Chapter 7 in Digital Video Image Quality and Perceptual Coding
(H. R. Wu, and K. R. Rao, eds.), Marcel Dekker Series in Signal
Processing and Communications, Nov. 2005.

Structural Similarity Based Image Quality Assessment

Zhou Wang, Alan C. Bovik and Hamid R. Sheikh

It is widely believed that the statistical properties of the natural visual environment play a fundamental role in the evolution, development and adaptation of the human visual system (HVS). An important observation about natural image signals is that they are highly structured. By "structured signal", we mean that the signal samples exhibit strong dependencies amongst themselves, especially when they are spatially proximate. These dependencies carry important information about the structure of the objects in the visual scene. The principle hypothesis of *structural similarity* based image quality assessment is that the HVS is highly adapted to extract structural information from the visual field, and therefore a measurement of structural similarity (or distortion) should provide a good approximation to perceived image quality.

In this chapter, we introduce structural similarity as an alternative design philosophy for objective image quality assessment methods. This is different from and complementary to the typical HVS-based approaches, which usually calculate signal difference between the distorted and the reference images, and attempt to quantify the difference "perceptually" by incorporating known HVS properties.

1.1 Structural Similarity and Image Quality

In *full-reference* image quality assessment methods, the quality of a test image is evaluated by comparing it with a reference image that is assumed to have perfect

quality. The goal of image quality assessment research is to design methods that quantify the strength of the perceptual similarity (or difference) between the test and the reference images. Researchers have taken a number of approaches to this end.

The first approach, which we call the error sensitivity approach, assumes that the test image signal is the sum of the reference image signal and an error signal. Assuming that the loss of perceptual quality is directly related to the visibility of the error signal, most HVS-based image quality assessment models attempt to weight and combine different aspects of the error signal according to their respective visual sensitivities, which are usually determined by psychophysical measurements. One problem with this approach is that larger visible differences may not necessarily imply lower perceptual quality. An example is shown in Figure 1.1, where the original "Einstein" image is altered with different distortions: contrast stretch, mean shift, JPEG compression, blurring, and impulsive salt-pepper noise contamination. We adjusted each type of distortion to yield the same mean squared error (MSE) relative to the original image, except for the JPEG compressed image, which has a slightly smaller MSE. Despite their nearly identical MSE, the images can be seen to have significantly different perceptual qualities. It is important to note that although the difference between the contrast stretched image (Figure 1.1(b)) and the reference image (Figure 1.1(a)) is easily discerned, the contrast stretched image has good perceptual quality.

The second approach is based on the conjecture that the purpose of the entire visual observation process is to efficiently extract and make use of the information represented in natural scenes, whose statistical properties are believed to play a fundamental role in the evolution, development and adaptation of the HVS (e.g., [1]). One distinct example of the second approach is the *Structural similarity* based image quality assessment method [2], which is motivated from the observation that natural image signals are highly "structured," meaning that the signal samples have strong dependencies amongst themselves, especially when they are spatially proximate. These dependencies carry important information about the structure of the objects in the visual scene. The principle premise of the structural similarity approach is that the major goal of visual observation is to extract such information, for which the HVS is highly adapted. Therefore, a measurement of structural information change or structural similarity (or distortion) should provide a good approximation to perceived image quality. Let us again take the contrast stretched image in Figure 1.1(b) as an example. Although its visible difference from the reference image is significant, it preserves almost all of the important information that reflects the structure of the objects represented in the image. In fact, the reference image can almost be fully recovered via a simple point-wise inverse linear luminance transform. Consequently, a high quality score should be assigned. On the other hand, some structural information in the original image is

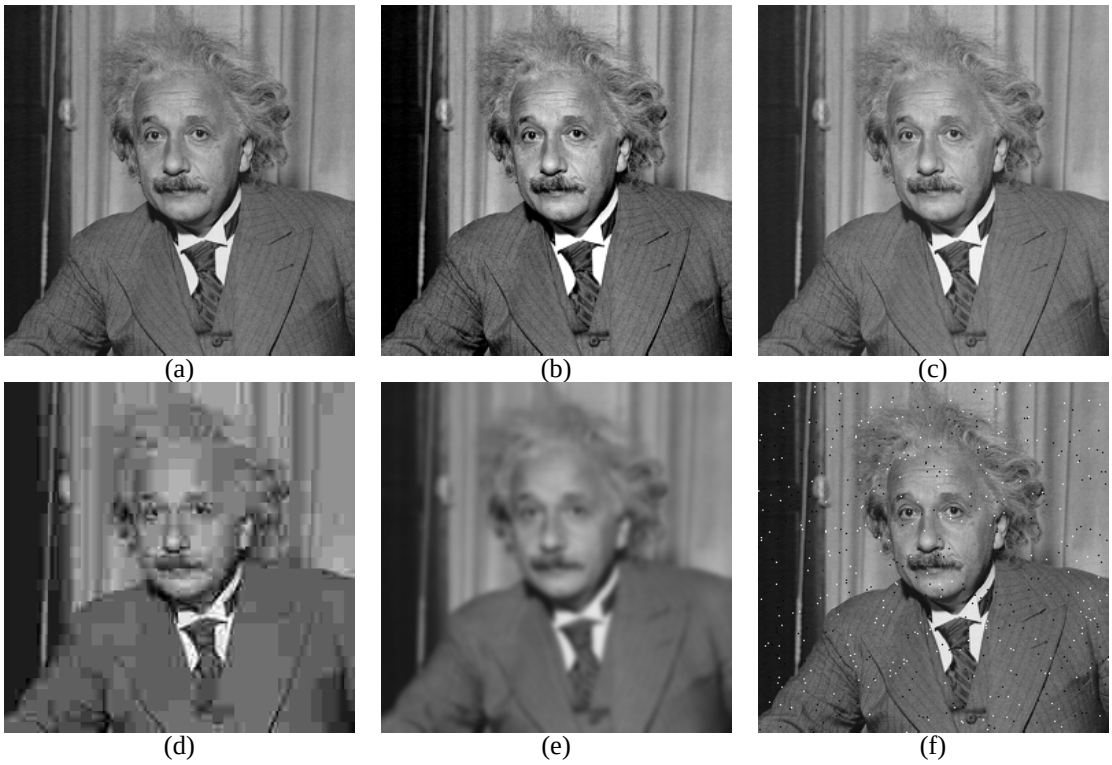

Figure 1.1: Comparison of 8bits/pixel "Einstein" images with different types of distortions. (a) original image, MSE = 0, MSSIM = 1; (b) contrast stretched image, MSE = 144, MSSIM = 0.9133; (c) mean shifted image, MSE = 144, MSSIM = 0.9884; (d) JPEG compressed image, MSE = 142, MSSIM = 0.6624; (e) blurred image, MSE = 144, MSSIM = 0.6940; (f) salt-pepper impulsive noise contaminated image, MSE = 144, MSSIM = 0.8317.

severely distorted and permanently lost in the JPEG compressed and the blurred images, and therefore they should be assigned lower quality scores.

The natural question that follows is then: What constitutes important information that reflects the structure of objects represented in an image? This is the key issue that will define the specific implementation of the image quality assessment algorithm. While it is difficult to directly provide a relatively small set of features that sufficiently describe the structural information in an image, it is worthwhile to consider its opposite: what is the information in an image that is not important for representing the structure of the objects? A simple answer comes 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 out the influence of illumination from the infor-

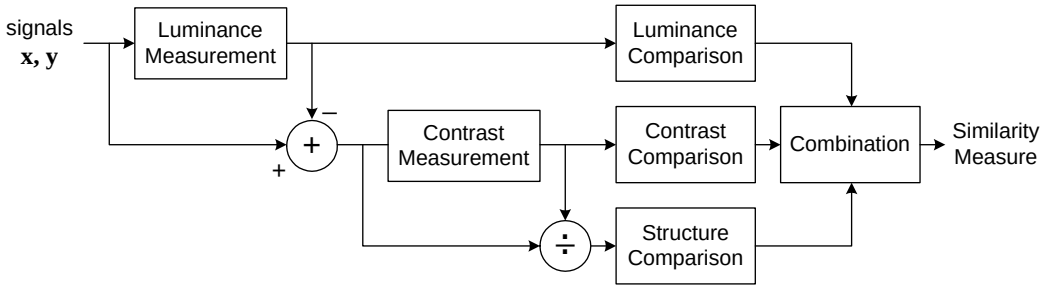

Figure 1.2: Diagram of the proposed similarity measurement system. (Adapted from [2])

mation that is more important for representing object structures. Intuitively, the major impact of illumination change is the variation of the average luminance and contrast in the image. Since luminance and contrast can vary across a scene, they are preferably measured locally. This leads to a localized image similarity measure that separates (and perhaps removes) the influence of luminance and contrast variation from the remaining attributes of the local image region.

The first instantiation of the structural similarity-based method was made in [3, 4] and promising results on simple tests were achieved. This method was further generalized and improved in [2, 5]. It was also adapted for video quality assessment in [6]. In Sections 1.2 and 1.3 of this chapter, we will mainly have a close look at the Structural SIMilarity (SSIM) index introduced in [2].

1.2 The Structural SIMilarity (SSIM) Index

The system diagram of the SSIM image quality assessment system is shown in Figure 1.2. Suppose $\mathbf{x}$ and $\mathbf{y}$ are two non-negative image signals, which have been aligned with each other (e.g., spatial patches extracted from each image). The purpose of the system is to provide a similarity measure between them. The similarity measure can serve as a quantitative measurement of the quality of one signal if we consider the other to have perfect quality. Here $\mathbf{x}$ and $\mathbf{y}$ can be either continuous signals with a finite support region, or discrete signals represented as $\mathbf{x} = \{ x_i \,|\, i = 1, 2, \cdots, N\}$ and $\mathbf{y} = \{ y_i \,|\, i = 1, 2, \cdots, N\}$, respectively, where i is the sample index and N is the number of signal samples (pixels).

The system separates the task of similarity measurement into three comparisons: luminance, contrast and structure. First, the luminance of each signal is

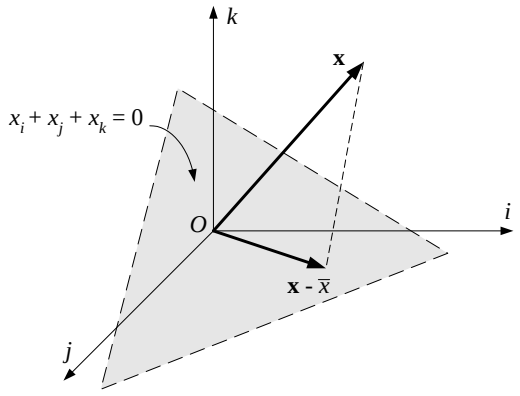

Figure 1.3: Projection onto the hyperplane of $\sum x_i = 0$. Note: this is an illustration in 3-D space. In practice, the number of dimensions is equal to the number of pixels.

compared. Assuming discrete signals, this is estimated as the mean intensity:

$$\mu_x = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \,. \tag{1.1}$$

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

$$l(\mathbf{x}, \mathbf{y}) = l(\mu_x, \mu_y) \,. \tag{1.2}$$

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

$$\sum_{i=1}^{N} x_i = 0 \,. \tag{1.3}$$

as illustrated in Figure 1.3. We use the standard deviation (the square root of variance) as an estimate of the signal contrast. An unbiased estimate in discrete form is given by

$$\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N} (x_i - \mu_x)^2 \right)^{1/2} \,. \tag{1.4}$$

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

$$c(\mathbf{x}, \mathbf{y}) = c(\sigma_x, \sigma_y) \,. \tag{1.5}$$

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:

$$s(\mathbf{x}, \mathbf{y}) = s\left(\frac{\mathbf{x} - \mu_x}{\sigma_x}, \frac{\mathbf{y} - \mu_y}{\sigma_y}\right). \tag{1.6}$$

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

$$S(\mathbf{x}, \mathbf{y}) = f(l(\mathbf{x}, \mathbf{y}), c(\mathbf{x}, \mathbf{y}), s(\mathbf{x}, \mathbf{y})). \tag{1.7}$$

An important point is that the three components are relatively independent. For example, the change of luminance and/or contrast has little impact on the structures of images.

In order to complete the definition of the similarity measure in Eq. (1.7), we need to define the three functions $l(\mathbf{x}, \mathbf{y})$, $c(\mathbf{x}, \mathbf{y})$, $s(\mathbf{x}, \mathbf{y})$, as well as the combination function $f(\cdot)$. We also would like the similarity measure to satisfy the following conditions:

1. Symmetry: $S(\mathbf{x}, \mathbf{y}) = S(\mathbf{y}, \mathbf{x})$. Since our purpose is to quantify the similarity between two signals, exchanging the order of the input signals should not affect the resulting similarity measurement.

2. Boundedness: $S(\mathbf{x}, \mathbf{y}) \leq 1$. Boundedness is a useful property for a similarity metric since an upper bound can serve as an indication of how close the two signals are to being perfectly identical. This is in contrast with most signal-to-noise ratio type of measurements, which are typically unbounded.

3. Unique maximum: $S(\mathbf{x}, \mathbf{y}) = 1$ if and only if $\mathbf{x} = \mathbf{y}$ (in discrete representations, $x_i = y_i$ for all $i = 1, 2, \cdots, N$). In other words, the similarity measure should quantify any variations that may exist between the input signals. The perfect score is achieved *only* when the signals being compared are exactly the same.

For luminance comparison, we define

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\,\mu_x\,\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}. \tag{1.8}$$

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

$$C_1 = (K_1\,L)^2, \tag{1.9}$$

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. Eq. (1.8) is easily seen to obey the three properties listed above.

Equation (1.8) is also connected with Weber's law, which has been widely used to model light adaptation (also called luminance masking) in the HVS. According to Weber's law, the magnitude of a just-noticeable luminance change ΔI is approximately proportional to the background luminance I for a wide range of luminance values. In other words, the HVS is sensitive to the *relative* luminance change, and not the absolute luminance change. Letting R represent the ratio of luminance change relative to background luminance, we rewrite the luminance of the distorted signal as $\mu_y = (1+R)\mu_x$. Substituting this into Eq. (1.8) gives

$$l(\mathbf{x}, \mathbf{y}) = \frac{2(1+R)}{1+(1+R)^2 + C_1/\mu_x^2} \ . \tag{1.10}$$

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 much more than the visibility threshold, which is out of the application scope of Weber's law.

The contrast comparison function takes a similar form:

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\,\sigma_x\,\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \,, \tag{1.11}$$

where C_2 is a non-negative constant

$$C_2 = (K_2\,L)^2 \,, \tag{1.12}$$

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

Structure comparison is conducted after luminance subtraction and contrast normalization. Specifically, we associate the direction of the two unit vectors $(\mathbf{x} - \mu_x)/\sigma_x$ and $(\mathbf{y} - \mu_y)/\sigma_y$, each lying in the hyperplane (Figure 1.3) defined by Eq. (1.3), with the structures of the two images. The correlation (inner product) between them is a simple and effective measure to quantify the structural similarity. Notice that the correlation between $(\mathbf{x} - \mu_x)/\sigma_x$ and $(\mathbf{y} - \mu_y)/\sigma_y$ is equivalent to the correlation coefficient between $\mathbf{x}$ and $\mathbf{y}$. Thus, we define the structure

comparison function as follows:

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x \, \sigma_y + C_3}. \tag{1.13}$$

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

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y) \, . \tag{1.14}$$

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

Finally, we combine the three comparisons of Eqs. (1.8), (1.11) and (1.13) and name the resulting similarity measure the Structural SIMilarity (SSIM) index between signals $\mathbf{x}$ and $\mathbf{y}$:

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = [l(\mathbf{x}, \mathbf{y})]^{\alpha} \cdot [c(\mathbf{x}, \mathbf{y})]^{\beta} \cdot [s(\mathbf{x}, \mathbf{y})]^{\gamma} \, , \tag{1.15}$$

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

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2 \, \mu_x \, \mu_y + C_1)(2 \, \sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \, . \tag{1.16}$$

The SSIM index may be better understood geometrically in a vector space of signal components as in Figure 1.4. These signal components can be either image pixel intensities or other extracted features such as transformed linear coefficients. Figure 1.4 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 two-dimensional 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 1.4(a) shows the result for a simple Minkowski metric. Each contour has the same size and shape (a circle here, as we are assuming an exponent of 2). That is, perceptual distance corresponds to Euclidean distance. Figure 1.4(b) shows a Minkowski metric in which different signal components are weighted differently. This could be, for example, weighting according to the contrast sensitivity function (CSF), as is common in many quality assessment models. Here the contours are ellipses,

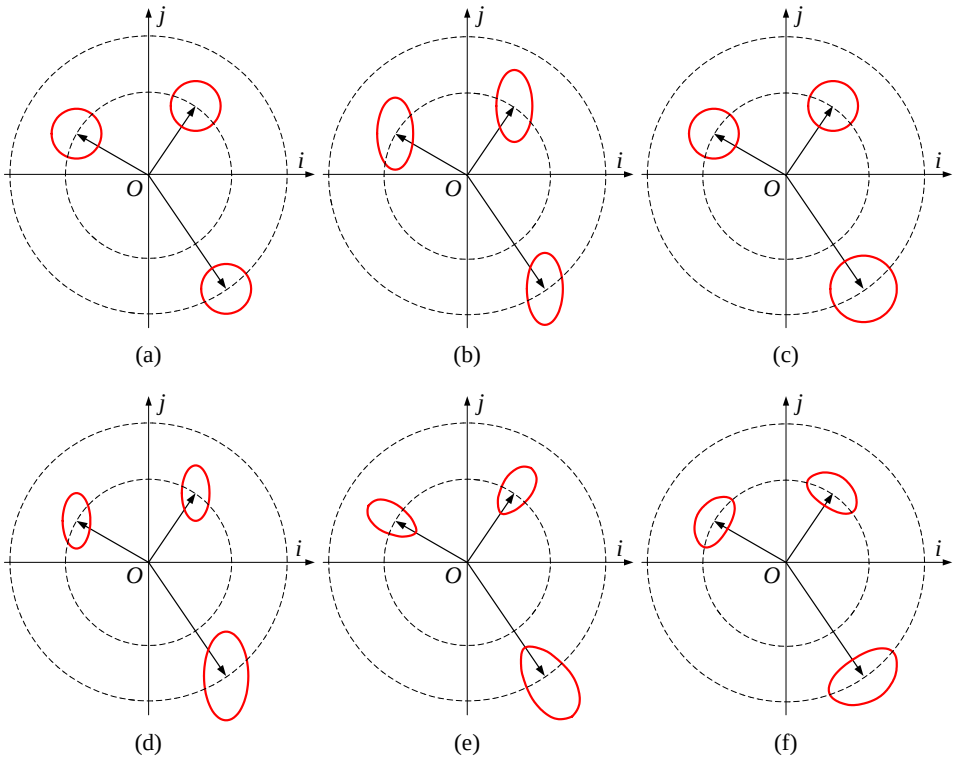

Figure 1.4: Equal-distortion contours for different quality measurement systems. (a) Minkowski error measurement systems; (b) component-weighted Minkowski error measurement systems; (c) magnitude-weighted Minkowski error measurement systems; (d) magnitude and component-weighted Minkowski error measurement systems; (e) SSIM measurement system (a combination of Eqs. (1.11) and (1.13)) with more emphasis on $s(\mathbf{x}, \mathbf{y})$; (f) SSIM measurement system (a combination of Eqs. (1.11) and (1.13)) with more emphasis on $c(\mathbf{x}, \mathbf{y})$. Each image is represented as a vector, whose entries are image components. Note: this is an illustration in 2-D space. In practice, the number of dimensions should be equal to the number of image components used for comparison (e.g, the number of pixels or transform coefficients). (From [2])

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 Figure 1.4(c). 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 1.4(d) shows a combination of contrast masking (magnitude weighting) followed by component weighting. In comparison of the vectors $\mathbf{x} - \mu_x$ and $\mathbf{y} - \mu_y$, the SSIM index corresponds to the comparison of two independent quantities: the vector lengths, and their angles. Thus, the contours will be aligned with the axes of a polar coordinate system. Figures 1.4(e) and 1.4(f) show two examples of this, computed with different exponents. 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.

1.3 Image Quality Assessment Based on the SSIM Index

For image quality assessment, it is useful to apply the SSIM index locally rather than globally. First, image statistical features are usually highly spatially non-stationary. Second, image distortions, which may or may not depend on the local image statistics, may also be space-variant. Third, at typical viewing distances, only a local area in the image can be perceived with high resolution by the human observer at one time instance (because of the foveation feature of the HVS, e.g., [7]). Fourth, localized quality measurement can provide a spatially varying quality map of the image, which delivers more information about the quality degradation of the image and may be useful in some applications.

In [3, 4], the local statistics μ_x, σ_x and σ_{xy} (Eqs. (1.1),(1.4) and (1.14)) 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 Figure 1.5(a). Such kind of "artifacts" are not desirable because it is created from the choice of the quality measurement system (local square window), but not from image distortions. In [2], a circular-symmetric Gaussian weighting function $\mathbf{w} = \{ w_i \,|\, i = 1, 2, \cdots, N\}$ with unit sum ($\sum_{i=1}^{N} w_i = 1$) is adopted. The estimates of local statistics μ_x, σ_x

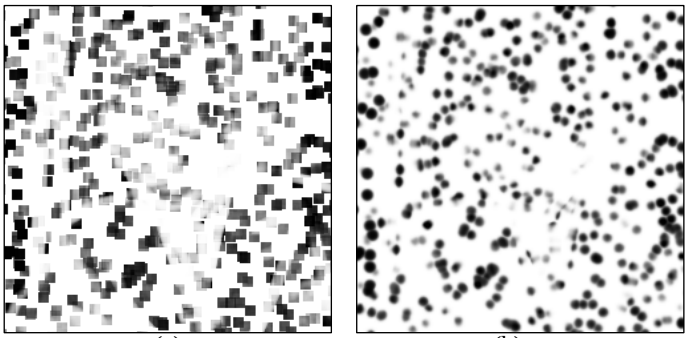

Figure 1.5: SSIM index maps of the impulse noise contaminated "Einstein" image (Figure 1.1(f)). Brightness indicates the magnitude of the local SSIM index value. (a) square windowing approach; (b) smoothed windowing approach.

and σ_{xy} are then modified accordingly as

$$\mu_x = \sum_{i=1}^{N} w_i \, x_i \, . \tag{1.17}$$

$$\sigma_x = \left(\sum_{i=1}^{N} w_i \, (x_i - \mu_x)^2 \right)^{1/2} \, . \tag{1.18}$$

$$\sigma_{xy} = \sum_{i=1}^{N} w_i \, (x_i - \mu_x)(y_i - \mu_y) \, . \tag{1.19}$$

With such a windowing approach, the quality maps exhibit a locally isotropic property, as shown in Figure 1.5(b).

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

$$\text{MSSIM} = \sum_{j=1}^{M} W_j \cdot \text{SSIM}_j \, , \tag{1.20}$$

where M is the number of samples in the quality map, SSIM_j is the SSIM index value at the j-th sample, and W_j is the weight given to the j-th sample and

$$\sum_{j=1}^{M} W_j = 1 \, . \tag{1.21}$$

If all the samples in the quality map are equally weighted, then $W_j = 1/M$ for all

j's. Depending on the application, it is also possible to assign spatially varying weights to different samples in the SSIM index map. For example, region-of-interest image processing systems may give different weights to different segmented regions in the image. For another example, it has been observed that different image textures attract human fixations with varying degrees (e.g., [8, 9]), and therefore a fixation probability model can be used to define the weighting model. Further, since the visual resolution decreases gradually as a function of the distance from the fixation point (e.g., [10]), a smoothly varying foveated weighting model can also be employed to define the weights. For the experiments described in this chapter, however, we use uniform weighting. A MatLab implementation of the SSIM index algorithm is available online at [11].

Many image quality assessment algorithms have been shown to behave consistently when applied to distorted images created from the same original image, using the same type of distortions (e.g., JPEG compression). However, the effectiveness of these models degrades significantly when applied to a set of images originating from different reference images, and/or including a variety of different types of distortions. Thus, cross-image and cross-distortion tests are critical in evaluating the effectiveness of an image quality metric. It is impossible to show a thorough set of such examples, but the images in Figure 1.1 provide an encouraging starting point for testing the cross-distortion capability of the quality assessment algorithms. The MSE and MSSIM measurement results are given in the figure caption. Obviously, the MSE performs very poorly in this case. The MSSIM values exhibit much better consistency with the qualitative visual appearance.

For a more thorough test, we apply the SSIM index algorithm to the LIVE database of JPEG and JPEG2000 compressed images that were evaluated by a number of subjects for perceptual quality [12]. The database was created with 29 high-resolution 24 bits/pixel RGB color images (typically 768×512 or similar size) compressed at a range of quality levels using either JPEG or JPEG2000, producing a total of 175 JPEG images and 169 JPEG2000 images. The bit rates were in the range of 0.150 to 3.336 and 0.028 to 3.150 bits/pixel for JPEG and JPEG2000 images, respectively, and were chosen non-uniformly such that the resulting distribution of subjective quality scores was approximately uniform over the entire range. Subjects viewed the images from comfortable viewing distances and were asked to provide their perception of quality on a continuous linear scale that was divided into five equal regions marked with adjectives "Bad", "Poor", "Fair", "Good" and "Excellent". Each JPEG and JPEG2000 compressed image was viewed by $13 \sim 20$ subjects and 25 subjects, respectively. The subjects were mostly male college students. Raw scores for each subject were normalized by the mean and variance of scores for that subject (i.e., raw values were converted to Z-scores [13]) and then scaled and shifted by the mean and variance of the entire subject pool to fill the range from 1 to 100. Mean opinion scores (MOSs) were then

computed for each image, after removing outliers (most subjects had no outliers). The image database, together with the subjective score and standard deviation for each image, has been made available on the Internet at [12].

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

Figure 1.6 shows two sample JPEG and JPEG2000 images from the database, together with their SSIM index maps and absolute error maps. By closer inspection of corresponding spatial locations in the SSIM index and the absolute error maps, we observe that the SSIM index is generally more consistent with perceived quality measurement. In particular, note that at low bit rates, the coarse quantization in JPEG and JPEG2000 algorithms often results in smooth representations of fine-detail regions in the image (e.g., the tiles in Figure 1.6(c) and the trees in Figure 1.6(d)). Compared with other types of regions, these regions may not be worse in terms of pointwise difference measures such as the absolute error. However, since the structural information of the image details are nearly completely lost, they exhibit poorer visual quality. Comparing Figure 1.6(e) with Figure 1.6(g), and Figure 1.6(f) with 1.6(h)), we can see that the SSIM index is better in capturing such poor quality regions.

The scatter plots of MOS versus PSNR and MSSIM image quality prediction are shown in Figure 1.7, where each sample point represents one test image. It can be observed that MSSIM supplies better prediction capability of the subjective scores than PSNR. In order to provide quantitative comparisons on the performance of the SSIM index measure, we use the logistic function adopted in the video quality experts group (VQEG) Phase I FR-TV test to provide a non-linear mapping between the objective/subjective scores [14]. The fitted curves are shown in Figure 1.7. After fitting, a set of quantitative measures are computed, which include the Pearson correlation coefficient (CC), the mean absolute error (MAE), the root mean squared error (RMS), the outlier ratio (OR, defined as the proportion of predictions that are outside the range of two times of the standard error in the subjective test), and the Spearman rank-order correlation coefficient (SROCC). Readers can refer to [14, 15] for details about how these measures are calculated. It can be seen that MSSIM outperforms PSNR in all these comparisons by clear margins.

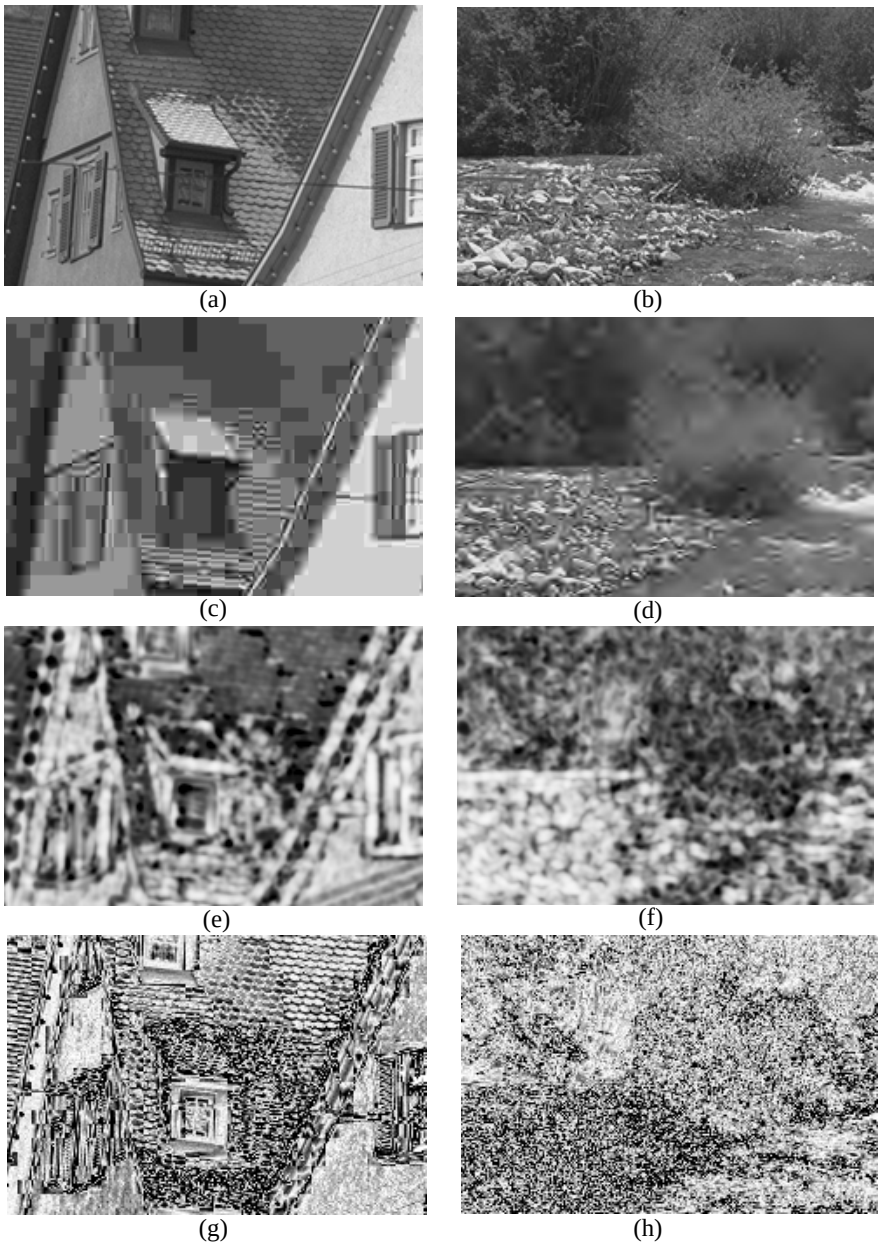

Figure 1.6: Sample JPEG and JPEG2000 compressed images and quality maps (cropped from 768×512 to 240×160 for visibility). (a) and (b) are the original "Buildings" and "Stream" images, respectively. (c) JPEG compressed "Buildings" image, 0.2673 bits/pixel; (d) JPEG2000 compressed "Stream" image, 0.1896 bits/pixel; (e) and (f) show SSIM maps of the compressed images, where brightness indicates the magnitude of the local SSIM index (squared for visibility). (g) and (h) show absolute error maps of the compressed images, where brighter point indicates smaller error (for easier comparison with the SSIM map).

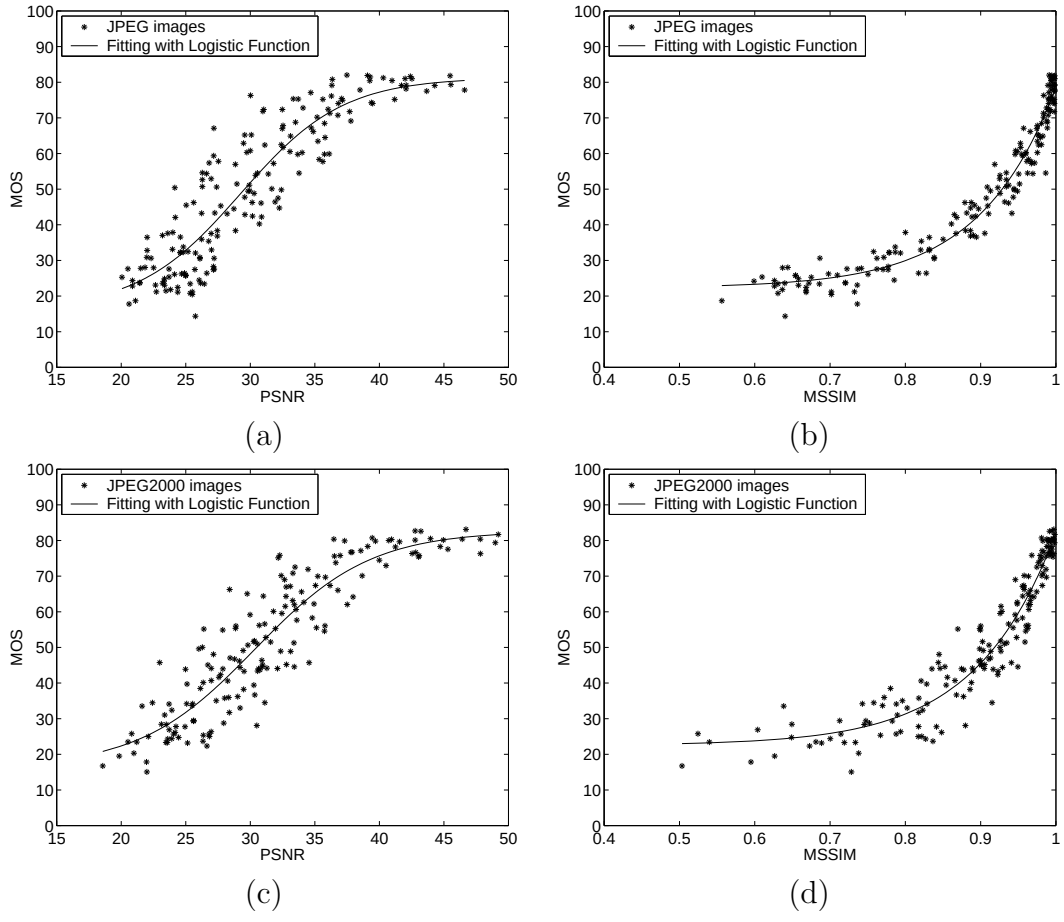

Figure 1.7: Scatter plots of subjective mean opinion score (MOS) versus model prediction. Each sample point represents one test image. (a) PSNR prediction for JPEG images; (b) MSSIM prediction for JPEG images; (c) PSNR prediction for JPEG2000 images; (d) MSSIM prediction for JPEG2000 images.

Table 1.1: Performance comparison of PSNR and MSSIM using the JPEG and JPEG2000 image databases. CC: correlation coefficient; MAE: mean absolute error; RMS: root mean squared error; OR: outlier ratio; SROCC: Spearman rank-order correlation coefficient.

Database	JPEG		JPEG2000	
Model	PSNR	MSSIM	PSNR	MSSIM
CC	0.904	0.978	0.910	0.958
MAE	6.769	3.324	6.300	4.352
RMS	8.637	4.176	8.062	5.540
OR	0.200	0.006	0.095	0.024
SROCC	0.893	0.973	0.906	0.955

1.4 Discussions

This chapter discusses the motivation, the general idea, and a specific SSIM index algorithm of the structural similarity-based image quality assessment method. It is worthwhile to look into the relationship between this method and the traditional error sensitivity based image quality assessment algorithms.

On the one hand, we consider "structural similarity" as a substantially different design principle for image quality assessment and would like to emphasize two distinct features of this method in comparison with the error sensitivity-based models. First, in terms of the nature of the distortions that a quality measure attempts to capture, it is targeted at *perceived structural information variation*, instead of *perceived error*. Second, in terms of the construction of the quality assessment system, it is a *top-down* approach that mimics the hypothesized functionality of the overall HVS, as opposed to a *bottom-up* approach that simulates the function of relevant early-stage components in the HVS.

On the other hand, we also view the structural similarity-based methods as being complementary to, rather than opposed to, the typical error sensitivity based approaches. Notice that error sensitivity based methods often involve signal decompositions based on linear transforms such as the wavelet transforms (e.g., [16, 17, 18, 19]). Such signal decompositions can be thought of as specific descriptive representations of the signal "structures". In this sense, the error between transformed wavelet coefficients implicitly suggests the structural change between the image signals being compared. On the other hand, the SSIM indexing method as described in the previous sections might be converted into an equivalent "error" measure in a specific coordinate system, only that such a coordinate system is locally adaptive, non-linear, and input-dependent. It needs to be mentioned that

certain divisive-normalization based masking models (e.g., [20, 21]) exhibit input-dependent behavior in measuring signal distortions, which leads to a departure from the distortion contours shown in Figures 1.4(a)-(d), although precise alignment with the axes of a polar coordinate system as in Figures 1.4(e) and 1.4(f) is not observed. Although not clear at this moment, we think it is possible that the two types of approaches may eventually converge into similar solutions.

The SSIM indexing algorithm is quite encouraging not only because it supplies good quality prediction accuracy in the current tests, but also because of its simple-formulation and low complexity implementation. This is in contrast with many complicated HVS-based quality assessment systems. Its simplicity makes it much more tractable in the context of algorithm and parameter optimizations for the development of perceptually-optimized image processing and coding systems.

Finally, we would like to point out that the SSIM indexing approach is only a particular implementation of the philosophy of structural similarity, from an image formation point of view. Under the same philosophy, other approaches may emerge that could lead to algorithms significantly different from the SSIM index. Creative investigation of the concepts of structural information and structural distortion is likely to drive the success of these innovations.

Bibliography

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

[2] 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.

[3] Z. Wang, *Rate scalable foveated image and video communications*, Ph.D. dissertation, Dept. of ECE, The University of Texas at Austin, Dec. 2001.

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

[5] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *Proc. IEEE Asilomar Conf. Signals, Systems & Computers*, Nov. 2003.

[6] Z. Wang, L. Lu and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, special issue on video quality metrics, vol. 19, no. 1, Jan. 2004.

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

[8] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 9, pp. 970-982, Sep. 2000.

[9] U. Rajashekar, L. K. Cormack and A. C. Bovik, "Image features that draw fixations," *IEEE Inter. Conf. Image Processing*, vol. 3, pp. 313-316, Sep. 2003.

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

[11] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "The SSIM Index for Image Quality Assessment," *http://www.cns.nyu.edu/~lcv/ssim/*, 2003.

[12] H. R. Sheikh, Z. Wang, A. C. Bovik and L. K. Cormack,"Image and video quality assessment research at LIVE, *http://live.ece.utexas.edu/research/quality/*, 2003.

[13] A. M. van Dijk, J. B. Martens and A. B. Watson, "Quality assessment of coded images using numerical category scaling," *Advanced Image and Video Communications and Storage Technologies*, Proc. SPIE, vol. 2451, 1995.

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

[15] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment, Phase II," *ftp://ftp.its.bldrdoc.gov/dist/ituvidq/frtv2_final_report/*, Aug. 2003.

[16] J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital images and human vision*, A. B. Watson, ed., pp. 163-178, The MIT Press, 1993.

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

[18] P. C. Teo and D. J. Heeger, "Perceptual Image Distortion," *Human Vision, Visual Processing, and Digital Display V*, Proc. SPIE, vol. 2179, pp. 127-141, 1994.

[19] 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.

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

[21] I. Epifanio, J. Gutirrez and J. Malo, "Linear transform for simultaneous diagonalization of covariance and perceptual metric matrix in image coding," *Pattern Recognition*, vol. 36, no. 8, pp.1679-1923, Aug. 2003.