in *Handbook of Image and Video Processing*, 2nd edition, Al Bovik, ed., Academic Press, 2005

8.3

Structural Approaches to Image Quality Assessment

Zhou Wang

New York University

Alan C. Bovik

The University of Texas at Austin

Eero P. Simoncelli

New York University

1 Introduction

2 The Structural Similarity Index

3 Image Quality Assessment Using the Structural Similarity Index

4 Validating Image Quality Measures

5 Concluding Remarks

References

1 Introduction

Digital image signals are typically represented as two-dimensional (2D) arrays of discrete signal samples. If we rearrange the signal samples into a one-dimensional (1D) vector, then every image becomes a single point in a high-dimensional *image space*, whose dimension equals the number of samples in the image signal. It has been pointed out that the cluster of *natural* image signals occupies an extremely tiny portion of such an image space [1, 2]. During its long evolution and development processes, the human visual system (HVS) has been extensively exposed to the *natural* visual environment, and a variety of evidence has shown that the HVS is highly adapted to extract useful information from natural scenes [3]. An image-quality metric, which aims to predict the quality evaluation behaviour of the HVS, would also need to be "adapted" to the properties of natural image signals.

One distinct feature that makes natural image signals different from a "typical" image randomly picked from the image space is that they are highly structured – the signal samples exhibit strong dependencies amongst themselves. These dependencies carry important information about the structures of objects in the visual scene. An image-quality metric that ignores such dependencies may fail to provide effective predictions of image quality. We will use the Minkowski error metric as an example. In the spatial domain, the Minkowski metric between a reference image **x** (assumed to have perfect quality) and a distorted image **y** is defined as

$$\mathrm{E_p} = \left(\sum_{i=1}^{N} | x_i - y_i |^p \right)^{1/p}, \qquad (1)$$

where x_i and y_i are the i-th samples in images **x** and **y**, respectively, N is the number of image samples, and p refers to the degree of power and varies in the range of $p \in [1, \infty)$. In Fig. 1, we show two distorted images generated from the same original image. The first distorted image was obtained by adding a constant

number to all signal samples, and the second was generated using the same method except that the signs of the constant are randomly chosen to be positive or negative. It can be easily shown that the Minkowski metrics between the original image and both of the distorted images are exactly the same, no matter what power p is used. However, the visual quality of the two distorted images is drastically different. Another example is shown in Fig. 2, where image B was generated by adding independent white Gaussian noise to the original texture image A. In image C, the signal sample values remained the same as in image A, but the spatial ordering of the samples has been changed (through a sorting procedure). Image D was obtained from image B, by following the same reordering procedure used to create image C. Again, the Minkowski metrics between images A and B and images C and D are exactly the same, no matter what power p is chosen. However, image D appears to be significantly noisier than image B.

There are different ways to explain the apparent failure of the Minkowski metric in the above examples. One way is to use a set of psychophysical features of human vision (see the discussions about ''Contrast Sensitivity Function" and "Contrast Masking" in Chapter 8.2). Here, we provide a more direct explanation based on the mathematical properties of the Minkowski metric. Notice that an implicit assumption of the Minkowski metric is that all signal samples are independent. As a result, the ordering of the signal samples has no effect on the overall distortion measurement. This is in sharp contrast to the fact that natural image signals are highly structured; indeed, the ordering and pattern of the signal samples carry most of the visual information in the image. Consequently, a "correct" image-quality measure would need to be able to capture the structural information or sense the structural changes in the image signals.

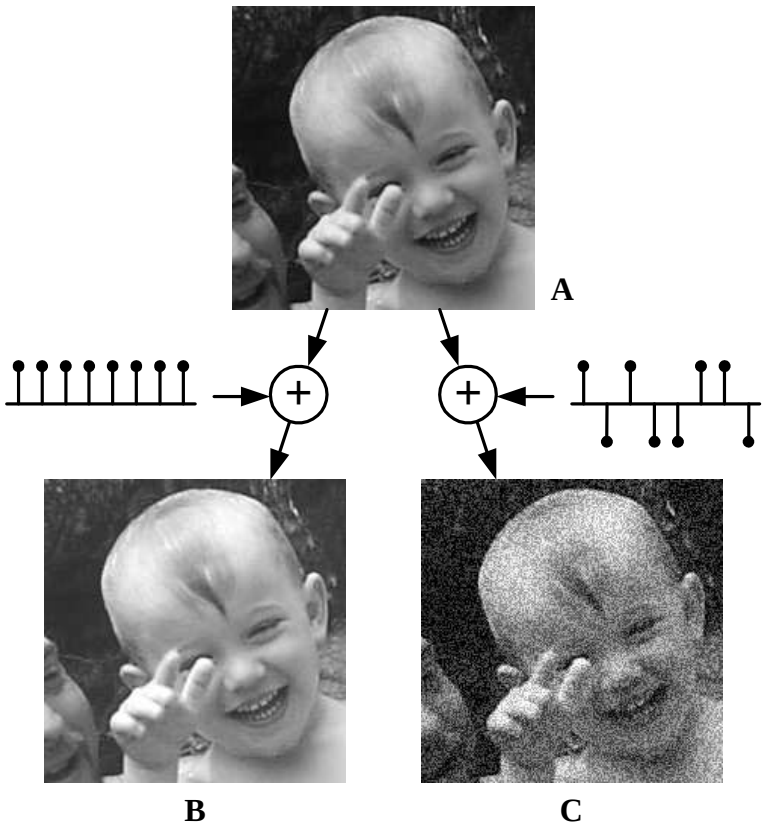

FIGURE 1 Failure of the Minkowski metric for image quality prediction. **A:** original image; **B** distorted image by adding a positive constant; **C** distorted image by adding the same constant, but with random sign. Images **B** and **C** have the same Minkowski metric with respect to image **A**, but drastically different visual quality.

Figure 3 shows one potential solution to overcome this. The idea is to apply an image transform prior to the Minkowski metric, so that the signal samples in the transform domain become independent (or at least decorrelated). An additional requirement of the transform T is that it must be lossless or "visually" lossless, in the sense that all the important visual information is preserved after the transform (presumably, there should exist an inverse transform that can reconstruct the image signals in the spatial domain). Since such a transform can decouple the dependencies between image signal samples without losing

important visual information, one may say that the "structure" of the image signal is well captured by the transform domain representation.

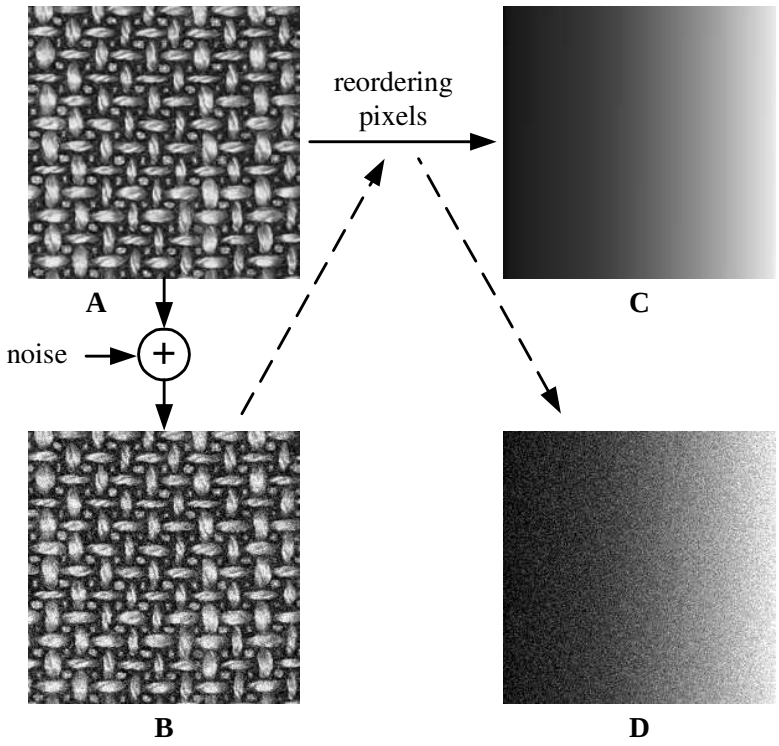

FIGURE 2 Failure of the Minkowski metric for image quality prediction. **A:** original texture image; **B** distorted image by adding independent white Gaussian noise; **C** reordering of the pixels in image **A** (by sorting pixel intensity values); **D** reordering of the pixels in image **B**, by following the same reordering used to create image **C**. The Minkowski metrics between images **A** and **B** and images **C** and **D** are the same, but image **D** appears much noisier than image **B**.

The framework shown in Fig. 3 presents interesting an analogy to the framework of perceptual image quality metrics presented in Chapter 8.2, in which, if all the processing stages before "error pooling" are combined into a single image transform, then the two frameworks can be made identical. Interestingly, the framework presented there originates from a substantially different motivation – simulating the computational aspects of the early stages of the HVS (see Chapter

4.1). Such an analogy is sensible from the viewpoint of computational neuroscience. In that context, it has been conjectured decades ago that the role of early biological sensory systems is to remove redundancies in the sensory input, resulting in a set of neural responses that are statistically independent, known as the "efficient coding" principle [3, 4].

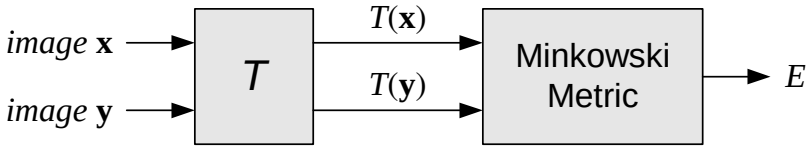

FIGURE 3 An image transform prior to an Minkowski metric may potentially reduce the dependencies between signal samples, thus improve an image quality metric.

The question that follows is then: can the image transforms (prior to the Minkowski error pooling stage) based on the current understanding of the HVS effectively decouple the dependencies between the input signal samples? Note that most recent models of early vision are based on multi-scale, bandpass and oriented linear transforms. These transforms, loosely referred to as "wavelet transforms," can reduce the correlations between signal samples as compared to spatial domain representations. However, empirical studies have shown that strong dependencies still exist between the intra- and interchannel wavelet transform coefficients of natural images (see Chapter 4.7). In fact, state-of-the-art wavelet image compression techniques achieve their success by exploiting these strong dependencies (see Chapter 5.4). In order to further reduce such signal dependencies, nonlinear operations must be applied. In fact, it has been shown that adding certain nonlinear gain control processes after the front end of linear wavelet transforms can significantly reduce signal dependencies [5-8]. The parameters of these gain control models may be tuned using psychophysical experimental data to account for visual masking effects [5, 6] (see Chapter 8.2 for a

description of visual masking). They may also be optimized to maximize the statistical dependencies between the wavelet coefficients obtained from a set of training natural images [7, 8]. Recent models have also been developed to jointly optimize statistical and perceptual dependencies [9, 10]. It remains to be seen the degree to which these models can improve the performance of current image quality assessment systems.

This chapter focuses on a different approach to image quality assessment: *structural similarity-based* methods [11]. Instead of attempting to develop an ideal transform that can fully decouple signal dependencies as suggested in Fig. 3, these methods replace the Minkowski error metric with different measurements that are adapted to the structures of the reference image signal. In the next section, we formulate structural similarity index algorithms and describe the intuition behind their design. We demonstrate how these algorithms are applied to image quality assessment in Section 3. In Section 4, we discuss an efficient approach to test the performance of image quality measures. This approach effectively reveals the perceptual implications of the structural similarity approach. Finally, concluding remarks are given in Section 5.

2 The Structural Similarity Index

The most fundamental principle underlying structural approaches to image quality assessment is that the HVS is highly adapted to extract structural information from the visual scene, and therefore a measurement of structural similarity (or distortion) should provide a good approximation to perceptual image quality. Depending on how *structural information* and *structural distortion* are defined, there may be different ways to develop image quality assessment algorithms. The structural similarity (SSIM) index is a specific implementation from the perspective of image formation.

To understand the intuition of the SSIM index method, let us again examine the image space described in the last section. In Fig. 4, a reference image (original "Einstein" image) is represented as a vector in the image space. Any image distortion can be interpreted as adding a distortion vector to the central reference image vector. In particular, the distortion vectors with the same length define an equal-mean squared error (MSE) hypersphere in the image space. However, as shown in Fig. 4, images that reside on the same hypersphere may have dramatically different visual quality. This implies that the length of a distortion vector does not suffice as a useful image quality measure, and that the directions of these vectors have more important perceptual meanings. Some insights can be found from the perspective of image formation. Recall that the luminance of the surface of an object being observed is the product of the illumination and the reflectance, but the structures of the objects in the scene are independent of the illumination. Consequently, we wish to separate the influence of illumination from the remaining information that represents object structures. Intuitively, the major impact of illumination change in the image is the variation of the average local luminance and contrast, and such variation should not have a strong effect on perceived image quality. This is confirmed by Fig. 4, where the images with only luminance or contrast changes have much better quality than the other images with severe "structural" distortions.

Figure 5 illustrates how luminance and contrast changes can be separated from structural distortions in the image space. Luminance changes can be characterized by moving along the direction defined by $x_1 = x_2 = \cdots = x_N$, which is perpendicular to the hyperplane of $\sum_{i=1}^{N} x_i = 0$. Contrast changes are defined by the direction $\mathbf{x} - \bar{x}$. In the image space, the two vectors that determine luminance and contrast changes span a 2D subspace (a plane), which is adapted to the reference image vector **x**. For example, in Fig. 4, the plane determined by the reference image A contains not only the reference image itself, but also images B

and C. The remaining image distortion corresponds to rotating such a plane by a certain angle, which we interpret as the structural change in Fig. 5.

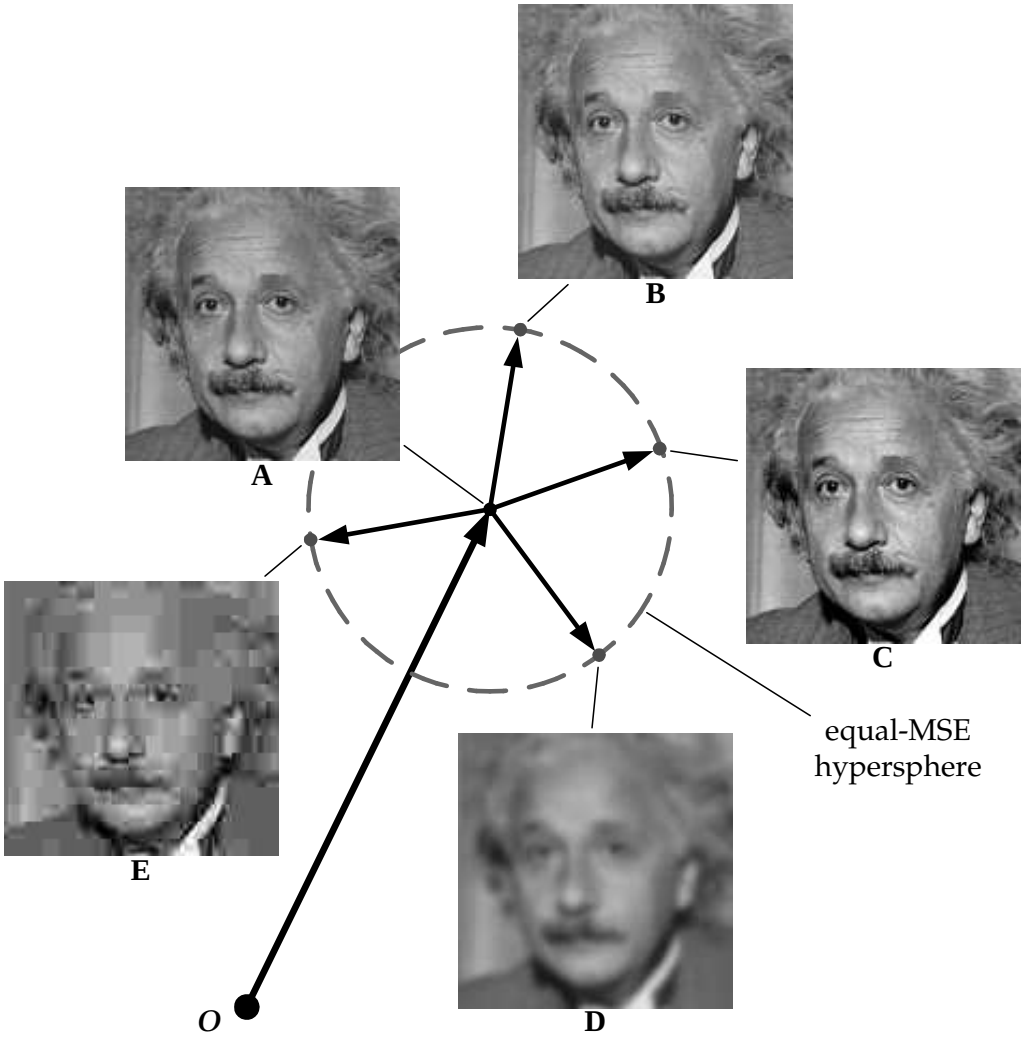

FIGURE 4 An image can be represented as a vector in the image space, whose dimension equals the number of pixels in the image. Images with the same mean squared error (MSE) with respect to the original image constitute a hypersphere in the image space, but images reside on the same hypersphere have dramatically different visual quality. **A:** original image; **B** mean shifted image, MSE = 144; **C** contrast stretched image, MSE = 144; **D** blurred image, MSE = 144; **E** JPEG compressed image, MSE = 142.

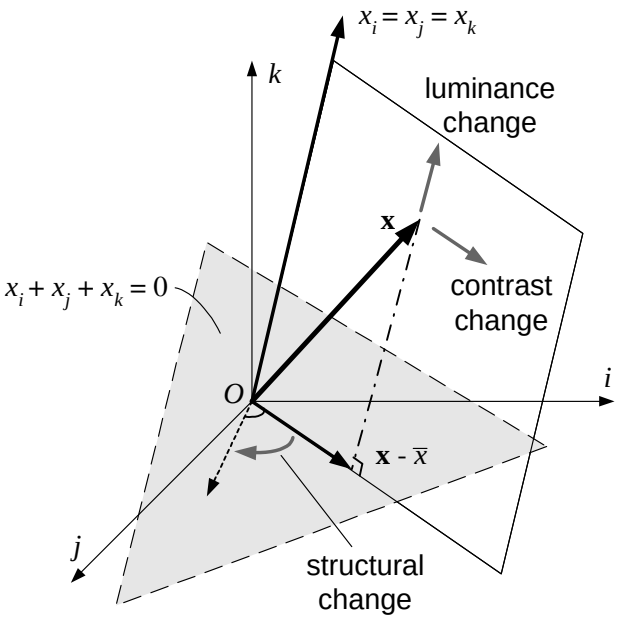

FIGURE 5 Separation of luminance, contrast and structural changes from a reference image **x** in the image space. This is an illustration in three-dimensional space. In practice, the number of dimensions is equal to the number of image pixels.

A system diagram of the SSIM index algorithm is shown in Fig. 6. First, the luminance of each signal is estimated as the mean intensity:

$$\mu_x = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \,. \tag{1}$$

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

$$l(\mathbf{x}, \mathbf{y}) = l(\mu_x, \mu_y) \,. \tag{2}$$

Second, we remove the mean intensity from the signal. The resulting signal $\mathbf{x} - \mu_x$ corresponds to the projection of vector **x** onto the hyperplane of $\sum_{i=1}^{N} x_i = 0$, as illustrated in Fig. 5. We use the standard deviation as an estimate of the signal contrast. An unbiased estimate in discrete form is given by

$$\sigma_x = \left(\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)^2 \right)^{1/2} . \tag{3}$$

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

$$c(\mathbf{x}, \mathbf{y}) = c(\mu_x, \mu_y) . \tag{4}$$

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

$$s(\mathbf{x}, \mathbf{y}) = s\left(\frac{\mathbf{x} - \mu_x}{\sigma_x}, \frac{\mathbf{y} - \mu_y}{\sigma_y} \right) . \tag{5}$$

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

$$S(\mathbf{x}, \mathbf{y}) = f(l(\mathbf{x}, \mathbf{y}), c(\mathbf{x}, \mathbf{y}), s(\mathbf{x}, \mathbf{y})) . \tag{6}$$

An important point is that the three components are relatively independent, which is physically sensible because the change of luminance and/or contrast has little impact on the structures of the objects in the scene.

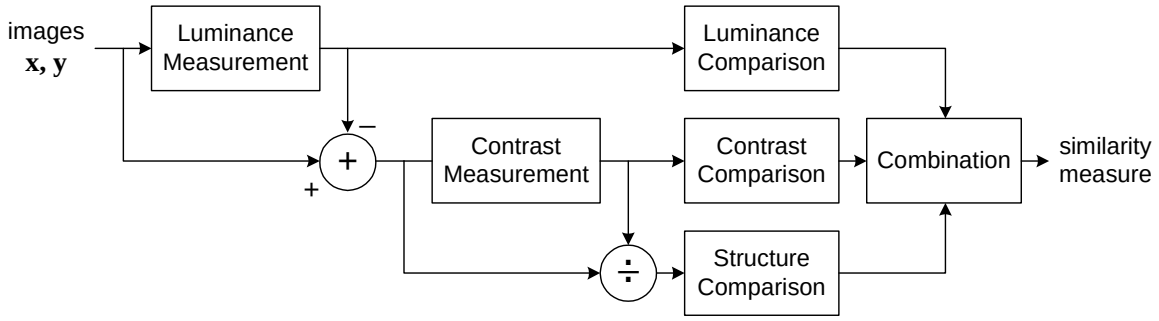

FIGURE 6 Diagram of image similarity measurement system. (Adapted from [11].)

To complete the definition of the similarity measure in Eq. (6), we need to define the three functions $l(\mathbf{x}, \mathbf{y})$, $c(\mathbf{x}, \mathbf{y})$ and $s(\mathbf{x}, \mathbf{y})$, as well as the combination function $f(\cdot)$. In addition, we also would like the similarity measure to satisfy the following conditions:

1. Symmetry: $S(\mathbf{x}, \mathbf{y}) = S(\mathbf{y}, \mathbf{x})$. When quantifying the similarity between two signals, exchanging the order of the input signals should not affect the resulting measurement.
2. Boundedness: $S(\mathbf{x}, \mathbf{y}) \leq 1$. An upper bound can serve as an indication of how close the two signals are to being perfectly identical. Notice that signal-to-noise ratio type of measurements is typically unbounded.
3. Unique maximum: $S(\mathbf{x}, \mathbf{y}) = 1$ if and only if $\mathbf{x} = \mathbf{y}$. The perfect score is achieved only when the signals being compared are identical. In other words, the similarity measure should quantify any variations that may exist between the input signals.

The luminance comparison is defined as

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{7}$$

where the constant C_1 is included to avoid instability when $\mu_x^2 + \mu_y^2$ is very close to zero. One good choice is

$$C_1 = (K_1 L)^2, \tag{8}$$

where L is the dynamic range of the pixel values (255 for 8-bit grayscale images), and $K_1 \ll 1$ is a small constant. Similar considerations also apply to contrast comparison and structure comparison as described later. Equation (7) is easily seen to obey the three properties listed above.

Equation (7) is also connected with Weber's law, which has been widely used to model light adaptation (also called luminance masking) in the HVS (see chapter 8.2). According to Weber's law, the magnitude of a just-noticeable luminance change ΔI is approximately proportional to the background luminance I for a wide range of luminance values. In other words, the HVS is sensitive to the relative rather than the absolute luminance change. Letting R represent the ratio of the

luminance of the distorted signal relative to the reference signal, then we can write $\mu_y = R\mu_x$. Substituting this into Eq. (7) gives

$$l(\mathbf{x},\mathbf{y}) = \frac{2R}{1+R^2+C_1/\mu_x^2} \ . \tag{9}$$

If we assume C_1 is small enough (relative to μ_x^2) to be ignored, then $l(\mathbf{x}, \mathbf{y})$ is a function only of R instead of $\Delta I = \mu_y - \mu_x$. In this sense, it is qualitatively consistent with Weber's law. In addition, it provides a quantitative measurement for the cases when the luminance change is higher than the visibility threshold, which is out of the application scope of Weber's law.

The contrast comparison function takes a similar form:

$$c(\mathbf{x},\mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2+\sigma_y^2+C_2} \ , \tag{10}$$

where C_2 is a non-negative constant

$$C_2 = (K_2 L)^2 \tag{11}$$

and K_2 satisfies $K_2 \ll 1$. This definition again satisfies the three properties listed above. An important feature of this function is that with the same amount of $\Delta\sigma = \sigma_y - \sigma_x$, this measure is less sensitive to the case of high base contrast σ_x than low base contrast. This is related to the contrast masking feature of the HVS (see Chapter 8.2).

Structure comparison is conducted after luminance subtraction and contrast normalization. Geometrically, we can associate the structures of the two images with the direction of the two unit vectors $(\mathbf{x}-\mu_x)/\sigma_x$ and $(\mathbf{x}-\mu_y)/\sigma_y$, each lying in the hyperplane $\sum_{i=1}^{N} x_i = 0$ as illustrated in Fig. 5. The angle between the two vectors provides a simple and effective measure to quantify structural similarity. In particular, the correlation coefficient between **x** and **y** corresponds to the cosine of the angle. Thus, we define the structure comparison function as:

$$s(\mathbf{x},\mathbf{y}) = \frac{2\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \ . \tag{12}$$

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

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) \ . \tag{13}$$

Notice that $s(\mathbf{x}, \mathbf{y})$ can take on negative values. As will be shown in later examples, the negative structural similarity values correspond to the cases that the local image structures are inverted.

Finally, we combine the three comparisons of Eqs. (7), (10) and (12). The result is a class of image similarity measures which we collectively *Structural SIMilarity* (*SSIM*) *Indices* between signals **x** and **y**:

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^{\alpha} \cdot [c(\mathbf{x},\mathbf{y})]^{\beta} \cdot [s(\mathbf{x},\mathbf{y})]^{\gamma} \tag{14}$$

where $\alpha > 0$, $\beta > 0$ and $\gamma > 0$ are parameters used to adjust the relative importance of the three components. It is easy to verify that this definition satisfies the three conditions given above. In what follows, we set $\alpha = \beta = \gamma = 1$ and $C_3 = C_2 / 2$. This results in a specific SSIM index [11]:

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \ . \tag{15}$$

The difference between the SSIM indices and previous error metrics proposed for image quality assessment may be better understood geometrically in a vector space of signal components as in Fig. 7. Here, the signal components can be either image pixel intensities or other extracted features such as transformed linear coefficients. Figure 7 shows equal-distortion contours drawn around three different example reference vectors, each of which could, for example, represent the local content of one reference image. For the purpose of illustration, we show only a 2D space, but in general the dimensionality should match that of the signal components being compared. Each contour represents a set of test signals with

equal distortion relative to the respective reference signal. Figure 7A shows the result for a simple Minkowski metric. Each contour has the same size and shape (a circle here, as we are assuming $p = 2$). That is, perceptual distance corresponds to Euclidean distance. Figure 7B shows a Minkowski metric in which different signal components are weighted differently. This could be, for example, weighting according to the contrast sensitivity function, as is common in many quality assessment models. Here the contours are ellipses, but still are all the same size. More advanced quality measurement models may incorporate contrast masking behaviors, which has the effect of rescaling the equal-distortion contours according to the signal magnitude, as shown in Fig. 7C. This may be viewed as a simple type of *adaptive* distortion measure: it depends not just on the difference between the signals, but also on the signals themselves. Figure 7D shows a combination of contrast masking (magnitude weighting) followed by component weighting.

The SSIM index gives a different picture. In the hyperplane of $\sum_{i=1}^{N} x_i = 0$, the SSIM index compare the vectors $(\mathbf{x} - \mu_x)$ and $(\mathbf{x} - \mu_y)$ with two independent quantities: the vector lengths, and their angles. Thus, the contours will be aligned with the axes of a polar coordinate system. Figures 7E and F show two examples of this, computed with different exponents (for β and γ). Again, this may be viewed as an *adaptive* distortion measure, but unlike the other models being compared, both the size and the shape of the contours are adapted to the underlying signal.

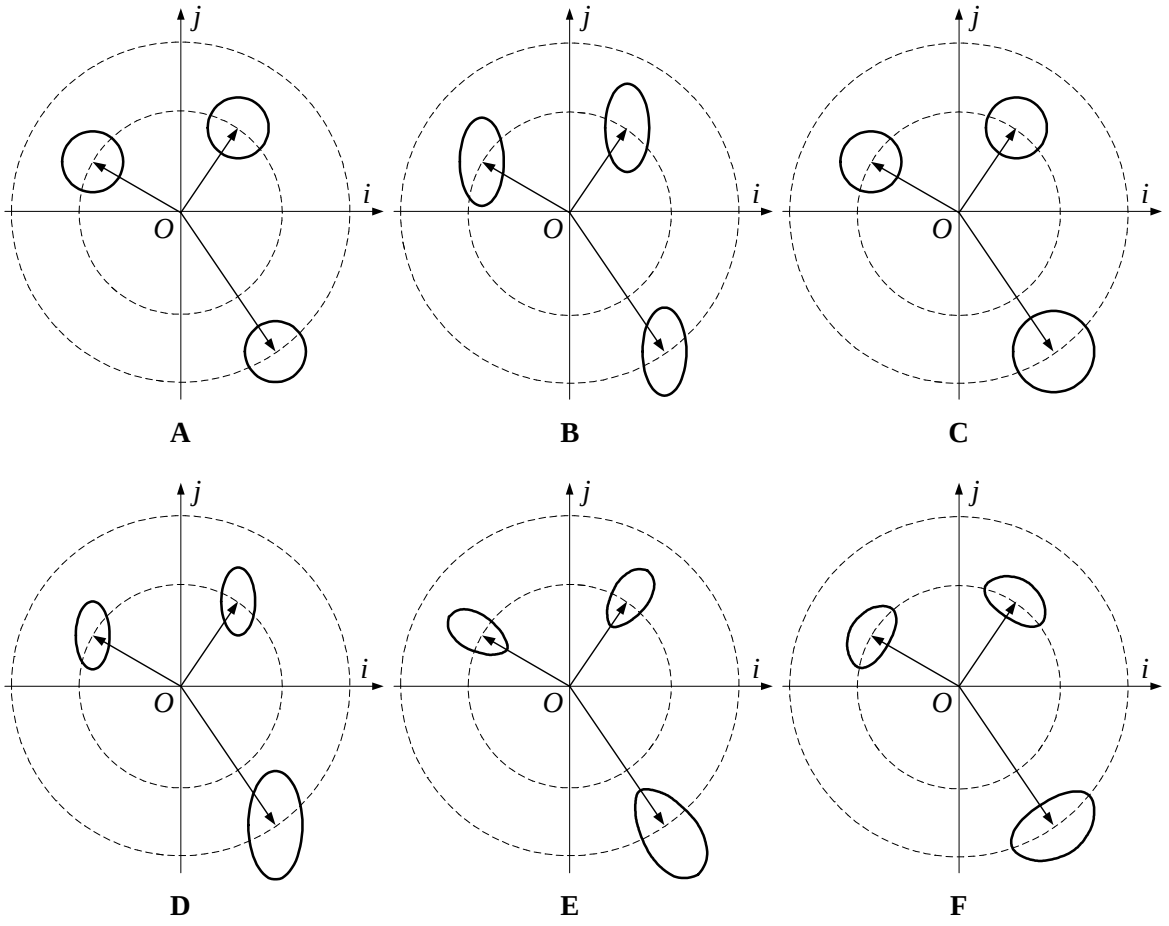

FIGURE 7 Equal-distortion contours in the image space for different quality measurement systems. **A:** Minkowski error measurement systems (assuming $p = 2$ in the illustration); **B** component-weighted Minkowski error measurement systems; **C** magnitude-weighted Minkowski error measurement systems; **D** magnitude- and component-weighted Minkowski error measurement systems; **E** structural similarity index (SSIM) measurement system (with more emphasis on structural comparison); **F** SSIM measurement system (with more emphasis on contrast comparison). Each image is represented as a vector, whose entries are image components. This is an illustration in two-dimensional space. In practice, the number of dimensions is equal to the number of image components used for comparison (e.g, the number of pixels or transform coefficients). (From [11].)

3 Image Quality Assessment Using a Structural Similarity Index

The SSIM indices measure the structural similarity between two image signals. If one of the image signals is regarded as of perfect quality, then the SSIM index can be viewed as an indication of the quality of the other image signal being compared. When applying the SSIM index approach to large-size images, it is useful to compute it locally rather than globally. The reason is manifold. First, statistical features of images are usually spatially nonstationary. Second, image distortions, which may or may not depend on the local image statistics, may also vary across space. Third, due to the non-uniform retinal sampling feature of the HVS (see Chapters 4.1), at typical viewing distances, only a local area in the image can be perceived with high resolution by the human observer at one time instance. Finally, localized quality measurement can provide a spatially varying quality map of the image, which delivers more information about the quality degradation of the image. Such a quality map can be used in different ways. It can be employed to indicate the quality variations across the image. It can also be used to control image quality for space-variant image processing systems, e.g., region-of-interest image coding and foveated image processing [12].

In early instantiations of the SSIM index approach [13, 14], the local statistics μ_x, σ_x and σ_{xy} [Eqs. (1), (3) and (13)], are computed within a local 8×8-square window. The window moves pixel-by-pixel from the top-left corner to the bottom-right corner of the image. At each step, the local statistics and SSIM index are calculated within the local window. One problem with this method is that the resulting SSIM index map often exhibits undesirable "blocking" artifacts as exemplified by Fig. 8C. Such kind of "artifacts" are not desirable because it is created from the choice of the quality measurement method (local square window), but not from image distortions. In [11], a circular-symmetric Gaussian

weighting function $\mathbf{w} = \{w_i \mid i = 1,2,\cdots N\}$ with unit sum ($\sum_{i=1}^{N} w_i = 1$) is adopted. The estimates of local statistics, μ_x, σ_x and σ_{xy}, are then modified accordingly:

$$\mu_x = \sum_{i=1}^{N} w_i x_i \ , \tag{16}$$

$$\sigma_x = \left(\sum_{i=1}^{N} w_i (x_i - \mu_x)^2 \right)^{1/2} , \tag{17}$$

$$\sigma_{xy} = \sum_{i=1}^{N} w_i (x_i - \mu_x)(y_i - \mu_y) \ . \tag{18}$$

With such a smoothed windowing approach, the quality maps exhibit a locally isotropic property as demonstrated in Fig. 8D.

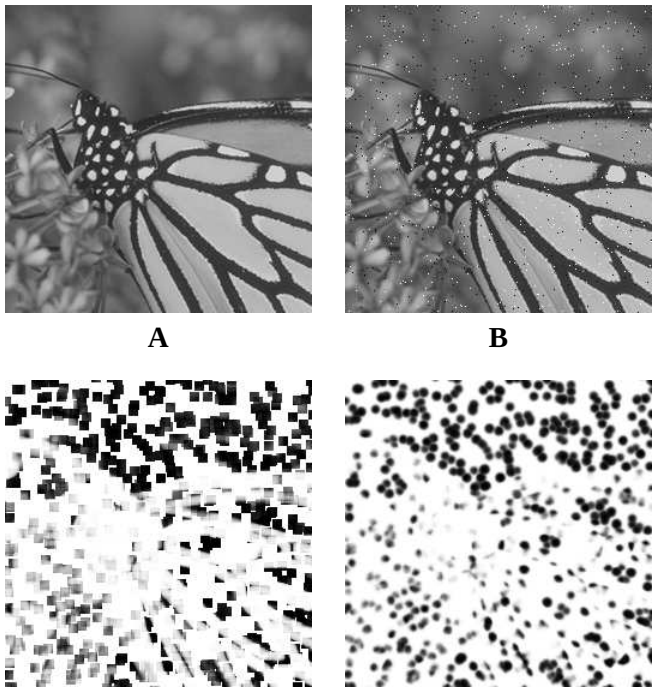

FIGURE 8 The effect of local window shape on structural similarity index (SSIM) index map. **A:** original image; **B** impulsive noise contaminated image; **C** SSIM index map using square windowing approach; **D** SSIM index map using smoothed windowing approach. In both SSIM index maps, brighter indicates better quality.

Figure 9 shows the SSIM index maps of a set of sample images with different types of distortions. The absolute error map for each distorted image is also included for comparison. The SSIM index and absolute error maps have been adjusted, so that brighter always indicates better quality in terms of the given quality/distortion measure. It can be seen that the distorted images exhibit variable quality across space. For example, in image B, the noise over the face region appears to be much more significant than that in the texture regions. However, the absolute error map (D) is completely independent of the underlying image structures. By contrast, the SSIM index map (C) gives perceptually consistent prediction. In image F, the bit allocation scheme of low bit-rate JPEG2000 compression leads to smooth representations of detailed image structures. For example, the texture information of the roof of the building as well as the trees is lost. This is very well indicated by the SSIM index map (G), but cannot be predicted from the absolute error map (H). Some different types of distortions are caused by low bit-rate JPEG compression. In image J, the major distortions we observe are the blocking effect in the sky and the artifacts around the outer boundaries of the building. Again, the absolute error map (L) fails to provide useful prediction, and the SSIM index map (K) successfully predicts image-quality variations across space. From these sample images, we see that an image-quality measure as simple as the SSIM index can adapt to various kinds of image distortions and provide perceptually consistent quality predictions.

The final step of an image-quality measurement system is to combine the quality map into one single quality score for the whole image. A convenient way is to use a weighted summation. Let $\mathbf{X}$ and $\mathbf{Y}$ be the two images being compared, and $\mathrm{SSIM}(\mathbf{x}_j, \mathbf{y}_j)$ be the local SSIM index evaluated at the j-th local sample [i.e., $\mathrm{SIM}(\mathbf{x}_j, \mathbf{y}_j)$ for all j's constitutes a SSIM index map as demonstrated in Fig. 9], then the SSIM index between $\mathbf{X}$ and $\mathbf{Y}$ is defined as:

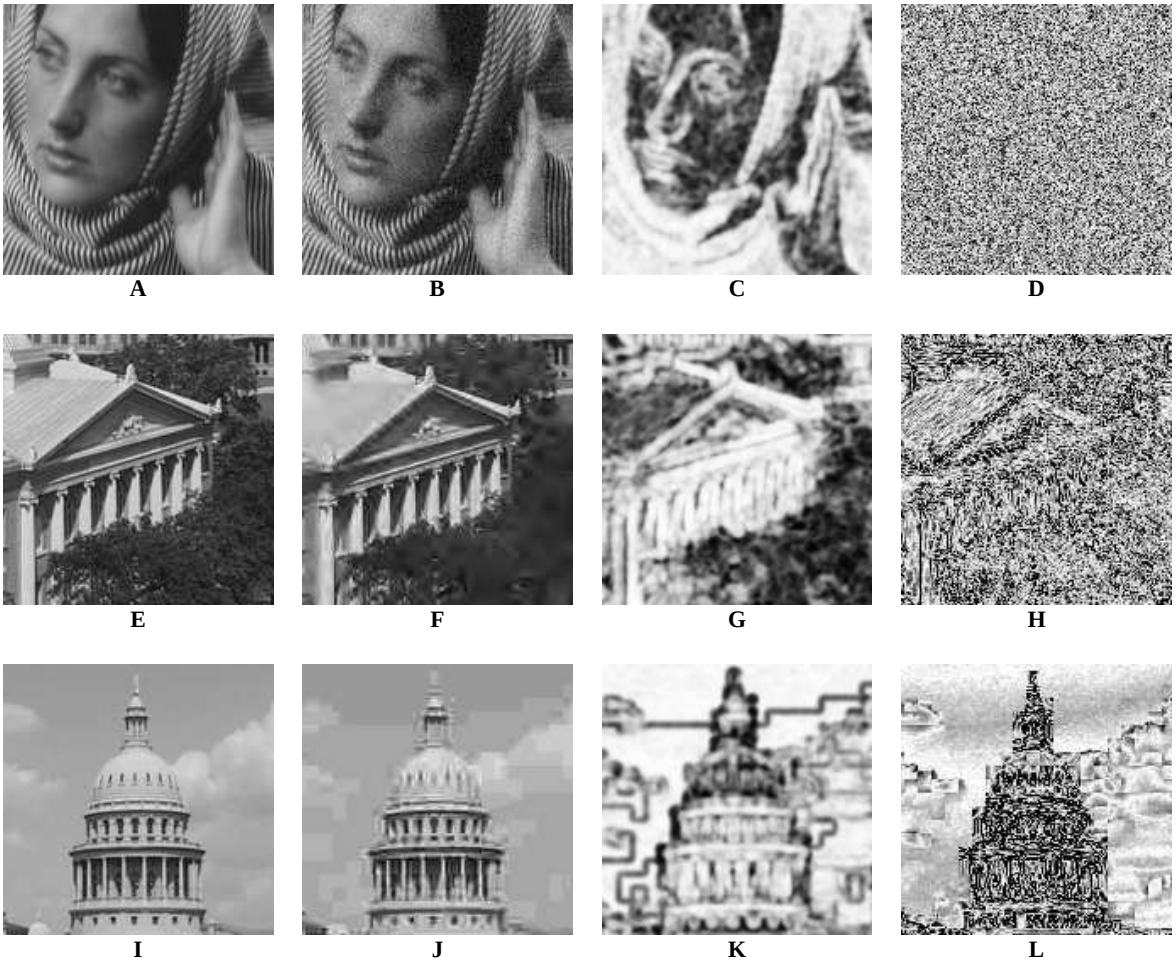

FIGURE 9 Sample distorted images and their quality/distortion maps (images are cropped to 160×160 for visibility); (**A, E, I**) original images; **B** Gaussian noise contaminated image; **F** JPEG2000 compressed image; **J** JPEG compressed images; (**C, G, K**) structural similarity index (SSIM) index maps of the distorted images, where brightness indicates the magnitude of the local SSIM index (squared for visibility); (D, H, L): absolute error maps of the distorted images, where darkness indicates the absolute value of the local pixel difference. Note that in all quality/distortion maps (**C, D, G, H, K** and **L**), brighter indicates better quality in terms of the underlying quality/distortion measure.

$$\mathrm{SSIM}(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{j=1}^{N_S} W_j(\mathbf{x}_j, \mathbf{y}_j) \cdot \mathrm{SSIM}(\mathbf{x}_j, \mathbf{y}_j)}{\sum_{j=1}^{N_S} W_j(\mathbf{x}_j, \mathbf{y}_j)}, \tag{19}$$

where N_S is the number of samples in the quality map, and $W_j(\mathbf{x}_j, \mathbf{y}_j)$ is the weight given to the j-th sample. If all the samples in the quality map are equally weighted, then $W_j(\mathbf{x}_j, \mathbf{y}_j) \equiv 1$. This results in the mean SSIM (MSSIM) measure employed in [11].

There are two cases in which nonuniform weighting is desirable. First, depending on the application, some prior knowledge about the importance of different regions in the image is available, and such an importance map can be converted into a weighting function. For example, object-based region-of-interest image processing systems often segment the objects in the scene and give different objects different importance. In a foveated image-processing system [12], the weighting function can be determined according to the foveation feature of the HVS (i.e., the visual resolution decreases gradually as a function of the distance from the fixation point). Note that the weighting function here is determined only by the spatial location j, but independent of the local image content $\mathbf{x}_j$ and $\mathbf{y}_j$. In the second case, the image content also plays a role. It has been observed that different image textures attract human fixations with varying degrees, and therefore different weights can be assigned. In [15], a variance-weighted weighting function is used, where

$$W(\mathbf{x}, \mathbf{y}) = \sigma_x^2 + \sigma_y^2 + C_2. \tag{20}$$

It was observed that this weighting function is useful to balance the extreme case that severe high-variance distortions concentrate at some small areas in the image.

4 Validating Image Quality Measures

Validation is an important step towards successful development of practical image quality measurement systems. The most standard form of validation is to compare objective quality measures with ratings by human subjects on an extensive database of images. Typically, the total number of images in a database of reasonable size is in the order of hundreds. Gathering reliable data in such a large-scale subjective experiment is a very expensive task. To hold down the number of subjective comparisons, the form of image distortions is usually highly restricted. Therefore, despite the substantial time involved in collecting psychophysical data in these experiments, there is no guarantee that the test results on these restricted databases provide a sufficient test for a "general-purpose" image quality assessment algorithm.

In [15], an alternative approach was formulated to evaluate the relative strength and weaknesses of image-quality measures with a much smaller number of subjective comparisons. The key idea is to conduct subject test on *synthesized* images that best differentiate two candidate image-quality measures. These synthesized images are obtained by a searching procedure in the image space, rather than collecting a large number of images with known types of distortions. In previous work, the idea of synthesizing images for subjective testing has been employed by the "synthesis-by-analysis" methods of assessing statistical texture models, in which the model is used to generate a texture with statistics matching an original texture, and a human subject then judges the similarity of the two textures [16-20]. In the context of image quality assessment, a similar concept has also been used for qualitatively demonstrating the performance [5, 11, 13] and calibrating parameter settings [21]. These synthesis methods provide a very powerful and efficient means to reveal the strength and weaknesses of a model. They also provide the added benefit that the resulting images may suggest improvements of the model.

Figure 10 shows the framework of the image synthesis-based system for performance comparison of two image-quality measures, which are denoted as M_1 and M_2, respectively. For each reference image and a given initial distortion level l, the system generates two pairs of synthesized image stimuli. First, the reference image is altered according to the given initial distortion level l (e.g., white noise of a specified variance, σ_l^2, is added) to generate an initial distorted image. Second, the quality of the initial image is calculated using the two given measures M_1 and M_2, respectively. Third, the system searches for the best-/worst-quality images in terms of M_2 while constraining the value of M_1 to remain fixed. The result is a pair of images that have the same M_1 value, but potentially very different M_2 values. This procedure is also applied with the roles of the two metrics reversed, to generate the best-/worst-quality images in terms of M_1 while constraining the value of M_2 to be fixed. Finally, subjects compare the quality of the resulting two synthesized image pairs.

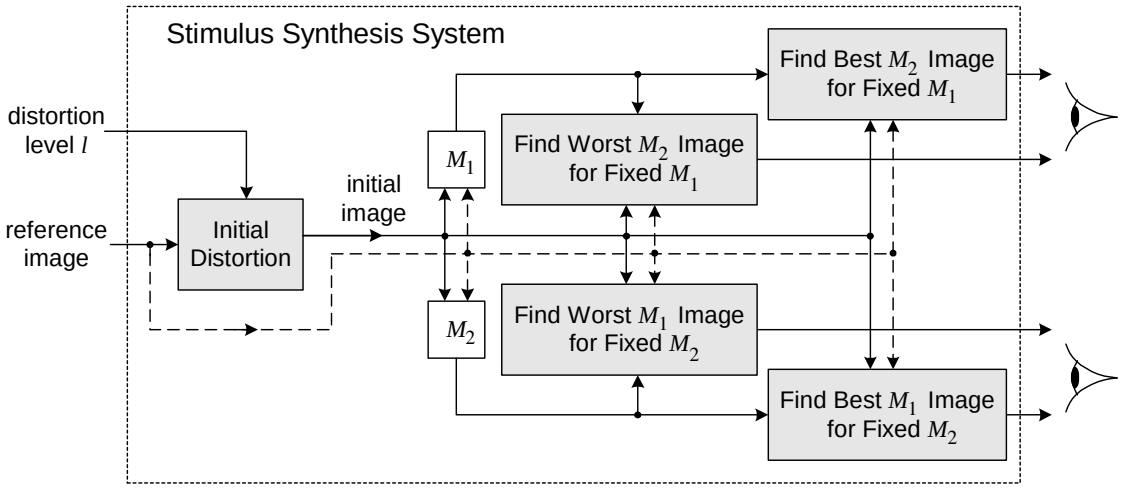

FIGURE 10 Diagram of image synthesis-based system for performance comparison of two image quality measures M_1 and M_2. (Adapted from [15].)

Figure 11 gives a more straightforward illustration of the method in the image space, in which the initial image is a common point of a level set of M_1 and a level set of M_2. As demonstrated in Fig. 11A, the goal of the image synthesis system is to start from the initial image, and move (perhaps iteratively) along the direction of increasing/decreasing the M_2 measure while constrained on the M_1 level set. Figure 11B demonstrates the reverse procedure for searching the best/worst M_1 images along the M_2 level set.

Depending on the specific formulation and complexity of the quality measures being compared, there may be various methods of finding the best-/worst-quality image in terms of one of the measures while constraining the other to be fixed. Figure 12 illustrates a single step of a constrained iterative gradient ascent/descent algorithm for optimization of M_2. Here, we denote the reference image $\mathbf{X}$ and the distorted image at the n-th iteration $\mathbf{Y}_n$ (with $\mathbf{Y}_0$ representing the initial distorted image). We compute the gradient of the two quality measures, evaluated at $\mathbf{Y}_n$:

$$\mathbf{G}_{1,n} = \vec{\nabla}_Y M_1(\mathbf{X},\mathbf{Y})\,|_{\mathbf{Y}=\mathbf{Y}_n} \qquad \mathbf{G}_{2,n} = \vec{\nabla}_Y M_2(\mathbf{X},\mathbf{Y})\,|_{\mathbf{Y}=\mathbf{Y}_n}. \tag{21}$$

We define a modified gradient direction, $\mathbf{G}_n$, by projecting out the component of $\mathbf{G}_{2,n}$ that lies in the direction of $\mathbf{G}_{1,n}$:

$$\mathbf{G}_n = \mathbf{G}_{2,n} - \frac{\mathbf{G}_{2,n}^T \mathbf{G}_{1,n}}{\mathbf{G}_{1,n}^T \mathbf{G}_{1,n}} \mathbf{G}_{1,n}. \tag{22}$$

A new distorted image is computed by moving in the direction of this vector:

$$\mathbf{Y}_n' = \mathbf{Y}_n + \lambda \mathbf{G}_n. \tag{23}$$

Finally, the gradient of M_1 is evaluated at $\mathbf{Y}_n'$, and an appropriate amount of this vector is added in order to guarantee that the new image has the correct value of M_1:

$$\mathbf{Y}_{n+1} = \mathbf{Y}_n' + \nu \mathbf{G}_{1,n}' \quad \text{s.t.} \quad M_1(\mathbf{X},\mathbf{Y}_{n+1}) = M_1(\mathbf{X},\mathbf{Y}_0). \tag{24}$$

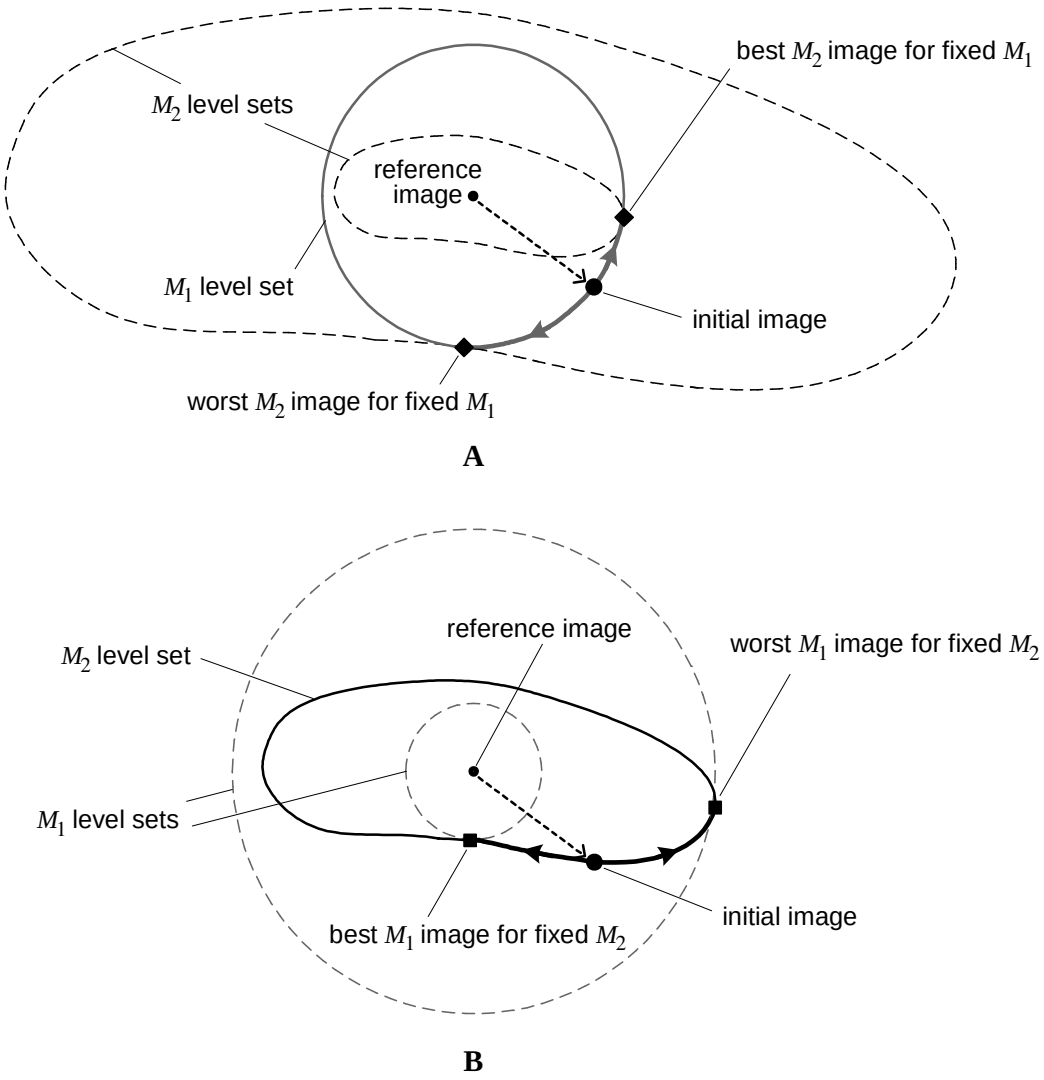

FIGURE 11 Illustration of the image synthesis approach in the image space. **A:** searching the best/worst M_2 image on M_1 level set; **B** searching the best/worst M_1 image on M_2 level set. This illustration is in two-dimensional space. In practice, the dimension equals the number of image pixels. (Adapted from [15].)

For the case of MSE, the selection of ν is straightforward, but in general it might require a 1D (line) search. During the iterations, the parameter λ is used to control the speed of convergence and ν must be adjusted dynamically so that the resulting vector does not deviate from the level set of M_1. The iteration continues until it satisfies certain convergence condition (e.g., mean squared change in the

distorted image in two consecutive iterations is less than some threshold). If Metric M_2 is differentiable, then this procedure will converge to a local maximum/minimum of M_2. In general, however, to find the global maximum/minimum is difficult (note that the dimension of the search space is equal to the number of pixels in the image), unless the quality measure satisfies certain properties (e.g., convexity or concavity).

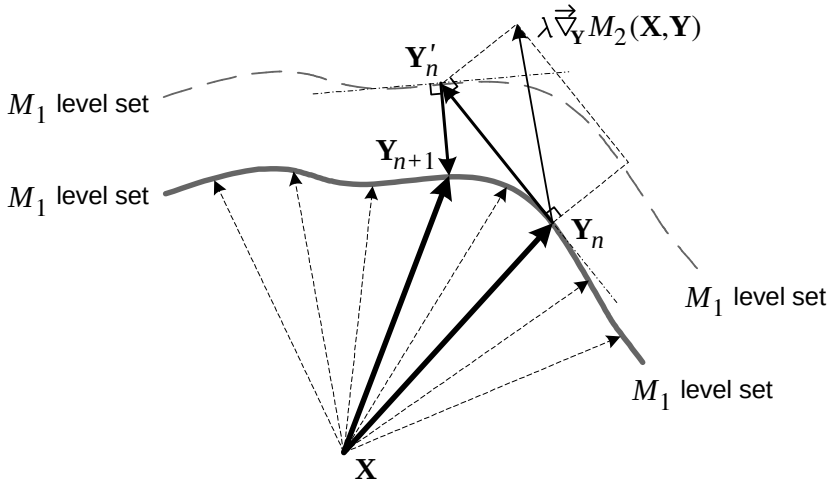

FIGURE 12 Illustration of the *n*-th iteration of the gradient ascent/descent search procedure for optimizing M_2 while constraining on the M_1 level set. This illustration is in two-dimensional space. In practice, the dimension equals the number of image pixels. (From [15].)

In Figs. 13 and 14, we demonstrate comparison of MSE against SSIM using the image synthesis-based approach. Note that the gradient-based searching approach described above requires calculating the gradients of both image quality measures with respect to the image during each iteration. Fortunately, both MSE and SSIM measures are simple enough, such that their gradients can be computed explicitly [15]. Figure 13 shows the synthesized images for performance comparison of MSE and SSIM, where the initial image is obtained by adding white Gaussian noise (σ_I^2 =1024) and variance-weighted pooling as of Eq. (20) is used for the SSIM measure. These synthesized images immediately reveal the perceptual

implications of both quality measures. In particular, MSE is blind in distinguishing between structural distortions (D) (in fact, many local image structures in image D are inverted, resulting in negative SSIM values) and luminance/contrast changes (C), but perceptually, image C has much better quality than image D. On the other hand, although the best/worst MSE images for fixed SSIM (images E and F) exhibit very different types of image distortions, the best MSE image E does not appear to be obviously better than the worst MSE image F. Similar comparisons remain consistently across a wide variety of image types, as is demonstrated in Fig. 14. Thus, we conclude that SSIM performs much better than MSE in this test.

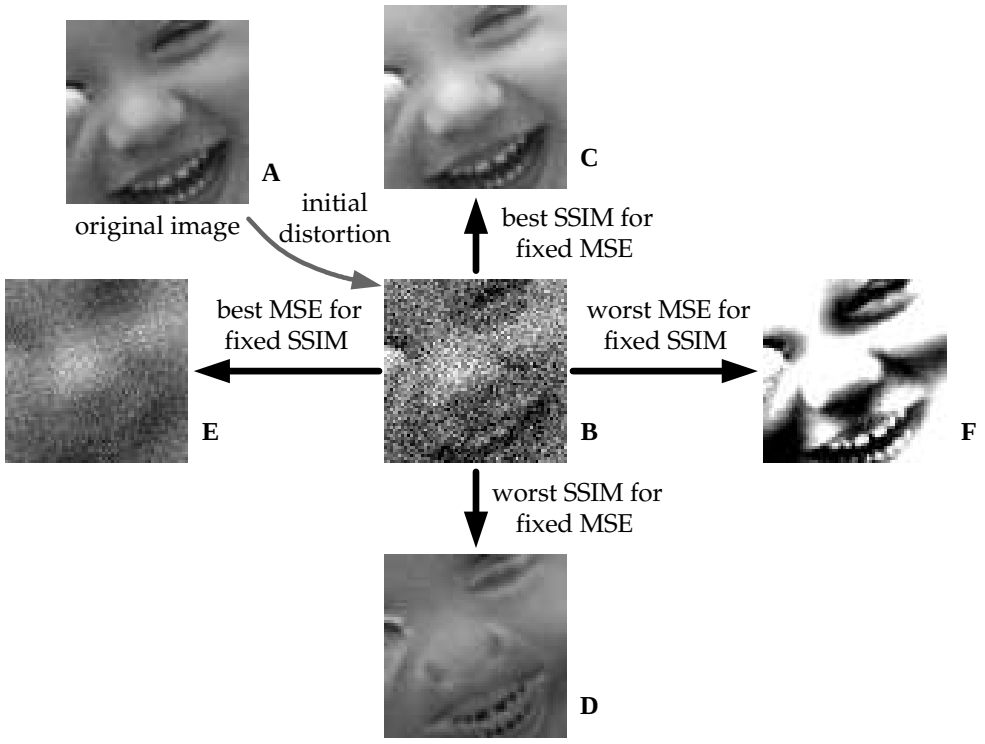

FIGURE 13 Demonstration of image stimulus synthesis for performance comparison of mean squared error and structural similarity index.

FIGURE 14 Synthesized image stimuli for performance comparison of mean squared error and structural similarity index. (Adapted from [15].)

5 Concluding Remarks

This chapter introduces the basic ideas and algorithms of structural approaches for image quality assessment. We have attempted to describe the concepts, the SSIM index algorithm, as well as the image synthesis-based performance evaluation algorithm in the image space. We demonstrate that image distortions along different directions in the image space have different perceptual meanings. The structural approaches attempt to separate the directions associated with structural distortions from those with non-structured distortions. This separation gives a new coordinate system in the image space. The new coordinate

system is not fixed as in traditional image decomposition frameworks (e.g., Fourier and wavelet types of transforms), but adapted to the underlying image structures.

In terms of the construction of image quality assessment systems, most traditional HVS-based methods are based on a *bottom-up* philosophy, which attempts to simulate the functions of the relevant components in the HVS and combine them together, in the hope that the combined system can predict the behavior of the overall HVS. The effectiveness of these methods depends on how much the HVS is understood and how accurately the simulation can be implemented. By contrast, the structural approaches are based on a *top-down* philosophy, which starts from the very top level — simulating the hypothesized functionality of the overall HVS. A top-down approach may lead to significantly simplified algorithm, but relies on the goodness of the underlying hypothesis. In particular, the basic assumption made by the structural approaches is that the HVS is highly adapted to extract structural information from the visual scene, and therefore structural distortion measure should give good prediction of perceived image quality. Current experiments have demonstrated very promising results.

Although the structural approaches are based on a substantially different design principle, we view them as complementary to, rather than opposed to, the traditional HVS-based methods. Notice that the traditional approaches often involve linear signal decompositions (e.g., the wavelet transforms), followed by local nonlinear normalization processes. These normalized transform coefficients may be considered as specific representations of the image structures. In this sense, the errors measured in normalized transform coefficients implicitly suggest the structural changes between the image signals being compared. On the other hand, the adaptive coordinate system (as demonstrated in Figs. 5 and 7) used by the SSIM approach may also be converted into an image transform, and then the SSIM index may become equivalent to an error metric after the transform. Thus,

the same framework of Fig. 3 is followed, only the image transform is adapted to local image structures. Interestingly, some divisive-normalization based masking models exhibit similar input-dependent behavior [9, 10], although precise alignment as in Figs. 7E and F is not observed. Although it is not clear at this moment, we believe it is possible that the two types of approaches may eventually converge into similar solutions.

The paradigm of structural image quality assessment is still at a preliminary stage. The current approaches can be extended in many directions. Direct extensions include video quality assessment [22], colour image quality assessment [23] and multi-scale image quality assessment [21]. Robustness to local image translation, scaling and rotation is another important issue to be studied because these transformations usually do not cause significant changes in perceived image structure (and quality). Furthermore, the SSIM index approach is quite encouraging not only because of its good image quality prediction accuracy, but also its simple formulation and low computational complexity. The simplicity makes it much more tractable than traditional methods in optimization tasks (e.g., its derivative with respect to the image can be explicitly computed [15]). Consequently, the SSIM index, and other structurally-oriented image quality assessment algorithms have great potential to be used in the future development of perceptually-optimized image processing, coding and communication systems.

References

[1] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of Optical Society of America*, vol. 4, no. 12, pp. 2379-2394, 1987.

[2] D. L. Ruderman, "The statistics of natural images," *Network: Computation in Neural Systems*, vol. 5, pp. 517-548, 1996.

[3] E. P. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, pp. 1193-1216, 2001.

[4] H. B. Barlow, "Possible principles underlying the transformation of sensory messages," in *Sensory Communication*, W. A. Rosenblith, ed., pp. 217-234, MIT Press, Cambridge, MA, 1961.

[5] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. SPIE*, vol. 2179, pp. 127-141, SPIE Press, Bellingham, WA, 1994.

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

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

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

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

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

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

[12] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Transactions on Image Processing*, vol. 10, no. 10, pp. 1397-1410, Oct. 2001.

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

[14] Z. Wang, A. C. Bovik and L. Lu, "Why is image quality assessment so difficult?" *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Proc.*, vol. 4, pp. 3313-3316, May 2002.

[15] Z. Wang and E. P. Simoncelli, "Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics," *Human Vision and Electronic Imaging IX, Proc. SPIE*, vol. 5292, San Jose, Jan. 2004.

[16] O. D. Faugeras and W. K. Pratt, "Decorrelation methods of texture feature extraction," *IEEE Pat. Anal. Mach. Intell.* 2(4), pp. 323–332, 1980.

[17] A. Gagalowicz, "A new method for texture fields synthesis: Some applications to the study of human vision," *IEEE Pat. Anal. Mach. Intell.* 3(5), pp. 520–533, 1981.

[18] D. Heeger and J. Bergen, "Pyramid-based texture analysis/synthesis," in *Proc. ACM SIGGRAPH*, pp. 229–238, Association for Computing Machinery, August 1995.

[19] S. Zhu and D. Mumford, "Prior learning and Gibbs reaction-diffusion," *IEEE Pat. Anal. Mach. Intell.* 19(11), 1997.

[20] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int'l Journal of Computer Vision* 40, pp. 49–71, December 2000.

[21] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," 37th *IEEE Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, Nov. 2003.

[22] Z. Wang, L. Lu and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*,

special issue on "Objective video quality metrics", vol. 19, no. 2, pp. 121-132, Feb. 2004.

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