice in their own theories and applications, and whether it is time to look for alternatives.

Digital Object Identifier 10.1109/MSP.2008.930649

1053-5888/09/$25.00©2009IEEE

WHAT IS THE MSE?

We begin with a discussion of the MSE as a signal fidelity measure. The goal of a signal fidelity measure is to compare two signals by providing a quantitative score that describes the degree of similarity/fidelity or, conversely, the level of error/distortion between them. Usually, it is assumed that one of the signals is a pristine original, while the other is distorted or contaminated by errors.

Suppose that $\mathbf{x} = \{x_i | i = 1, 2, \cdots, N\}$ and $\mathbf{y} = \{y_i | i = 1, 2, \cdots, N\}$ are two finite-length, discrete signals (e.g., visual images), where N is the number of signal samples (pixels, if the signals are images) and x_i and y_i are the values of the ith samples in x and y, respectively. The MSE between the signals is

$$\mathrm{MSE}(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2. \tag{1}$$

In the MSE, we will often refer to the error signal $e_i, = x_i - y_i$, which is the difference between the original and distorted signals. If one of the signals is an original signal of acceptable (or perhaps pristine) quality, and the other is a distorted version of it whose quality is being evaluated, then the MSE may also be regarded as a measure of signal quality. Of course, a more general form is the l_p norm

$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{N} |e_i|^p \right)^{1/p}.$$

In the literature of image processing, MSE is often converted into a peak signal-to-noise ratio (PSNR) measure

$$\mathrm{PSNR} = 10 \log_{10} \frac{L^2}{\mathrm{MSE}},$$

where L is the dynamic range of allowable image pixel intensities. For example, for images that have allocations of 8 b/pixel of gray-scale, $L = 2^8 - 1 = 255$. The PSNR is useful if images having different dynamic ranges are being compared, but otherwise contains no new information relative to the MSE.

WHY DO WE LOVE THE MSE?

The MSE has many attractive features:

1) It is simple. It is parameter free and inexpensive to compute, with a complexity of only one multiply and two additions per sample. It is also memoryless—the squared error can be evaluated at each sample, independent of other samples.

2) All l_p norms are valid distance metrics in $\mathbf{R}^N$, which satisfy the following convenient conditions, and allow for consistent, direct interpretations of similarity:

— nonnegativity: $d_p(\mathbf{x}, \mathbf{y}) \geq 0$

— identity: $d_p(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$

— symmetry: $d_p(\mathbf{x}, \mathbf{y}) = d_p(\mathbf{y}, \mathbf{x})$

— triangular inequality: $d_p(\mathbf{x}, \mathbf{z}) \leq d_p(\mathbf{x}, \mathbf{y}) + d_p(\mathbf{y}, \mathbf{z})$.

In particular, the $p = 2$ case (proportional to the square root of the MSE) is the ordinary distance metric in N-dimensional Euclidean space.

3) It has a clear physical meaning—it is the natural way to define the energy of the error signal. Such an energy measure is preserved after any orthogonal (or unitary) linear transformation, such as the Fourier transform (Parseval's theorem). The energy preserving property guarantees that the energy of a signal distortion in the transform domain is the same as in the signal domain. This property distinguishes d_2 from the other l_p energy measures, which are not energy preserving.

4) The MSE is an excellent metric in the context of optimization. The MSE possesses the very satisfying properties of convexity, symmetry, and differentiability. Minimum-MSE (MMSE) optimization problems often have closed-form analytical solutions, and when they don't, iterative numerical optimization procedures are often easy to formulate, since the gradient and the Hessian matrix of the MSE are easy to compute.

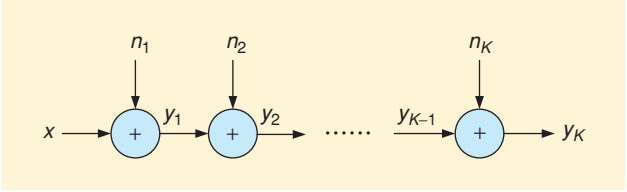
5) The MSE is also a desirable measure in the statistics and estimation framework (where the sample average in (1) is replaced by statistical expectation). The MSE in this form was first introduced by C.F. Gauss, who also noted its arbitrary nature relative to actual loss in applications, as well as its conveniences [1]. The MSE is additive for independent sources of distortions. This is illustrated in Figure 1, where a zero-mean random source x passes through a cascade of K additive independent zero-mean distortions $n_1, n_2, \ldots, n_K$, resulting in $y_1, y_2, \ldots, y_K$, i.e.,

$$y_k = x + \sum_{i=1}^{k} n_i \quad \text{for} \quad k = 1, 2, \ldots, K.$$

The overall MSE is simply the sum of the MSEs from the individual distortion stages

$$\begin{aligned} \mathrm{MSE}(x, y_K) &= E\left[(x - y_K)^2\right] \\ &= E\left[\left(\sum_{k=1}^{K} n_k\right)^2\right] \\ &= \sum_{k=1}^{K} E\left[n_k^2\right] \\ &= \mathrm{MSE}(x, y_1) + \mathrm{MSE}(y_1, y_2) + \cdots \\ &\quad + \mathrm{MSE}(y_{K-1}, y_K), \end{aligned}$$

so that the contribution from each source of distortion may be analyzed independently. When a squared error function is combined with Gaussian assumptions on the source and noise models, the optimal signal estimate is



[FIG1] Independent additive sources of distortions and the additive property of the MSE.

often analytical and linear. An excellent example is the Wiener filter for signal deconvolution and denoising (that also requires second-order stationary assumptions about the signal and the noise).

6) Finally, the MSE is widely used simply because it is a convention. Historically, it has been employed extensively for optimizing and assessing a wide variety of signal processing applications, including filter design, signal compression, restoration, denoising, reconstruction, and classification. Moreover, throughout the literature, competing algorithms have most often been compared using the MSE/PSNR. It therefore provides a convenient and extensive standard against which the MSE/PSNR results of new algorithms may be compared. This saves time and effort but further propagates the use of the MSE.

SO WHAT'S WRONG WITH THE MSE?

It is apparent that the MSE possesses many favorable properties for application and analysis, but the perspicacious reader might point out that a more fundamental issue has been missing. That is, does the MSE really measure signal fidelity? Given all of its above-mentioned attractive features, a signal processing practitioner might opt for the MSE if it proved to be a reasonable signal fidelity measure. But is that the case?

Unfortunately, the converse appears true when the MSE is used to predict human perception of image fidelity and quality [2]–[5]. An illustrative example is shown in Figure 2, where an original Einstein image is altered by different types of distortion: a contrast stretch, mean luminance shift, contamination by additive white Gaussian noise, impulsive noise distortion, JPEG compression, blur, spatial scaling, spatial shift, and rotation. In

[FIG2] Comparison of image fidelity measures for "Einstein" image altered with different types of distortions. (a) Reference image. (b) Mean contrast stretch. (c) Luminance shift. (d) Gaussian noise contamination. (e) Impulsive noise contamination. (f) JPEG compression. (g) Blurring. (h) Spatial scaling (zooming out). (i) Spatial shift (to the right). (j) Spatial shift (to the left). (k) Rotation (counter-clockwise). (l) Rotation (clockwise).

Figure 2, both MSE values and values of another quality index, the structural similarity (SSIM) index, are given. The SSIM index is described in detail later where it also refers to this figure. Note that the MSE values [relative to the original image (a)] of several of the distorted images are nearly identical [images (b)–(g)], even though the same images present dramatically (and obviously) different visual quality. Also notice that images that undergo small geometrical modifications [images (h)–(i)] may have very large MSE values relative to the original, yet show a negligible loss of perceived quality. So a natural question is: "What's the problem with the MSE?"

IMPLICIT ASSUMPTIONS WHEN USING THE MSE

We'll look at this topic from three different viewpoints. First, we'll examine the following underlying implicit assumptions that an engineer is making when she/he decides to use the MSE (or any l_p metric) to evaluate signal fidelity:

1) Signal fidelity is independent of temporal or spatial relationships between the samples of the original signal. In other words, if the original and distorted signals are randomly re-ordered in the same way, then the MSE between them will be unchanged.
2) Signal fidelity is independent of any relationship between the original signal and the error signal. For a given error signal, the MSE remains unchanged, regardless of which original signal it is added to.
3) Signal fidelity is independent of the signs of the error signal samples.
4) All signal samples are equally important to signal fidelity.

All of the above implicit assumptions are very strong, since they impose significant limitations on the signal samples, how they interact with each other, and how they interact with the error. But are they accurate? Are they useful or damaging in the context of measuring signal fidelity?

Unfortunately, not one of them holds (even roughly) in the context of measuring the visual perception of image fidelity. Dramatic, visual examples of the failures of the MSE with respect to the veracity of these assumptions are demonstrated in Figure 3.

In Figure 3(a), the bottom-left image was created by adding independent white Gaussian noise to the original image (top-left). In the top-right image, the spatial ordering of the pixels was changed (through a sorting procedure), but without changing any of the pixel values from those in the original. The bottom-right image was obtained by applying the same reordering procedure to the bottom-left image. Of course, the MSE (or any l_p metric) between the two left images, and between the two right images, are identical. Yet, the bottom-right image appears significantly noisier than the bottom-left image—the perceived visual fidelity of the bottom-right image is much poorer than that of the bottom-left image. Apparently, MSE Assumption 1 is not a good one when measuring the fidelity of images. This is an excellent example of the failure of the MSE (and all other l_p metrics) to take into account the dependencies (textures, orderings, patterns, etc.) that occur between signal samples. Since natural image signals are highly structured—the ordering of the signal samples carries important perceptual structural information about the contents of the visual scene—this is a severe shortcoming of the MSE for image fidelity measurement. By extension, other perceptual signals, such as sound and speech signals, also contain perceptually important structures, the fidelity of which might not be well measured by the MSE.

Figure 3(b) conveys a dramatic example of the failure of MSE Assumption 2. In the figure, the same error signal was added to both original images (top left) and (top right). The error signal was created to be fully correlated with the top-left image. Both distorted images have exactly the same MSE and l_p metrics (no matter what p is chosen) with respect to their originals, but the visual distortion of the bottom-right image is much stronger than that of the bottom-left image. Clearly, the correlation (and dependency) between the error signal and the underlying image signal significantly affects perceptual image distortion—an important feature that is completely ignored by any l_p metric.

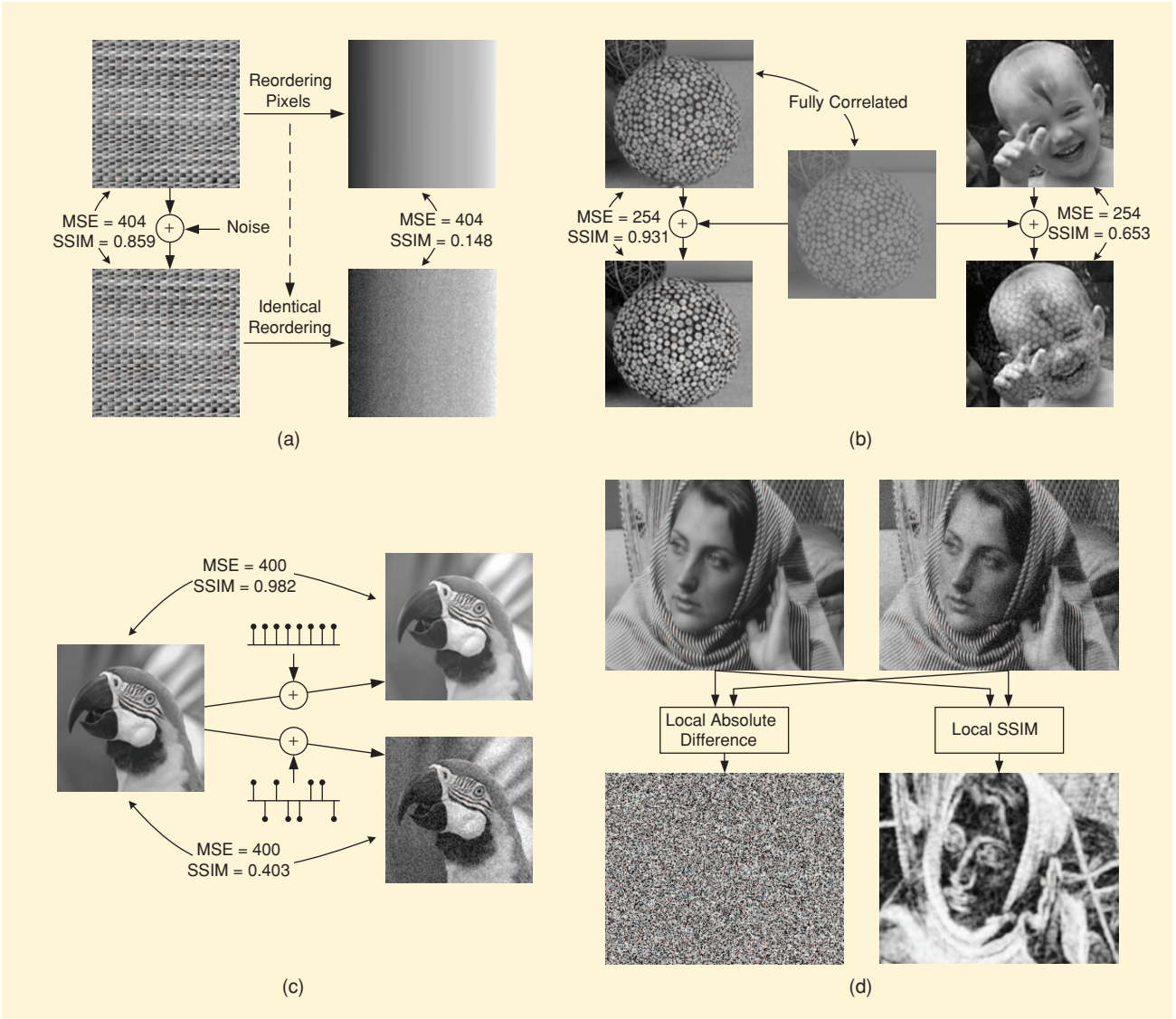
Figure 3(c) depicts the failure of the underlying MSE Assumption 3. In the figure, the first distorted image was obtained by adding a constant value to all pixels in the original image, while the second distorted image was generated by the same method, except that the signs of the constant were randomly chosen to be positive or negative. The visual fidelity of the two distorted images is drastically different. Yet, the MSE (or any l_p metric) ignores the effect of signs and reports the same fidelity measure for both distorted images.

Figure 3(d) supplies a particularly instructive example of both MSE Assumptions 1 and 4. Distorted image (top right) was created by adding independent white Gaussian noise to the original image (top left). Clearly, the degree of noise-induced visual distortion varies significantly across the spatial coordinates of the image. In particular, the noise in the facial (and other smooth-intensity) regions appears rather severe, yet is visually negligible in other regions containing patterns and textures. The perceived fidelity of the distorted image varies over space, although the error signal (bottom left) has a uniform energy distribution across space. Since all image pixels are treated equally in the formulation of the MSE (and all l_p metrics), such image content-dependent variations in image fidelity cannot be accounted for.

OBSERVING THE MSE IN SIGNAL SPACE

The second method in examining the problem of signal fidelity measurement is to look at it in N-dimensional signal space, where each coordinate represents the value of a signal sample and N is the number of samples in the signal. Thus, each signal is represented by a single point in signal space. Any distortion can be interpreted as adding a distortion vector to the signal point. The set of all distortion vectors of the same length constitutes an equal-MSE hypersphere in signal space. For example, in image space, Figure 2(b)–(g) reside near the surface of the same equal-MSE hypersphere centered about Figure 2(a). Since images on the same hypersphere can have substantially different perceptual image fidelity (as in Figure 2), the length of

a distortion vector does not suffice as a good indication of image fidelity. Apparently, the *directions* of these vectors are also important. This interpretation has been probed by allowing two signal fidelity measures to compete with each other, by maximizing/minimizing one measure while holding the other fixed [6]. This was used to define an optimal signal synthesis algorithm that seeks to maximally differentiate one measure from another. An example is shown in Figure 4, where an iterative procedure is used for image synthesis. First, an initial distorted image is generated by adding a random vector in the image space to the reference image. Starting from the initial image, the algorithm iteratively moves along the direction of increasing/decreasing a fidelity metric (in this case, the SSIM index [7] that will be detailed later), while constraining the movement to be within the equal-MSE hypersphere. The iteration continues until it converges to the best/worst SSIM images (shown in Figure 4 along with intermediate images). This example provides a strong visual demonstration of the failing of

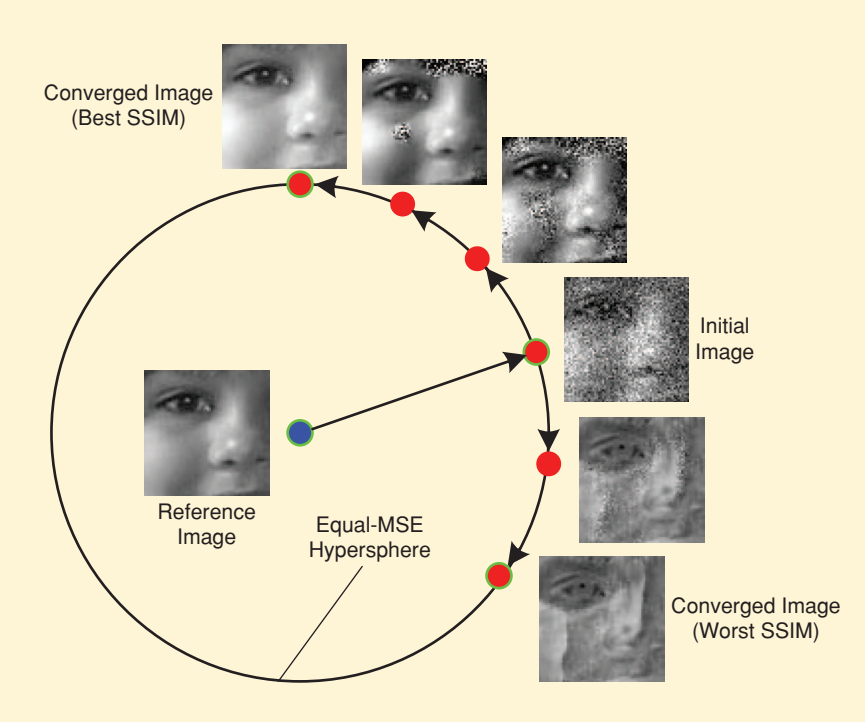


[FIG3] Failures of the MSE and other l_p metrics. (a) An original image (top left) is distorted by adding independent white Gaussian noise (bottom left). In the top-right image, the pixels are reordered by sorting pixel intensity values. The same reordering process is applied to the bottom-left image to create the bottom-right image. The MSE (and any l_p metric) between the two left images and between the two right images are the same, but the bottom-right image appears much noisier than the bottom-left image. (b) Two original images (top left and top right) are distorted by adding the same error image (middle), which is fully correlated with the top-left image. The MSE (and any l_p metric) between the two left images and between the two right images are the same, but the perceived distortion of the bottom-right image is much stronger than that of the bottom-left image. (c) An original image (left) is distorted by adding a positive constant (top right) and by adding the same constant, but with random signs (bottom right). The MSE (or any l_p metric) between the original and any of the right images are the same, but the right images exhibit drastically different visual distortions. (d) An original image (top left) is distorted by adding independent white Gaussian noise (top right). The energy distribution of the absolute difference signal (bottom left, enhanced for visibility), which is the basis in computing all l_p metrics, is uniform. However, the perceived noise level is space variant, which is reflected in the SSIM map (bottom right, enhanced for visibility).

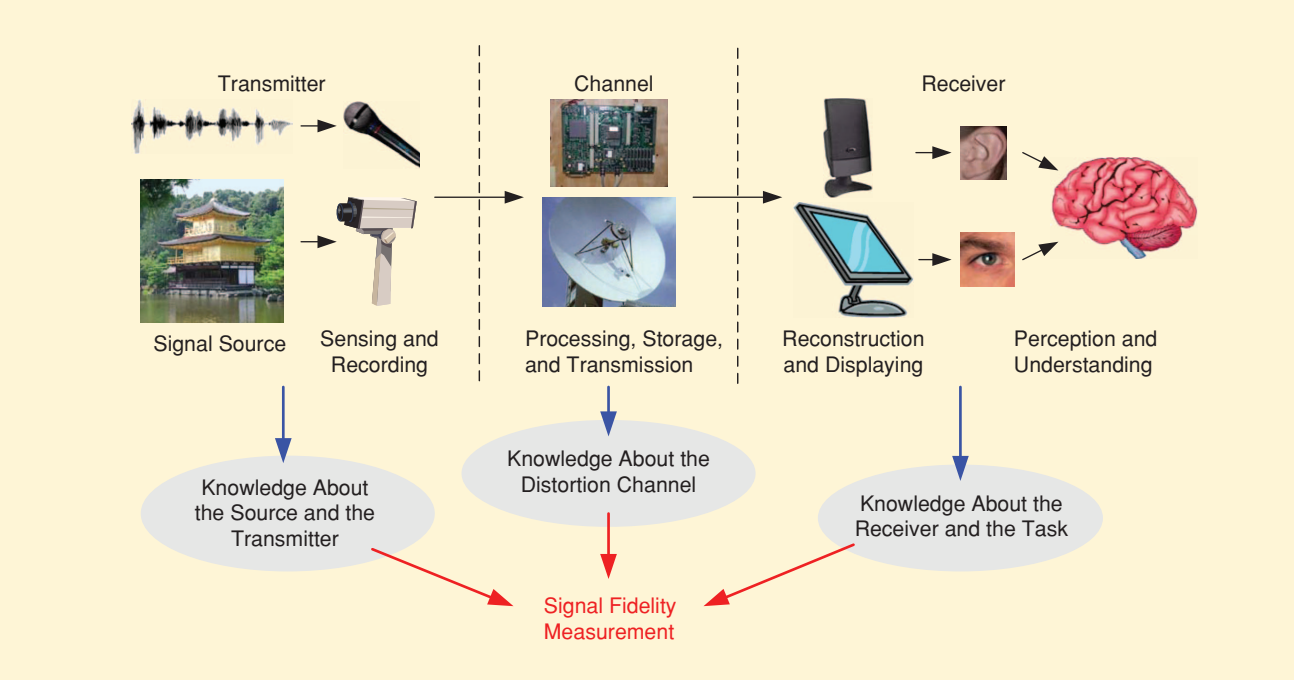
the MSE, as the various synthesized images with fixed MSE exhibit astonishingly different image fidelities. The key distinction from the earlier examples (Figures 2 and 3) is that here, the images are not hand-designed, but automatically synthesized in an optimal way.

SIGNAL FIDELITY IN AN INFORMATION COMMUNICATION FRAMEWORK

The problem of signal fidelity measurement may also be thought of within an information communication framework. An illustration is given in Figure 5. The general view is that any signal that is ultimately consumed by a receiver originated from a transmitter and is passed through a channel. Here the transmitter, the channel, and the receiver should be regarded in a broad sense. For example, if visual images are the signals under consideration, then the transmitter may include objects in the visual environment, the light source(s), atmosphere conditions, and the sensing/recording techniques and devices. The channel may include any storage and transmission processing that may alter the image signal. For example, an image communication/networking system may involve lossy compression, noise contamination, packet loss, and/or pre- or postprocessing. The receiver includes image mapping and display devices for the intended consumers (for example, human eyes and ears). Correspondingly, there are three types of knowledge that can be used for signal fidelity measurement: information source/transmitter, distortion channel, and intended receiver. Simple metrics



[FIG4] Finding the maximum/minimum SSIM images along the equal-MSE hypersphere in image space.



[FIG5] Signal fidelity measurement expressed within an information communication framework.

such as the MSE cannot account for these types of knowledge (another failing of the MSE); however, as we shall see, it is possible to include, either explicitly or implicitly, all three types of knowledge into the design of signal fidelity measures.

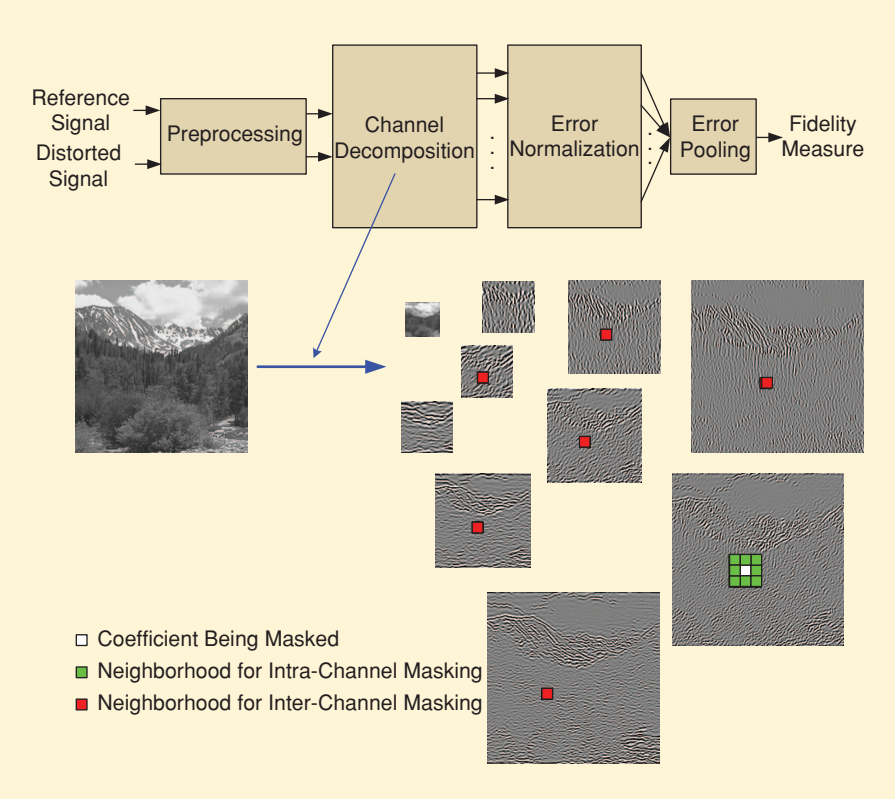
WHAT ARE THE ALTERNATIVES?

Before one decides to love or leave the MSE, another question worth asking is: "What are the alternatives?" Following the communication framework previously described, a good signal fidelity measure would need to be able to effectively and efficiently make use of knowledge about the transmitter, channel, and receiver. Depending on the application field and the type of the signals being considered, the nature of this knowledge might vary considerably: it is unlikely that there is a universal signal fidelity measure that works in all situations. Therefore, as we proceed, we use the natural image transmitter and the human visual receiver as examples as we discuss general approaches to image and video fidelity measurement.

SIMULATION OF PERCEPTUAL SYSTEMS

The most obvious type of information to incorporate into a signal fidelity system is probably receiver information, and this has been the approach taken by researchers in the fields of image quality assessment (IQA) and human speech quality assessment (SQA). In human applications, the ultimate receivers are perceptual systems such as the human visual system and the human auditory system. In the past century there has been considerable progress in increasing our depth of understanding of the functions of our perceptual systems, and in expanding these psychophysical and neurophysiological findings into mathematical models of these functions. While our overall knowledge of human perception remains in its nascent stages, current models of biological information processing mechanisms have become sufficiently sophisticated that it is of great interest to explore whether it is possible to deploy them to predict the performance of simple human behaviors, such as evaluating perceptual signal fidelity. Not surprisingly, the most common approach to perceptual signal fidelity measurement is to mathematically model each functional perceptual component, then integrate the component models, as basic building blocks, into an overall system model. The hope, of course, is that the integrated system will perform in a manner similar to the human perceptual system in assessing signal fidelity.

A typical framework for this type of approach is illustrated in Figure 6. First, the reference and the distorted signals are subject to a preprocessing stage. In the context of image fidelity measurement, this may include image registration, color space transformations, and a low-pass filtering process that simulates the point spread function of the optics of the eye. Following preprocessing, the signal is decomposed into multiple channels (or subbands) using a biologically motivated linear transform. In particular, it is well known that a large number of neurons in the primary visual cortex are tuned to visual stimuli with specific spatial locations, frequencies, and orientations, thus a wavelet transform [8], [9] is ideal for this task, since its bases are localized, band-pass, oriented filters. An example is shown in Figure 6, where a steerable pyramid wavelet decomposition [9] is applied to a natural image. In the next stage, the subband signal samples (or transform coefficients) are normalized in a perceptually meaningful way. For visual images, the most commonly used normalization elements are the contrast sensitivity function (CSF) and masking effects. The CSF accounts for the variation of visual sensitivity as a function of spatial frequency. Thus, different weights may be assigned to the subbands at different levels of decomposition. Masking effects describe the reduction of visibility of an image component due to neighboring components in space, frequency, and/or orientation. A masking effect in the wavelet domain is illustrated in Figure 6, where the masker



[FIG6] A prototypical perceptual signal fidelity measurement system and an example of image decomposition and visual masking.

components can come from the spatial neighbors in the same subband (intra-channel masking) or from the nearby subbands in frequency and orientation (inter-channel masking). In the final stage, the normalized error signal is pooled to form a single signal fidelity score. The most commonly used pooling methods adopt an l_p form, possibly with adaptive spatial weighting. Pioneering work of this general approach dates back as early as the 1970s [10], with a large number of variations being proposed since then. Representative models include [11]–[15]. Most of these methods are general purpose, in the sense that they do not assume any specific distortion type. They are intended to be flexible enough to be used in a variety of different applications. There are also many methods that are designed for specific applications. For example, many image fidelity measurement methods have been developed specifically for block-discrete cosine transfer (DCT) [16], [17] and wavelet-based image compression [18], [19]. Using information about the compression process can be viewed as incorporating channel knowledge into the design process. For tutorial reviews and more detailed descriptions about such methods, refer to [2]–[4].

There is little doubt that if all of the functional components of a human perceptual system were precisely simulated, then an accurate prediction of perceived signal fidelity could be achieved. However, this is quite difficult to accomplish for a number of reasons. First, our knowledge of the functional architecture of biological perceptual systems is still quite limited. Secondly, human perceptual systems are quite complicated and contain many nonlinearities, while most existing computational models are linear or quasi-linear and have been developed based on restricted, simplistic stimuli and threshold psychophysics (where the visual sensitivities are defined and measured at the contrast detection threshold levels). An interesting recent algorithm called the visual signal-to-noise ratio (VSNR) [20] first determines whether the distortions are below the threshold of visual detection, and then quantifies the distortions that are beyond the threshold in a separate stage. This has led to significant improvement over traditional signal fidelity measures [20], and the algorithm appears to be generally competitive with other IQA algorithms.

STRUCTURAL SIMILARITY

Using the framework of image fidelity measurement as an image communication problem, the perceptual simulation methods just discussed primarily seek to model the receiver. Few approaches have attempted to account for the characteristics of the transmitter. As with any communication system problem, the more that is known about the transmitter, then, the better job of communication that can be accomplished, particularly if the signal source is highly specific. Thus, it is relevant that the cluster of natural images occupies an extremely tiny portion in the image space [21]. This potentially provides strong prior information about what an original image should look like, which should be a precious source of information for the design of signal fidelity measures. In the literature of computational neuroscience, it has been long hypothesized that the human visual system is highly adapted to the natural visual environment [21], not only in terms of short-term adaptation ability (e.g., to lighting conditions), but also through long-term neural evolution and development. As a result, we may view the modeling of the natural image source (the transmitter) and of the human visual systems (the receiver) as dual problems [21]. Of course, such hypothesis may also be adopted with respect to other biological perceptual systems.

One recently proposed approach to image fidelity measurement, which may also prove highly effective for measuring the fidelity of other signals, is the SSIM index [7]. SSIM actually takes a variety of forms, depending on whether it is implemented at a single scale [7], [22], over multiple scales [23], or in the wavelet domain [24]. Regardless of specifics, the SSIM approach was originally motivated by the observation that natural image signals are highly structured, meaning that the samples of natural image signals have strong neighbor dependencies, and these dependencies carry important information about the structures of the objects in the visual scene. Of course, the same may be said of most other signals.

The principle philosophy underlying the original SSIM approach is that the human visual system is highly adapted to extract structural information from visual scenes. Therefore, at least for image fidelity measurement, the retention of signal structure should be an important ingredient. Equivalently, an algorithm may seek to measure structural distortion to achieve image fidelity measurement. Figure 7 helps illustrate the distinction between structural and nonstructural distortions. In the figure, the nonstructural distortions (a change of luminance or brightness, a change of contrast, Gamma distortion, and a spatial shift) are caused by ambient environmental or instrumental conditions occurring during image acquisition and display. These distortions do not change the structures of images of the objects in the visual scene. However, other distortions (additive noise and blur and lossy compression) significantly distort the structures of images of the objects. If we view the human visual system as an ideal information extractor that seeks to identify and recognize objects in the visual scene, then it must be highly sensitive to the structural distortions and automatically compensates for the nonstructural distortions. Consequently, an effective objective signal fidelity measure should simulate this functionality.

The main ideas of SSIM were introduced in [22], and more formally distilled in [7] and [3]. The basic form of SSIM is very easy to understand. Suppose that $\mathbf{x}$ and $\mathbf{y}$ are local image patches taken from the same location of two images that are being compared. The local SSIM index measures the similarities of three elements of the image patches: the similarity $l(\mathbf{x}, \mathbf{y})$ of the local patch luminances (brightness values), the similarity $c(\mathbf{x}, \mathbf{y})$ of the local patch contrasts, and the similarity $s(\mathbf{x}, \mathbf{y})$ of the local patch structures. These local similarities are expressed using simple, easily computed statistics, and combined together to form local SSIM [7]

$$S(\mathrm{x},\mathrm{y}) = l(\mathrm{x},\mathrm{y}) \cdot c(\mathrm{x},\mathrm{y}) \cdot s(\mathrm{x},\mathrm{y}) = \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right) \cdot \left(\frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right) \cdot \left(\frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}\right), \quad (2)$$

where μ_x and μ_y are (respectively) the local sample means of x and y, σ_x and σ_y are (respectively) the local sample standard deviations of x and y, and σ_{xy} is the sample cross correlation of x and y after removing their means. The items C_1, C_2, and C_3 are small positive constants that stabilize each term, so that near-zero sample means, variances, or correlations do not lead to numerical instability. Actually, even if $C_1 = C_2 = C_3 = 0$, SSIM usually works quite well. This choice of constants defined the first and simplest version of SSIM with the universal image quality index (UQI) [22].

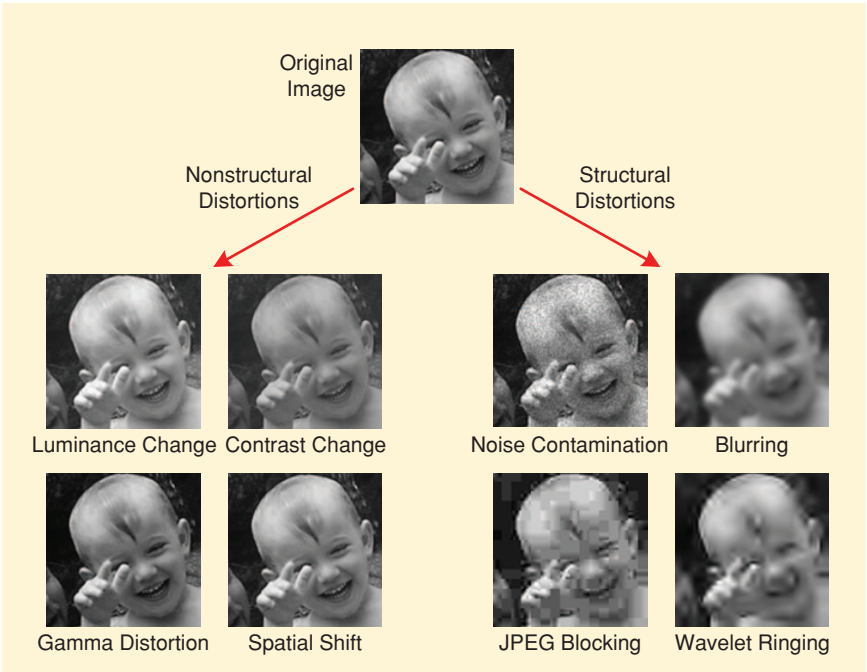
The SSIM index is symmetric: $S(\mathrm{x},\mathrm{y}) = S(\mathrm{y},\mathrm{x})$, so that two images being compared give the same index value regardless of their ordering. It is also bounded: $-1 < S(\mathrm{x},\mathrm{y}) \leq 1$, achieving maximum value $S(\mathrm{x},\mathrm{y}) = 1$ if and only if $\mathrm{x} = \mathrm{y}$. The SSIM index is computed locally within a sliding window that moves pixel-by-pixel across the image, resulting in a SSIM map. The SSIM score of the entire image is then computed by pooling the SSIM map, e.g., by simply averaging the SSIM values across the image [7]. More sophisticated adaptive space-variant weighting can also be used [25]. A recent extension expressed SSIM using an adaptive basis decomposition framework that allows for explicit separation of structural and nonstructural distortions [26]. The best performance using the SSIM framework (to date) is obtained when the SSIM score is computed over a range of scales [23], [27]. This makes sense for a couple of reasons: first, images, like other natural signals, contain structures that occur over a range of spatial scales, and second, the human visual system decomposes visual data into multiscale spatial channels early in the visual pathway [3].

Despite its simplicity, the SSIM index performs remarkably well across a wide variety of image and distortion types as has been shown in intensive human studies [27]. By example, Figure 2 shows the SSIM scores of images having near identical MSE values. Without much effort, it can be seen that the SSIM scores are much more consistent than the MSE scores relative to visual perception. Luminance-shifting and contrast-stretching, which generally does not degrade image structure, lead to very high SSIM values, while noise contamination and excessive JPEG-compression lead to low SSIM values. Similarly, in Figures 3(a)–(c), in which the MSE was demonstrated to be highly problematic, the SSIM index provides relative scores that are much more consonant with perception. Figure 4 describes an algorithm that seeks to maximally differentiate the SSIM index from the MSE, where extremely high- and low-SSIM (and correspondingly high- and low-quality) images were found on the same equal-MSE hypersphere, providing a dramatic example of the failings of the MSE.

The effectiveness of the structure-centric SSIM method is better revealed by comparing the SSIM maps with the absolute error maps in Figure 3(d) and Figure 8. In both types of maps, brightness corresponds to goodness-of-fidelity for each measure (i.e., brighter = better). As shown in Figure 3(d), the SSIM index handles the texture masking visual effect quite well, although the added noise is applied uniformly across the image, the visual appearance of the image is most highly degraded where the image is smooth, or relatively texture-less. While any l_p measurement would be likewise uniform (as shown), the SSIM scores accord with visual perception. Likewise, Figure 8 depicts JPEG-induced annoying pseudo-contouring effects (in the sky region) and blocking artifacts (along the boundaries of the building) that are successfully captured by the SSIM index, yet poorly predicted by the absolute error map.

SSIM has been used for evaluating image processing results in a rapidly increasing number of exciting applications. Some are listed as follows:

- image fusion [28]
- image compression [25]
- image watermarking [29]
- chromatic image quality [30]
- retinal and wearable displays [31]
- video hashing [32]

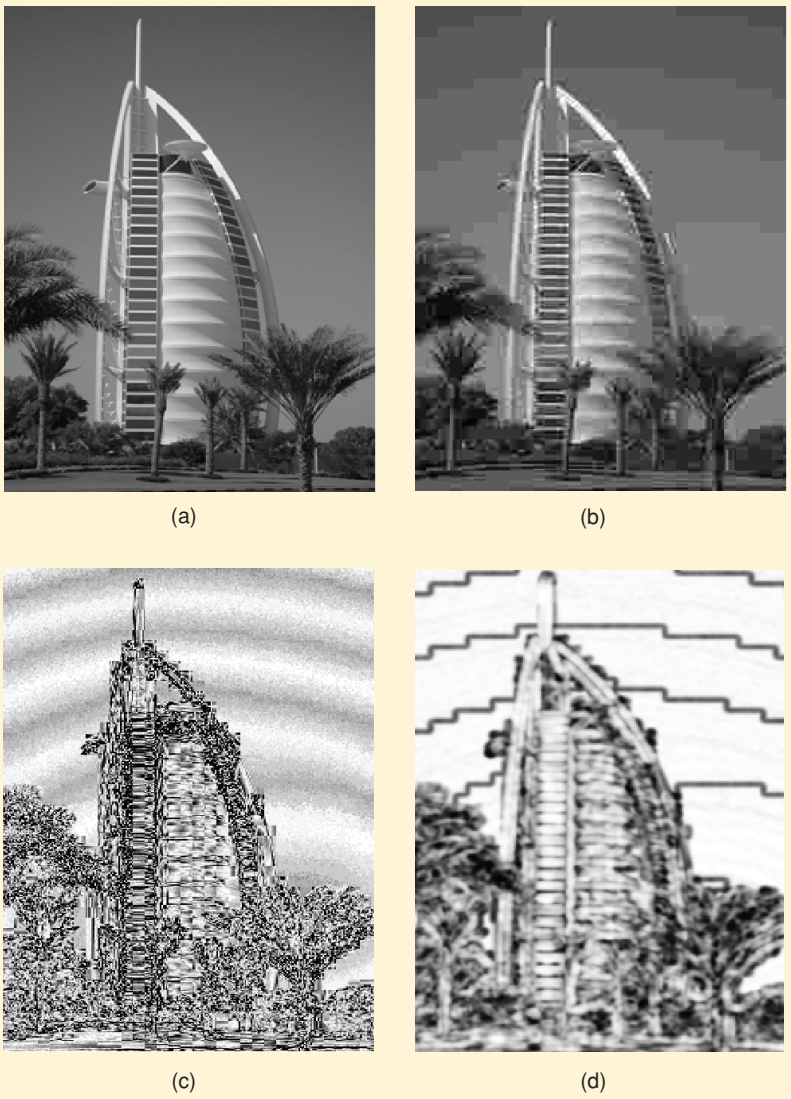


[FIG7] Examples of structural versus nonstructural distortions.

- wireless video streaming [33]
- visual surveillance [34]
- radar imaging [35]
- digital camera design [36]
- infrared imaging [37]
- MRI imaging [38]
- chromosome imaging [39]
- remote sensing [40]
- target recognition [41].

An exciting consideration is the possibility of numerous extended applications beyond image processing, since the SSIM index does not rely on specific image or visual models. The generic definition of SSIM suggests that it should find broad applicability.

A drawback of the basic SSIM index as previously described is its sensitivity to relative translations, scalings and rotations of images, as seen in Figure 2(h)–(l). This is undesirable and contradictory to the philosophy of structural similarity, since small geometric distortions are nonstructural. To handle such situations, a wavelet-domain version of SSIM, called the complex wavelet SSIM (CW-SSIM) index was developed [24]. The CW-SSIM index is also inspired by the fact that local phase contains more structural information than magnitude in natural images [42], while rigid translations of image structures leads to consistent phase shifts. In the complex wavelet transform domain, let $\mathbf{c}_x = \{c_{x,i} | i = 1, 2, \cdots N\}$ and $\mathbf{c}_y = \{c_{y,i} | i = 1, 2, \cdots N\}$, respectively, be two sets of coefficients extracted at the same spatial location in the same wavelet subbands of two images x, y being compared. Then, the CW-SSIM index [24] has a form similar to the spatial domain counterpart



[FIG8] **Comparison of image fidelity/distortion maps. (a) Reference image. (b) JPEG compressed image. (c) Absolute error map of the distorted image (enhanced for visibility). (d) SSIM index map of the distorted images (enhanced for visibility).**

$$\tilde{S}(\mathbf{c}_x, \mathbf{c}_y) = \tilde{m}(\mathbf{c}_x, \mathbf{c}_y) \cdot \tilde{p}(c_x, c_y) = \frac{2\sum_{i=1}^{N} |c_{x,i}||c_{y,i}| + K}{\sum_{i=1}^{N} |c_{x,i}|^2 + \sum_{i=1}^{N} |c_{y,i}|^2 + K} \cdot \frac{2\left|\sum_{i=1}^{N} c_{x,i} c_{y,i}^*\right| + K}{2\sum_{i=1}^{N} |c_{x,i} c_{y,i}^*| + K},$$

where c^* is the complex conjugate of c and K is a small positive constant (stabilizer). The first component $\tilde{m}(\mathbf{c}_x, \mathbf{c}_y)$ is completely determined by the magnitudes of the coefficients, with maximum value one achieved if and only $|c_{x,i}| = |c_{y,i}|$ for all i. Thus, this term is equivalent to the SSIM index applied to the magnitudes of the coefficients (note that the coefficients are zero-mean, owing to the bandpass nature of the wavelet filters). The second component $\tilde{p}(c_x, c_y)$ is determined by the consistency of phase changes between $\mathbf{c}_x$ and $\mathbf{c}_y$. It achieves maximum value one when the phase difference between $c_{x,i}$ and $c_{y,i}$ is constant for all i. This phase component of CW-SSIM effectively captures image structural similarity since local image structure is maintained by the relative phase patterns of local image frequencies (as sampled by the wavelet coefficients); moreover, a constant phase shift of all wavelet coefficients will not change the structure of local image feature. CW-SSIM is locally computed from each subband, then averaged over space and subbands, yielding an

overall CW-SSIM Index between the original and the distorted images. The CW-SSIM method is simultaneously robust with respect to luminance changes, contrast changes and translations [24], leading to robustness with respect to small scalings and rotations, since they can be locally approximated with translations. Referring again to Figure 2, it may be seen that CW-SSIM delivers high scores to luminance-shifted, contrast-stretched, space-shifted, scaled, and rotated images, and low scores to the images containing structural distortions.

WE HOPE TO INSPIRE SIGNAL PROCESSING ENGINEERS TO RETHINK WHETHER THE MSE IS TRULY THE CRITERION OF CHOICE IN THEIR OWN THEORIES AND APPLICATIONS, AND WHETHER IT IS TIME TO LOOK FOR ALTERNATIVES.

VISUAL INFORMATION FIDELITY

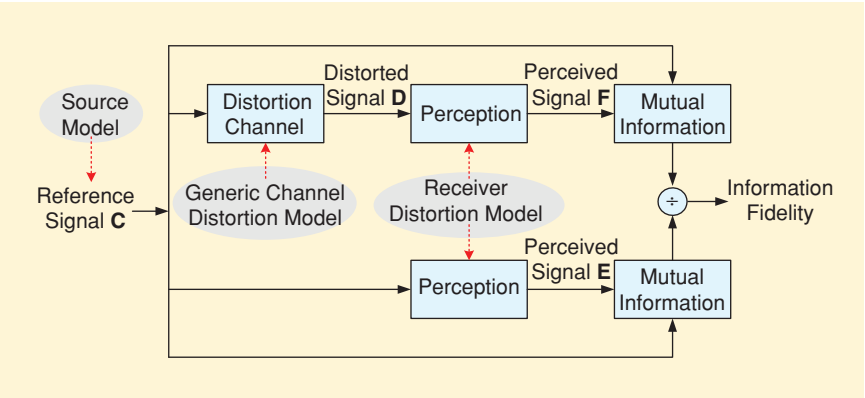
Visual information fidelity (VIF) methods explicitly incorporate statistical models of all the components in the communication system interpretation of signal fidelity measurement, viz., the transmitter, the channel, and the receiver [43], [44]. This approach attempts to relate signal fidelity to the amount of information that is shared between two signals. The shared information is quantified using the concept of mutual information, a widely used measure in information theory. Since mutual information is purely a statistical measure that might only be loosely related to human perception of information, it places fundamental limits on the amount of perceptually relevant information that could be extracted from a signal, provided that the hypothesized statistical models about the signal source, the channel distortion, and the receiver distortion are accurate.

The idea is better explained in Figure 9, which follows the framework of the VIF index designed for comparing visual images [44]. The reference image is modeled by a wavelet-domain Gaussian scale mixture (GSM) [45], which has been shown to quite effectively model the non-Gaussian marginal distributions of the wavelet coefficients of natural images, while also capturing the dependencies between the magnitudes of neighboring wavelet coefficients. Let $\mathbf{c}$ be a collection of M neighboring wavelet coefficients extracted from a local patch in a wavelet subband. Then the wavelet-domain GSM model is simple: model $\mathbf{c}$ as $\mathbf{c} = \sqrt{z}\,\mathbf{u}$, where $\mathbf{u}$ is a zero-mean Gaussian vector, and $\sqrt{z}$ is an independent scalar random variable. In other words, the vector $\mathbf{c}$ is a mixture of random Gaussian vectors that share the same covariance structure $\mathbf{C}_u$, but scale differently according to the magnitude of $\sqrt{z}$. The GSM model thus provides a simple, yet powerful VIF source/transmitter model for the signal fidelity measurement framework.

In VIF, a generic and simple image distortion model is used to model all distortions that may occur between the reference and the distorted image signals, including artificial distortions such as compression artifacts. The VIF distortion model assumes that the image distortion can be roughly described locally as a combination of a uniform wavelet-domain energy attenuation with a subsequent independent additive noise: $\mathbf{d} = g\mathbf{c} + \mathbf{v}$, where $\mathbf{c}$ and $\mathbf{d}$ are random vectors extracted from the same location in the same wavelet subband in the reference and the distorted images, respectively. Here g represents a scalar deterministic gain factor (viz., blur), while $\mathbf{v}$ is independent additive zero-mean white Gaussian noise with covariance $\mathbf{C}_v = \sigma_v^2\mathbf{I}$. Although such a model may be criticized for being very general, or for not directly accounting any specific distortion types (such as JPEG blocking), it provides a reasonable first approximation. Moreover, it makes the algorithm both mathematically tractable and computationally accessible. And most important of all, it lends exceedingly good performance to the VIF index over a wide range of distortion types [27], [44].

Finally, in the VIF receiver model, the visual distortion process is modeled as stationary, zero-mean, additive white Gaussian noise process in the wavelet transform domain, mainly to account for internal neural noise: $\mathbf{e} = \mathbf{c} + \mathbf{n}$ and $\mathbf{f} = \mathbf{d} + \mathbf{n}$. Here, $\mathbf{e}$ and $\mathbf{f}$ denote random coefficient vectors in the same wavelet subbands in the perceived reference and distorted images, respectively, and $\mathbf{n}$ is independent white Gaussian noise with covariance matrix $\mathbf{C}_n = \sigma_n^2\mathbf{I}$. This receiver model greatly simplifies prior image fidelity measurement algorithms that relied upon sophisticated computational models of the eye, the retina, and visual cortex.

Given the statistical models of source/transmitter, channel distortion, and receiver, the mutual information between $\mathbf{c}$ and $\mathbf{e}$, and between $\mathbf{c}$ and $\mathbf{f}$, are given by [44]



[FIG9] System diagram of the generic VIF method.

$$I(\mathbf{c};\mathbf{e}|z) = \frac{1}{2}\log\frac{|z\mathbf{C}_u+\sigma_n^2\mathbf{I}|}{|\sigma_n^2\mathbf{I}|} = \frac{1}{2}\sum_{j=1}^{M}\log\left(1+\frac{z\lambda_j}{\sigma_n^2}\right);$$

$$I(\mathbf{c};\mathbf{f}|z) = \frac{1}{2}\log\frac{|g^2z\mathbf{C}_u+(\sigma_v^2+\sigma_n^2)\mathbf{I}|}{|(\sigma_v^2+\sigma_n^2)\mathbf{I}|} = \frac{1}{2}\sum_{j=1}^{M}\log\left(1+\frac{g^2z\lambda_j}{\sigma_v^2+\sigma_n^2}\right).$$

In these expressions, we have factored the covariance matrix $\mathbf{C}_u = \mathbf{Q}\Lambda\mathbf{Q}^T$, where Λ is a diagonal matrix whose diagonal entries are the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_M$. The mutual information is computed at each spatial location in each subband of the image, using local maximum-likelihood estimates of z, g and σ_v [44]. Assuming independence between the distortion parameters across local coefficient patches and independence between the coefficients across different wavelet subbands, the overall mutual information can be computed by simple summation. Finally, the VIF index is defined as the ratio of the summed mutual information

$$\text{VIF} = \frac{I(\mathbf{C};\mathbf{F}|\mathbf{z})}{I(\mathbf{C};\mathbf{E}|\mathbf{z})} = \frac{\sum_{i=1}^{N} I(\mathbf{c}_i;\mathbf{f}_i|\mathbf{z}_i)}{\sum_{i=1}^{N} I(\mathbf{c}_i;\mathbf{e}_i|\mathbf{z}_i)},$$

where i is the index of local coefficient patches, with all subbands included.

The VIF measure has been extensively tested across a wide variety distortion types. A detailed report can be found in [27], where ten well-known or state-of-the-art image quality/fidelity measures were tested using the laboratory for image and video engineering (LIVE) image quality assessment database that is maintained by the Laboratory of Image and Video Engineering at The University of Texas at Austin (the complete database, along with extensive subjective ratings of the images, are available at http://live.ece.utexas.edu). According to this study, which is, to date, the most extensive that has been done, the VIF index exhibits superior performance relative to all other image fidelity measurement algorithms.

FROM IMAGES TO VIDEOS

Naturally, the question arises whether new developments in image fidelity measurement can be extended to moving images or video. Indeed, there is strong demand for video fidelity measures in the multimedia communication industry that has motivated the formation of the industry-controlled video quality experts group (VQEG), which seeks to develop, validate, and standardize objective measures for video quality.

A simple and obvious way to implement a video fidelity measure is to apply a still image measure on a frame-by-frame basis, then average over all frames. Indeed, SSIM has been deployed in this manner with rather good results [46], and is now deployed as a basic video quality assessment tool in popular public-domain software such as the Moscow State University video quality measurement tool (www.compression.ru/video/quality_measure/) and the award-winning freeware H.264 codec *x*.264 (www.videolan.org/developers/x264.html).

However, two important aspects of video are missing in frame-by-frame implementations of image fidelity metrics: 1) there are strong correlations between adjacent video frames that define temporal and spatio-temporal signal structures; and 2) video contains perceptually important structured motion. The most common method to take into account temporal correlations is temporal or spatiotemporal filtering [4], [47], [48]. The general approach is similar to the framework presented in Figure 6. Linear filters or filter banks are first applied along the spatial and temporal directions, and the filtered signals are normalized to reflect additional perceptual effects such as the temporal CSF (human visual sensitivity as a function of temporal frequency) and temporal masking. A video version of VIF has also been proposed that includes statistical modeling of the temporal filter coefficients, followed by an information fidelity measure [49].

AN OPTIMIZED SYSTEM IS ONLY AS GOOD AS THE OPTIMIZATION CRITERION USED TO DESIGN IT.

Naturally, motion is one of the most important contributors to the information content of videos. Yet, only relatively few existing video quality assessment (VQA) algorithms detect motion explicitly and use motion information directly [46], [50]–[53]. Direct use of motion is desirable since temporal filtering cannot fully capture motion. Indeed, motion does not create all temporal intensity variations in video. Moreover, the responses of temporal filters cannot supply the speed of motion, since intensity attributes affect the filter responses.

More recent methods seek to detect motion and convert the motion information into spatiotemporal weighting factors in the pooling stage of video fidelity measurement [46], [50], [51]. For example, in [51], visual perception of motion is modeled using a recent psychophysical study of human visual speed perception [54] that has provided critical information about the prior and likelihood probability distributions of visual speed perception. The general approach of motion-based weighting has proven consistently effective in improving the performance of video fidelity measures over standard (i.e., simple averaging) MSE/PSNR and SSIM measures when tested using the VQEG Phase I database (available at www.vqeg.org) [46], [50], [51].

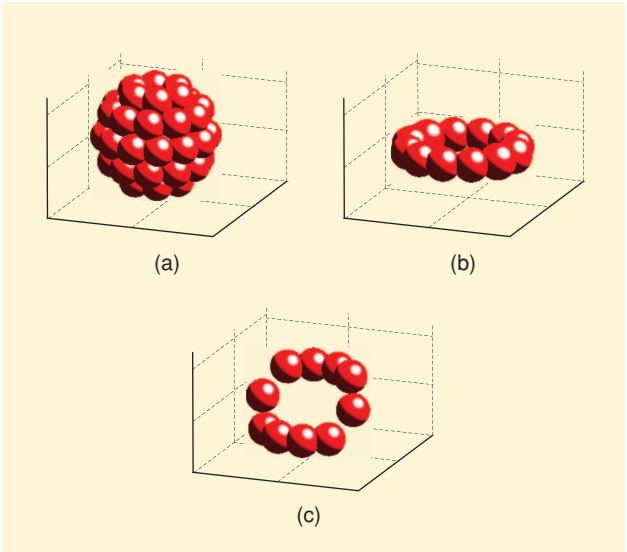
A substantially different approach involves using optical flow estimation [52], [53] to adaptively guide spatiotemporal filtering using three-dimensional (3-D) Gabor filterbanks. The key difference of this method is that a subset of filters are selected adaptively at each location based on the direction and speed of motion, such that the major axis of the filter set is

oriented along the direction of motion in the frequency domain, as illustrated in Figure 10. The video fidelity evaluation process is carried out with coefficients computed from these selected filters only. Distortions in the video that are purely spatial, meaning intra-frame distortions, result in changes in the frequency components along the plane, and are captured in the Gabor filter outputs. Distortions in the video that are purely temporal, meaning inter-frame distortions, result in a changes in the axis along which the plane intersects the Gabor filters. This approach has been incorporated into the VIF [52] and the CW-SSIM [53] measures, with improved performance relative to frame-based VIF and SSIM.

FROM VISUAL TO AUDIO SIGNALS

Algorithms for computing objective measures of perceptual signal quality assessment have also been a focus in the audio signal processing field, where standards such as the International Telecommunications Union (ITU) standard BS.1387–Perceptual Evaluation of Audio Quality (PEAQ) [55], and the ITU standard P.862–Perceptual Evaluation of Speech Quality (PESQ) [56] have been developed after many years of careful collaboration between industry, government, and academic groups. Although there are differences, PEAQ and PESQ both share similarities with many of the earlier image quality assessment algorithms, in that channel decompositions are used to analyze the data across frequency bands, followed by a form of perceptual differencing (along with many other facets, such as aligning the signals and accounting for loudness).

It is unlikely that PEAQ or PESQ would be adaptable to other applications, since these perceptual indices are built on detailed audio perception models. However, it is possible that these well-established standards might benefit by some of the notions recently developed for image quality assessment, e.g., that the direct measurement of structural distortions might be effective for audio signals, since certainly such signals contain one-dimensional structures not dissimilar to those found in images, e.g., sudden changes in sound volume (similar to image intensity edges), periodic or quasi-periodic sounds (similar to many image textures), and so on. Likewise, information-theoretic approaches to audio signal quality assessment may be fruitful.



[FIG10] Frequency domain illustration of motion-guided 3-D filters for video fidelity assessment. (a) Full set of Gabor filter bank. (b) Filters selected when there is no motion. (c) Filters selected when there is motion (orientation depends on the direction and the speed of motion).

Towards this end, researchers at New Mexico State University have recently applied variations of SSIM for the assessment of audio signal quality [57]. Using a set of seven different (original) 44.1-KhZ audio signals, they generated a variety of distorted signals by adding noise, using different types of audio compression algorithms, and band-limiting the signals. Subjective tests using 15 subjects were conducted. The linear correlation of SSIM scores with the subjective scores when applied to the distorted data was found to be surprisingly high. Further variations of SSIM using temporal and time-frequency weightings resulted in even better correlations with the human data.

While there remains much work to be done towards determining the efficacy of SSIM (or VIF) for audio quality assessment, these results are quite promising, and suggest again, that differencing methods, whether perceptually weighted or otherwise, might be replaced by different distance indices.

THE NEXT STEP: OPTIMIZATION USING NEW SIGNAL FIDELITY MEASURES

The nontraditional methods for signal fidelity measurement that we have been discussing are well suited for many practical applications. In particular, they can be used to monitor signal fidelity as part of quality of service (QoS) efforts, and also to benchmark signal processing systems and algorithms in comparative studies. However, the potential application scope of these new measurement devices greatly exceeds QoS or benchmarking algorithms. In particular, they can be used as design criteria to optimize signal processing algorithms and systems.

WHAT IS OPTIMAL?

As mentioned earlier, the MSE is used not only to evaluate, but also to optimize a large variety of signal processing algorithms and systems. Of course, a good rule of thumb is that an optimized system is only as good as the optimization criterion used to design it. Therefore, it makes sense to carefully study the criteria employed in the design of current signal processing systems. In particular, a worthwhile direction is to deploy modern signal fidelity measures that are specifically suited to each application, then do (or redo) the optimization work. This is not a new idea, of course; in statistical communication theory, a basic concept is to inject as much information as possible regarding transmitter, channel and receiver, in order to deliver the best performance possible [58]. Thus, a priori statistical modeling can be used to refine or establish weights in the optimization process. The next logical step is to toss out the optimization criteria altogether and develop new, specific ones. There has already been interesting work

done in this vein in a number of fields, including speech coding [59], image halftoning [60], image segmentation and classification [61], image watermarking [62], and image/video compression [2], [13], [14], [63], [64]. It is no accident that these examples are perception-related, as models of visual and audio perception have gained rapid ground over the past few decades. As examples, simple perceptual models are employed in the development of the JPEG quantization table [65] and the JPEG 2000 visual optimization tools [66].

In many mature fields such as image compression, state-of-the-art algorithms have achieved a performance plateau levels and significant improvement has become difficult to attain. The key here, is relative to which performance criterion? Nominally, minimizing the MSE (or PSNR) is the optimization goal. However, the enormous difference in perceptual relevance between the MSE and modern signal fidelity measures suggests that there is good reason to be optimistic about further improvement in established algorithms. Thus far, progress in this direction has been only preliminary, yet very promising. In particular, the SSIM Index and its relatives have been found effective in growing list of image processing optimization problems, owing both to its computational simplicity and its relative analytic tractability (e.g., it is differentiable [6]). Figure 4 is a simple example of a SSIM-based iterative optimization algorithm that uses a gradient computation at each iteration.

IMAGE COMPRESSION

Image coding algorithms traditionally optimize the MSE under the constraint of a limited bit budget. The possibility of using perceptual cost functions is suggested by Figure 8, where the SSIM index does a better job of predicting local image quality. Perceptual image coding algorithms typically deploy perceptual models in a preprocessing stage. Perceptual normalization transforms the image into a perceptually uniform space, where all transform coefficients have equal perceptual importance. Standard coding is then applied uniformly to all coefficients.

Signal compression is a well-suited area in which signal fidelity metrics may supply enhanced results. For example, in [67], the SSIM index was applied in an iterative algorithm, where at each iteration, a bit allocation scheme is used to spatially redistribute the available bits based on the SSIM map obtained from the last iteration. The scheme seeks to improve the worst case scenario, so that the lowest quality region in the image is enhanced. In other words, a maximin (maximum of minima) SSIM criterion is the optimization goal. The philosophy of the approach is that visual attention is often attracted to image regions having annoying artifacts, thus degrading the perceptual quality of the entire image. In [67], the SSIM-based bit allocation scheme was embedded into the well-known set partitioning in hierarchical trees (SPIHT) image compression algorithm [68]. Figure 11 shows comparisons of the maximin SSIM-optimized coding result with the SPIHT algorithm,

[FIG11] Comparison of coding results of SPIHT and maximin SSIM-optimized algorithms at 0.2 b/pixel.

showing that the fidelity of the maximin SSIM-optimized compression is more uniformly distributed over the image space. Those regions having the worst quality in the SPIHT coded image showed the most improvement, while detailed structures (e.g., in the enlarged regions) were preserved.

IMAGE QUANTIZATION

One of the most challenging aspects of achieving perceptually optimized image or video compression is the selection of quantization parameters in the compression domain. For example, quantization of the block DCT coefficients in JPEG, MPEG, or subband compression algorithms has in the past been approached using measured responses of visual frequency response (contrast sensitivity). While this approach has merit, and does lead to improved performance [13], it is of considerable interest to attempt optimization of the quantization parameters using an objective criterion that correlates well with overall visual perception of quality, rather than using data measured from one facet of human vision. There remains much work to be done in this direction; however, bounds on the SSIM values of quantized image DCT coefficients have been developed which can be used to place bounds on the performance of uniformly quantized image DCT coefficients, and by extension for rate allocation in image and video coding [69].

SIGNAL COMPRESSION IS A WELL-SUITED AREA IN WHICH SIGNAL FIDELITY METRICS MAY SUPPLY ENHANCED RESULTS.

IMAGE RESTORATION

Another general class of problems that could clearly benefit from the use of signal fidelity metrics is signal restoration, meaning, removal of blur and/or noise from distorted signals. Given a linearly blurred version of an original signal $\mathbf{x}$ to which noise was somehow added

$$\mathbf{y} = \mathbf{g} * \mathbf{x} + \mathbf{n},$$

where $\mathbf{g}$ is a linear blur and $\mathbf{n}$ is additive noise, the goal of this classical problem is to attempt recovery of $\mathbf{x}$ given $\mathbf{g}$ and the observation $\mathbf{y}$, using a linear filter. The classical solution to this problem is the Wiener filter $\mathbf{h}_{\mathrm{MSE}}$ which minimizes the expectation of the squared error between the true signal $\mathbf{x}$ and the estimate $\hat{\mathbf{x}}$

$$E[(\hat{\mathbf{x}} - \mathbf{x})^2]$$

over all linear filtered solutions $\hat{\mathbf{x}} = \mathbf{h} * \mathbf{y}$. Within the context of our discussion, it is natural to consider replacing the MSE with a perceptual distortion measure. In fact, owing to its analytic simplicity, this can be done with relative ease with the SSIM index. In [70], a closed-form linear estimator was derived that maximizes a statistical version of the SSIM index, called stat-SSIM, as an alternative to the MMSE solution. The stat-SSIM index of two random vectors $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ is

$$\mathrm{Stat-SSIM}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \left(\frac{2\mu_{\mathbf{x}}\mu_{\mathbf{y}} + C_1}{\mu_{\mathbf{x}}^2 + \mu_{\mathbf{y}}^2 + C_1}\right) \times \left(\frac{2E\left[(\tilde{\mathbf{x}} - \mu_{\mathbf{x}})(\tilde{\mathbf{y}} - \mu_{\mathbf{y}})\right] + C_2}{E\left[(\tilde{\mathbf{x}} - \mu_{\mathbf{x}})^2\right] + E\left[(\tilde{\mathbf{y}} - \mu_{\mathbf{y}})^2\right] + C_2}\right),$$

where $\mu_{\mathbf{x}} = E[\tilde{\mathbf{x}}]$, $\mu_{\mathbf{y}} = E[\tilde{\mathbf{y}}]$. Finding the SSIM-optimal linear filter that minimizes the stat-SSIM index is a nonconvex optimization problem, but one that can be transformed into a quasi-convex problem that has a tractable solution that is of similar computational complexity as the MMSE filter [70]. Figure 12 shows four images: original, blurred with noise added, MSE-optimal linear filter, and SSIM-optimal linear filtered. The images were blurred using a Gaussian blur kernel with a standard deviation of five pixels, while the simulated additive noise is white Gaussian with a standard deviation of 50 [the gray-scale range is (0, 255)]. By visually comparing Figure 12(c) and (d), it can be seen that the SSIM-optimal solution is visually crisper, with better contrast, and better retention of visual details. The sample MSE (1) and sample SSIM value (2) computed from the restored images are likewise in accord with expectations: the MSE-optimal solution yields a lower MSE, while the SSIM-optimal yields a higher SSIM value. Naturally, the SSIM-optimal solution has a better appearance, since the SSIM value correlates more closely with visual perception of quality.

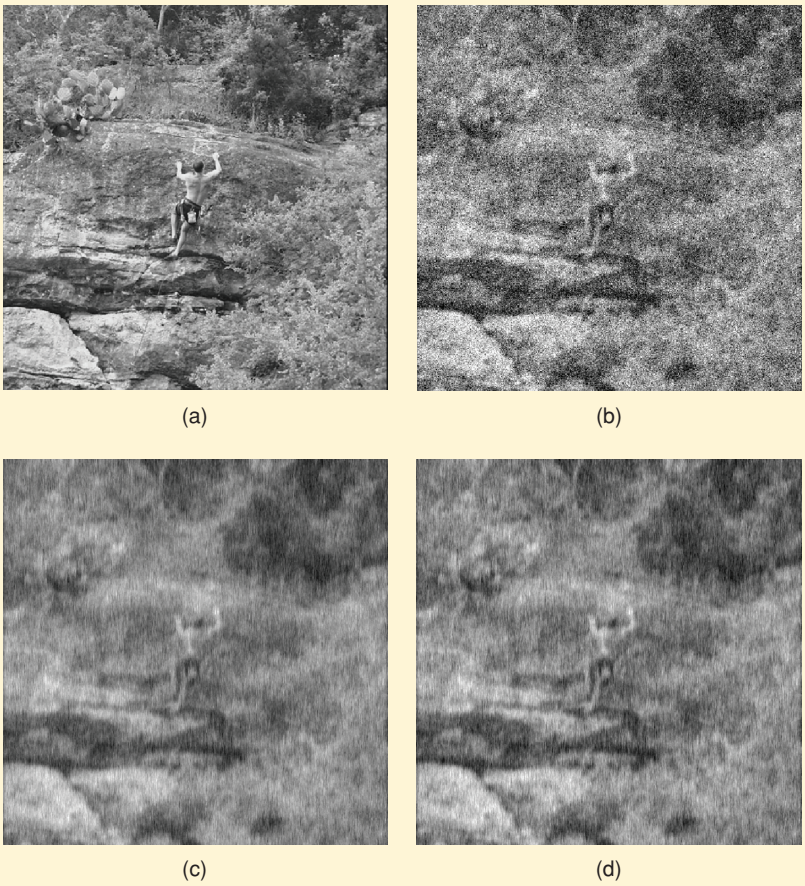
It must be remembered, of course, that the image restorations in these examples are accomplished by space-invariant linear filters only. Although the difference in performance depicted in Figure 12 is clearly visible (and representative [70]), better results should be obtainable using SSIM-based optimization of nonlinear filters, iterative restoration, and multiscale or wavelet-based restoration. The possibility of accomplishing image restoration, denoising, reconstruction, and other basic processes by optimizing criteria such as SSIM, VIF, or other perceptual criteria reveals important research questions, the answers to which may markedly elevate the performance of state-of-the-art image improvement algorithms.

PATTERN RECOGNITION

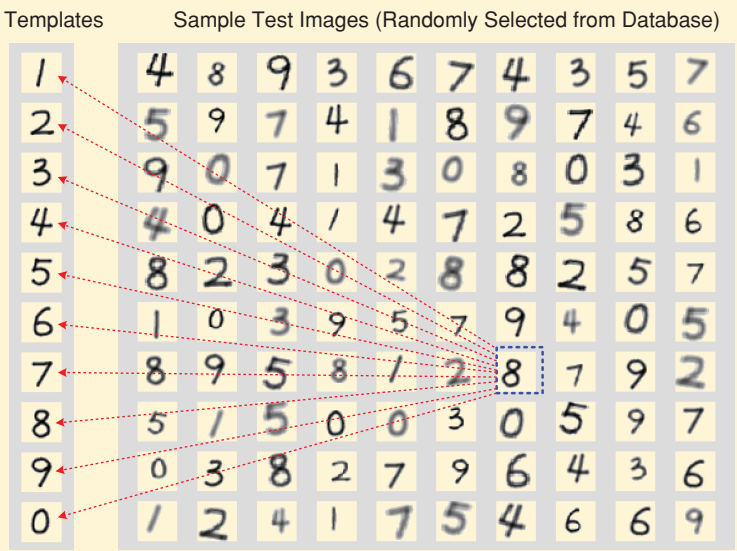
Image fidelity measurement, as we have previously defined it, involves making comparisons between original images and distorted versions. Since the algorithms perform image comparisons that accord with visual perception, it is natural to wonder if there are other applications of image comparison which these tools might be adapted for. One such very broad application is recognition of patterns or objects in images. The pattern recognition capability of biological perceptual systems is quite remarkable when cast against the best efforts made thus far by the engineering and computer science communities. Certainly our

perceptual systems are highly tuned for this purpose, although the mechanisms for visual recognition remain largely undiscovered. There is evidence, however, that the type of bandpass channels that we have been deploying for signal fidelity measurement, in for example, CW-SSIM [24], are also used in pattern recognition processes by humans [71]. This leads us to believe that signal fidelity measures that accord with human performance have the potential to achieve success in pattern recognition tasks. As we will discover, the CW-SSIM index has several nice properties that promote it for recognition tasks. In the following, we will show how it has been successfully used for several applications.

Printed character recognition is a classic problem in the pattern recognition field—one that is simple when the characters are machine printed, and much more difficult when they are printed by hand. As an example application, the CW-SSIM index was used in a digit recognition test which can be described as image matching without registration [24]. First, ten standard digit templates were created manually, as shown in Figure 13. A total of 2,430 distorted images (243 for each digit) were then generated by shifting, scaling, rotating, and blurring the standard



[FIG12] Linear image restoration. (a) Original image. (b) Blurred, noisy image (MSE = 3,131, SSIM = 0.27). (c) Image restored with linear MMSE filter (MSE = 1,064, SSIM = 0.35). (d) Restored using linear minimum stat-SSIM filter (MSE = 1,116, SSIM = 0.38).



Recognition Error Rate (%)

Digit	MSE	CW-SSIM
1	16.0	0
2	34.6	1.6
3	50.6	2.9
4	36.2	0
5	52.3	3.7
6	43.6	2.1
7	31.7	5.8
8	50.2	0.4
9	40.7	0
0	48.6	7.0
All	40.4	2.3

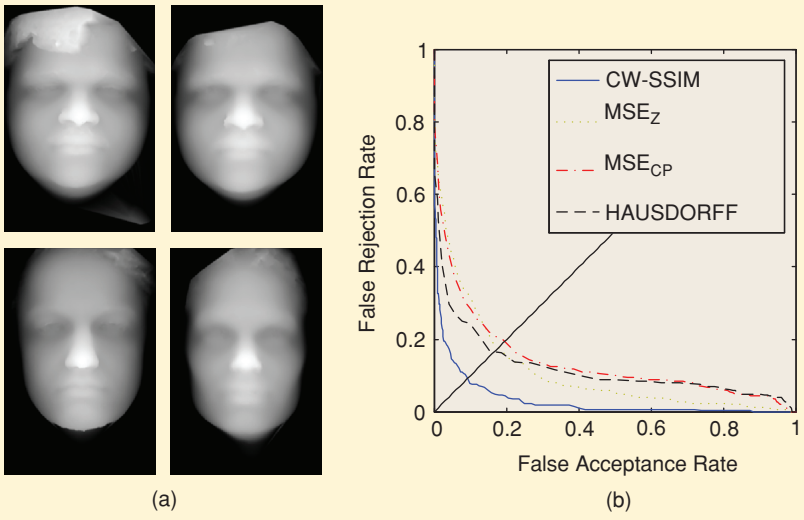
[FIG13] Image matching without registration-application to digit recognition.

templates (examples shown in Figure 13). Each distorted image is then recognized by direct image matching with the ten standard templates, without any prior alignment or registration process. The MSE and the CW-SSIM index were used as the matching standards. As expected, the MSE is sensitive to translation, scaling, and rotation of images, leading to poor recognition error rates. By contrast, the performance of the CW-SSIM index was surprisingly good, with an overall recognition error rate of only 2.3%. Of course, the CW-SSIM index was not developed for digit recognition, and there exist other sophisticated digit recognition systems that may perform better. However, the simple CW-SSIM approach just described performs rather impressively, given that it requires no training, nor any of the complex preprocessing steps often encountered in such systems, such as registration, edge detection, feature extraction, and contour processing, etc., nor any probabilistic modeling of the patterns or the distortions.

THERE IS LITTLE DOUBT THAT IF ALL OF THE FUNCTIONAL COMPONENTS OF A HUMAN PERCEPTUAL SYSTEM WERE PRECISELY SIMULATED, THEN AN ACCURATE PREDICTION OF PERCEIVED SIGNAL FIDELITY COULD BE ACHIEVED.

Automatic face recognition is a topic of considerable recent interest in biometrics and security applications. Here again, there are many sophisticated face recognition systems that utilize multiple stages of processing of two-dimensional (2-D) or 3-D images of human faces. In another experiment regarding its applicability to recognition problems, the CW-SSIM index was applied to the problem of recognizing human face images from 3-D range maps obtained using a binocular scanner [72]. 3-D face recognition is a compelling new application with great promise relative to 2-D systems. In the experiment, a test data set consisting of 360 face images was partitioned into a gallery set containing one image of each of the 12 subjects with a neutral expression, and a probe set of 348 images (29 for each subject) with a neutral or an arbitrary expression. Figure 14 shows several sample images from the database. Three popular face recognition measures were evaluated on the face image database along with the CW-SSIM index, and standard receiver operating characteristic (ROC) curves (false rejection rate versus false acceptance rate) generated for each algorithm. The algorithms deployed were the depth MSE (MSE_Z) [73], the close-point MSE (MSE_{CP}) [74], the Hausdorff distance [75], and the CW-SSIM index. The ROC curves for false rejection rate against false acceptance rate are shown in Figure 14, showing CW-SSIM to significantly outperform the other algorithms. The success of the CW-SSIM index in this application can likely also be attributed its robustness against small misregistrations. Moreover, since it does not involve any search procedure, it is less costly in terms of computation as compared to the MSE_{CP} measure and the Hausdorff distance [72]. The complexity of CW-SSIM is bounded by the complexity of calculating the wavelet coefficients, which for images of size $M \times N$ is $O(MN\log MN)$, whereas the search procedure in MSE_{CP} or using the Hausdorf distance have a natural complexity of $O(M^2N^2)$.

Medical pattern recognition is a broad area of increasing interest. An important problem in many biomedical imaging applications is to evaluate intra- and inter-observer agreement of experts (usually physicians) in identifying and localizing structures of interest in medical images, such as radiographs. An algorithm that has gained wide usage in medical imaging studies is the Dice similarity coefficient (DSC) [76]. As a third example of the applicability of perceptual metrics such as CW-SSIM for a broad array of recognitions problems, the CW-SSIM was compared with DSC using both simulated and clinical data sets [77]. The simulated images were artificially generated from binary images of line structures, followed by rotations of 0.1– 2° and translations from 0–4 pixels. In the test with the clinical data set, two radiologists marked structures of interest and outlined the lesions on a set of digitized mammograms images. One of the radiologists repeated the process. As reported in [77], the CW-SSIM measure proved to be much more robust than DSC on the simulated data set. For the clinical data, the



[FIG14] Face recognition using range images. (a) Sample range images of human face. The top two faces are from the same person. (b) ROC curves of the recognition results.

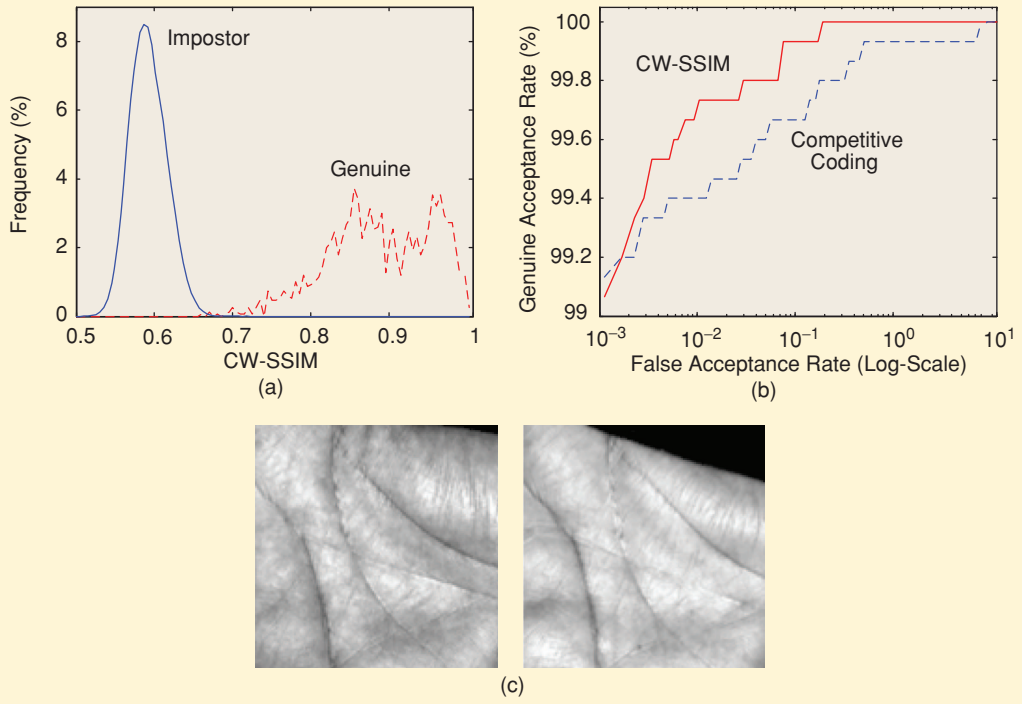
DSC measure failed to capture the obvious agreement between the two radiologists, while the CW-SSIM index indicated the agreement. Interestingly, the intra-observer agreement was consistently rated as higher than inter-observer agreement by CW-SSIM, which agrees with visual inspection of the images as well as the intuitive expectation that a human observer should agree more with himself in recognizing structures of interest than with another individual.

THE POSSIBILITY OF ACCOMPLISHING IMAGE RESTORATION, DENOISING, RECONSTRUCTION, AND OTHER BASIC PROCESSES BY OPTIMIZING CRITERIA SUCH AS SSIM, VIF, OR OTHER PERCEPTUAL CRITERIA REVEALS IMPORTANT RESEARCH QUESTIONS.

Palmprint verification is another promising biometric for which the CW-SSIM index has also been employed [78]. A test database of palmprints was created by the Biometrics Research Center at The Hong Kong Polytechnic University (www4.comp.polyu.edu.hk/~biometrics/) containing 600 palmprint images from 100 different palms (six each). Each palmprint image was matched with all the other images, resulting in a total of 179,700 pairs of comparisons. Figure 15(a) shows the genuine and the impostor distributions using the CW-SSIM index, revealing a clear separation between the two. The method was compared with the competitive coding scheme [79], one of the most successful palmprint verification algorithms. Figure 15(b) plots the ROC curves for genuine acceptance rate against false acceptance rate, showing that the CW-SSIM index again performed better in most cases. Finally, Figure 15(c) depicts two palmprint images from the same person that were correctly identified by the CW-SSIM method, but not by the competitive coding scheme.

FINAL COMMENTS

In this article, we have reviewed the reasons why we (collectively) want to love or leave the venerable (but perhaps hoary) MSE. We have also reviewed emerging alternative signal fidelity measures and discussed their potential application to a wide variety of problems. The message we are trying to send here is not that one should abandon use of the MSE nor to blindly switch to any other particular signal fidelity measure. Rather, we hope to make the point that there are powerful, easy-to-use, and easy-to-understand alternatives that might be deployed depending on the application environment and needs. While we expect (and indeed, hope) that the MSE will continue to be widely used as a signal fidelity measure, it is our greater desire to see more advanced signal fidelity measures being used, especially in applications where perceptual criteria might be relevant. Ideally, the performance



[FIG15] Palmprint verification. (a) Genuine and impostor distributions of CW-SSIM. (b) ROC curves for CW-SSIM and competitive coding. (c) Two palmprints from the same person that are correctly identified by CW-SSIM but not by the competitive coding algorithm.

of a new signal processing algorithm might be compared to other algorithms using several fidelity criteria. Lastly, we hope that we have given further motivation to the community to consider recent advanced signal fidelity measures as design criteria for optimizing signal processing algorithms and systems. It is in this direction that we believe that the greatest benefit eventually lies.

ACKNOWLEDGMENTS

The authors would like to thank Kalpana Seshadrinathan, Shalini Gupta, Dr. Mehul Sampat, Dr. Sumohana Channappayya, Dr. Lei Zhang, Zhenhua Guo, and Dr. David Zhang for their assistance in contributing advice and examples, and the anonymous reviewers for their careful reading and helpful comments.

AUTHORS

Zhou Wang (Z.Wang@ece.uwaterloo.ca) received the Ph.D. degree from the University of Texas at Austin. He is currently an assistant professor at the University of Waterloo, Canada. Previously, he was an assistant professor at the University of Texas at Arlington, a research associate at Howard Hughes Medical Institute and New York University, and a research engineer at AutoQuant Imaging, Inc. His research interests include image processing, coding, communication, and quality assessment; computational vision and pattern analysis; multimedia coding and communications, and biomedical signal processing. He has more than 60 publications and one U.S. patent in these fields and is an author of *Modern Image Quality Assessment* (Morgan & Claypool, 2006). He is an associate editor of *IEEE Signal Processing Letters and Pattern Recognition,* and a guest editor of *IEEE Journal of Selected Topics in Signal Processing.* He is a Member of the IEEE.

Alan C. Bovik (bovik@ece.utexas.edu) is the Curry/Cullen Trust Endowed Chair Professor at the University of Texas at Austin and director of the Laboratory for Image and Video Engineering. His research interests are digital video, image processing, and computational aspects of biological visual perception. He has published over 500 technical articles in these areas and holds two U.S. patents. He is the editor/author of *The Handbook of Image and Video Processing* (Academic Press, 2nd edition, 2005), *Modern Image Quality Assessment* (Morgan & Claypool, 2006), *The Essential Guide to Image Processing,* and *The Essential Guide to Video Processing*. He received the following IEEE Signal Processing Society awards: the 2008 Education Award, the 2005 Technical Achievement Award, the 2000 Distinguished Lecturer Award, and the 1998 Meritorious Service Award. He has participated in numerous professional society activities, including: Board of Governors, IEEE Signal Processing Society (1996 to 1998); editor-in-chief, *IEEE Transactions on Image Processing* (1996 to 2002); editorial board, *The Proceedings of the IEEE* (1998 to 2004); series editor for *Image, Video, and Multimedia Processing*, Morgan & Claypool (2003 to the present); and founding general chair of the First IEEE International Conference on Image Processing (1994). He also received the 2000 IEEE Third Millennium Medal as well as two journal paper awards from the Pattern Recognition Society in 1988 and 1993. He is a Fellow of the IEEE.

REFERENCES

[1] G. Casella and E.L. Lehmann, *Theory of Point Estimation*. New York: Springer-Verlag, 1999.

[2] T.N. Pappas, R.J. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Processing*, 2nd ed., May 2005.

[3] Z. Wang and A.C. Bovik, *Modern Image Quality Assessment*. San Rafael, CA: Morgan & Claypool, 2006.

[4] H.R. Wu and K.R. Rao, *Digital Image Video Quality and Perceptual Coding*. Boca Raton, FL: CRC, Nov. 2005.

[5] B. Girod, "What's wrong with mean-squared error?" in *Visual Factors of Electronic Image Communications*. Cambridge, MA: MIT Press, 1993.

[6] Z. Wang and E.P. Simoncelli, "Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities," *J. Vis.*, vol. 8, no. 12, pp. 1–13, Sept. 2008.

[7] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004 [Online]. Available: www.cns.nyu.edu/~lcv/ssim/

[8] S. Mallat, *A Wavelet Tour of Signal Processing*. New York: Academic, 1998.

[9] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, and D.J. Heeger, "Shiftable multiscale transforms," *IEEE Trans. Inform. Theory*, vol. 38, no. 9, pp. 587–607, 1992.

[10] J.L. Mannos and D.J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Inform. Theory*, vol. 20, no. 4, pp. 525–536, 1974.

[11] S. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A.B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 179–206.

[12] J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital Images & Human Vision*, A.B. Watson, Ed. Cambridge, MA: MIT Press,1993, pp. 163–178.

[13] R.J. Safranek and J.D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May 1989, pp. 1945–1948.

[14] A.B. Watson, G.Y. Yang, J.A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Processing*, vol. 6, no. 8, pp. 1164–1175, Aug. 1997.

[15] P.C. Teo and D.J. Heeger, "Perceptual image distortion," in *Proc. IEEE Int. Conf. Image Processing*, 1994, pp. 982–986.

[16] S.A. Karunasekera and N.G. Kingsbury, "A distortion measure for blocking artifacts in images based on human visual sensitivity," *IEEE Trans. Image Processing*, vol. 4, no. 6, pp. 713–724, June 1995.

[17] Z. Yu, H.R. Wu, S. Winkler, and T. Chen, "Vision-model-based impairment metric to evaluate blocking artifacts in digital video," *Proc. IEEE*, vol. 90, no. 1, pp. 154–169, Jan. 2002.

[18] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: Application to JPEG2000," *Signal Process. Image Commun.*, vol. 19, no. 2, pp. 163–172, Feb. 2004.

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

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

[21] E.P. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Annu. Rev. Neurosci.*, vol. 24, pp. 1193–1216, May 2001.

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

[23] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. Signals, Systems, Comput.*, Asilomar, CA, Nov. 2003, vol. 2, pp. 1398–1402.

[24] Z. Wang and E.P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Mar. 2005, pp. 573–576.

[25] Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 2006, pp. 2945–2948.

[26] Z. Wang and E.P. Simoncelli, "An adaptive linear system framework for image distortion analysis," in *Proc. IEEE Int. Conf. Image Processing*, Genoa, Italy, vol. 3, pp. 1160–1163, Sept. 2005.

[27] H.R. Sheikh, M.F. Sabir, and A.C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Processing*, vol. 15, no. 11, Nov. 2006, pp. 3449–3451.

[28] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proc. IEEE Int. Conf. Image Processing*, Barcelona, Spain, vol. 3, Sept. 2003, pp. 173–176.

[29] A.M. Alattar, E.T. Lin, and M.U. Celik, "Digital watermarking of low bit-rate advanced simple profile MPEG-4 compressed video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 8, pp. 787–800, Aug. 2003.

[30] A. Toet and M.P. Lucassen, "A new universal colour image fidelity metric," *Displays*, vol. 24, no. 4–5, pp. 197–207, Dec. 2003.

[31] M. von Waldkirch, P. Lukowicz, and G. Troster, "Effect of light coherence on depth of focus in head-mounted retinal projection displays," *Opt. Eng.*, vol. 43, no. 7, pp. 1552–1560, 2004.

[32] C.-Y. Hsu and C.-S. Lu, "Geometric distortion-resilient image hashing system and its application scalability," *IEEE Workshop on Multimedia and Security*, Magdeburg, Germany, 2004, pp. 81–92.

[33] V. Vukadinovi and G. Karlsson, "Trade-offs in bit-rate allocation for wireless video streaming," in *Proc. ACM Int. Symp. Modeling, Analysis, and Simulation of Wireless and Mobile Systems*, Quebec, Canada, 2005, pp. 349–353.

[34] L. Snidaro and G.L. Foresti, "A multi-camera approach to sensor evaluation in video surveillance," in *Proc. IEEE Int. Conf. Image Processing*, Genoa, Italy, 11–14 Sept. 2005, vol. 1, pp. 1101–1104.

[35] L. Bentabet, S. Jodouin, D. Ziou, and J. Vaillancourt, "Road vectors update using SAR imagery: A snake-based method," *IEEE Trans. Geosci. Remote Sensing*, vol. 41, pp. 1785–1803, Aug. 2003.

[36] W. Yu, "Practical anti-vignetting methods for digital cameras," *IEEE Trans. Consumer Electron.*, vol. 50, no. 4, pp. 975–983, Nov. 2004.

[37] J.E. Pezoa, S.N. Torres, J.P. Córdova, and R.A. Reeves, "An enhancement to the constant range method for nonuniformity correction of infrared image sequences," *Lecture Notes in Comput. Sci.*, vol. 3287, pp. 525–532, Jan. 2004.

[38] S.A. Reinsberg, S.J. Doran, E.M. Charles-Edwards, and M.O. Leach, "A complete distortion correction for MR images: II. Rectification of static-field inhomogeneities by similarity-based profile mapping," *Phys. Med. Biol.*, vol. 50, no. 11, pp. 2651–2661, June 2005.

[39] H. Choi, K.R. Castleman, and A.C. Bovik, "Color compensation of multi-color FISH images," *IEEE Trans. Med. Imaging*, to be published.

[40] E. Christophe, D. Leger, and C. Mailhes, "Quality criteria benchmark for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens*ing, vol. 43, no. 9, pp. 2103–2114, Sept. 2005.

[41] H. Chang and J. Zhang, "New metrics for clutter affecting human target acquisition," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 42, no. 1, pp. 361–368, Jan. 2006.

[42] T.S. Huang, J.W. Burdett, and A.G. Deczky, "The importance of phase in image processing filters," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 23, no. 6, pp. 529-542, Dec. 1975.

[43] H.R. Sheikh, A.C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

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

[45] J. Portilla, V. Strela, M. Wainwright, and E.P. Simoncelli, "Image denoising using scale mixtures of Gaussians in wavelet domain," *IEEE Trans. Image Processing*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.

[46] Z. Wang, L. Lu, and A.C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process. Image Commun.*, vol. 19, no. 2, pp. 121–132, Feb. 2004.

[47] C.J. van den Branden Lambrecht, D.M. Costantini, G.L. Sicuranza, and M. Kunt, "Quality assessment of motion rendition in video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 766–782, Aug. 1999.

[48] S. Winkler, *Digital Video Quality: Vision Models and Metrics*. Hoboken, NJ: Wiley, Mar. 2005.

[49] H.R. Sheikh and A.C. Bovik, "A visual information fidelity approach to video quality assessment," in *Proc. 1st Int. Workshop Video Processing & Quality Metrics for Consumer Electronics*, Scottsdale, AZ, 23–25 Jan. 2005.

[50] Z.K. Lu, W. Lin, X.K. Yang, E.P. Ong, and S.S. Yao, "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation," *IEEE Trans. Image Processing*, vol. 14, no. 11, pp. 1928–1942, 2005.

[51] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Amer. A*, vol. 24, no. 12, pp. B61–B69, Dec. 2007.

[52] K. Seshadrinathan and A.C. Bovik, "An information-theoretic video quality metric based on motion models," in *Proc. Int. Workshop Video Processing & Quality Metrics for Consumer Electronics*, Scottsdale, AZ, Jan. 2007.

[53] K. Seshadrinathan and A.C. Bovik, "A structural similarity metric for video based on motion models," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processessing*, Honolulu, HI, vol. 1, Apr. 2007, pp. 869–872.

[54] A.A. Stocker and E.P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception," *Nature Neurosci.*, vol. 9, pp. 578–585, Mar. 2006.

[55] "Methods for objective measurement of perceived audio quality," Rec. ITU-R BS.1387, 2001.

[56] "Perceptual evaluation of speech quality," Rec. ITU-T P.862, 2001.

[57] S. Kandadai, J. Hardin, and C.D. Creusere, "Audio quality assessment using the mean structural similarity measure," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Las Vegas, NV, 30 Mar.–4 Apr. 2008, pp. 221–224.

[58] D.R. Middleton, *Introduction to Statistical Communication Theory*. New York: McGraw-Hill, 1960.

[59] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Select. Areas Commun.*, vol. 10, no. 5, pp. 819–829, 1992.

[60] T.N. Pappas, J.P. Allebach, and D.L. Neuhoff, "Model-based digital halftoning," *IEEE Signal Processing Mag.* vol. 20, no. 4, pp. 14–27, July 2003.

[61] T.N. Pappas, J. Chen, and D. Depalov, "Perceptually-based techniques for image segmentation and semantic classification," *IEEE Commun. Mag.*, vol. 45, pp. 44–51, Jan. 2007.

[62] R.B. Wolfgang, C.I. Podilchuk, and E.J. Delp, "Perceptual watermarks for digital images and video," *Proc. IEEE*, vol. 87, no. 7, pp. 1108–1126, July 1999.

[63] I. Hontsch and L.J. Karam, "Locally adaptive perceptual image coding," *IEEE Trans. Image Processing*, vol. 9, no. 9, pp. 1472–1483, Sept. 2000.

[64] J.G. Ramos and S.S. Hemami, "Suprathreshold wavelet coefficient quantization in complex stimuli: psychophysical evaluation and analysis," *J. Opt. Soc. Amer. A*, vol. 18, no. 10, pp. 2385–2397, 2001.

[65] W.B. Pennebaker and J.L. Mitchell, *JPEG: Still Image Data Compression Standard*. New York: Springer-Verlag, Oct. 2006.

[66] W. Zeng, S. Daly, and S. Lei, "An overview of the visual optimization tools in JPEG 2000," *Signal Process. Image Commun.*, vol. 17, no. 1, pp. 85–104, Jan. 2002.

[67] Z. Wang, Q. Li, and X. Shang, "Perceptual image coding based on a maximum of minimal structural similarity criterion," in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, Sept. 2007, pp. 121–124.

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

[69] S.S. Channappayya, A.C. Bovik, and R.W. Heath, "Rate bounds on SSIM index of quantized images," *IEEE Trans. Image Process.*, vol. 17, no. 9, pp. 1624–1639, Sept. 2008.

[70] S.S. Channappayya, A.C. Bovik, C. Caramanis, and R.W. Heath, "Design of linear equalizers optimized for the structural similarity index," *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 857–872, June 2008.

[71] J.A. Solomon and D.G. Pelli, "The visual filter mediating letter identification," *Nature*, vol. 369, pp. 395–397, June 1994.

[72] S. Gupta, M.P. Sampat, Z. Wang, M.K. Markey, and A.C. Bovik, "Facial range image matching using the complex wavelet structural similarity metric," in *Proc. IEEE Workshop on Applications of Computer Vision*, Austin, TX, 21–22 Feb. 2007, pp. 4–9.

[73] G.G. Gordon, "Face recognition based on depth maps and surface curvature," *Proc. SPIE, Geometric Methods in Comput. Vis.*, vol. 1570, pp. 234–247, Sept. 1991.

[74] X. Lu, A.K. Jain, and D. Colbry, "Matching 2.5d face scans to 3d models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 31–43, 2006.

[75] B. Achermann and H. Bunke, "Classifying range images of human faces with hausdorff distance," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2000, vol. 2, pp. 809–813.

[76] L.R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, July 1945.

[77] M.P. Sampat, Z. Wang, G.J. Whitman, T. Stephens, M.K. Markey, and A.C. Bovik, "Measuring intra- and inter-observer agreement in identifying and localizing structures in medical images," in *Proc. IEEE Int. Conf. Image Processing*, Atlanta, GA, 8–11 Oct. 2006, pp. 81–84.

[78] L. Zhang, Z. Guo, Z. Wang, and D. Zhang, "Palmprint verification using complex wavelet transform," in *Proc. IEEE Int. Conf. Image Processing*, Sept. 2007, vol. 2, pp. 417–420.

[79] A. Kong and D. Zhang, "Competitive coding scheme for palmprint verification," in *Proc. IEEE Int. Conf. Pattern Recognition*, Cambridge, U.K., 2004, vol. 1, pp. 520–523.