in Multi-Sensor Image Fusion and Its Applications (R. S. Blum and Z. Liu, eds.),
CRC Press, July 2005.

Structural and Information Theoretic Approaches to Image Quality Assessment

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

1 Introduction

Digital images and video are prolific in the world today owing to the ease of acquiring, processing, storing and transmitting them. Several common image processing operations such as compression, dithering and printing affect the quality of the image. Advances in sensor and networking technologies, from the internet to wireless networks, has led to a surge of interest in image communication systems. Again, the communication channel tends to distort the image signal passing through it and the quality of the image needs to be closely monitored to meet requirements at the end receiver. The integrity of the image also needs to be preserved, irrespective of the specific display system that is used by the viewer. In this chapter, we describe state of the art objective quality metrics to assess the quality of digital images.

In all the applications mentioned above, the targeted receiver is the human eye. Compression and half-toning algorithms are generally designed to generate images that closely approximate the original reference image, *as seen by the human eye.* In fact, most of the images encountered in day to day life are meant for human consumption. The measure of quality depends on the intended receiver and we focus on applications where the ultimate receiver is the Human Visual System (HVS).

Subjective assessment of image quality involves studies where humans are asked to assign a score to an image after viewing it under certain fixed environment conditions such as viewing distance, display characteristics and so on. Typically, the same image is shown to a number of subjects and final scores are assigned after accounting for any variability in the assessment. However, subjective assessment studies are tedious and impossible to do for every possible image in the world. The goal of quality assessment research is to objectively predict the quality of an image to approximate the score a human might assign to it.

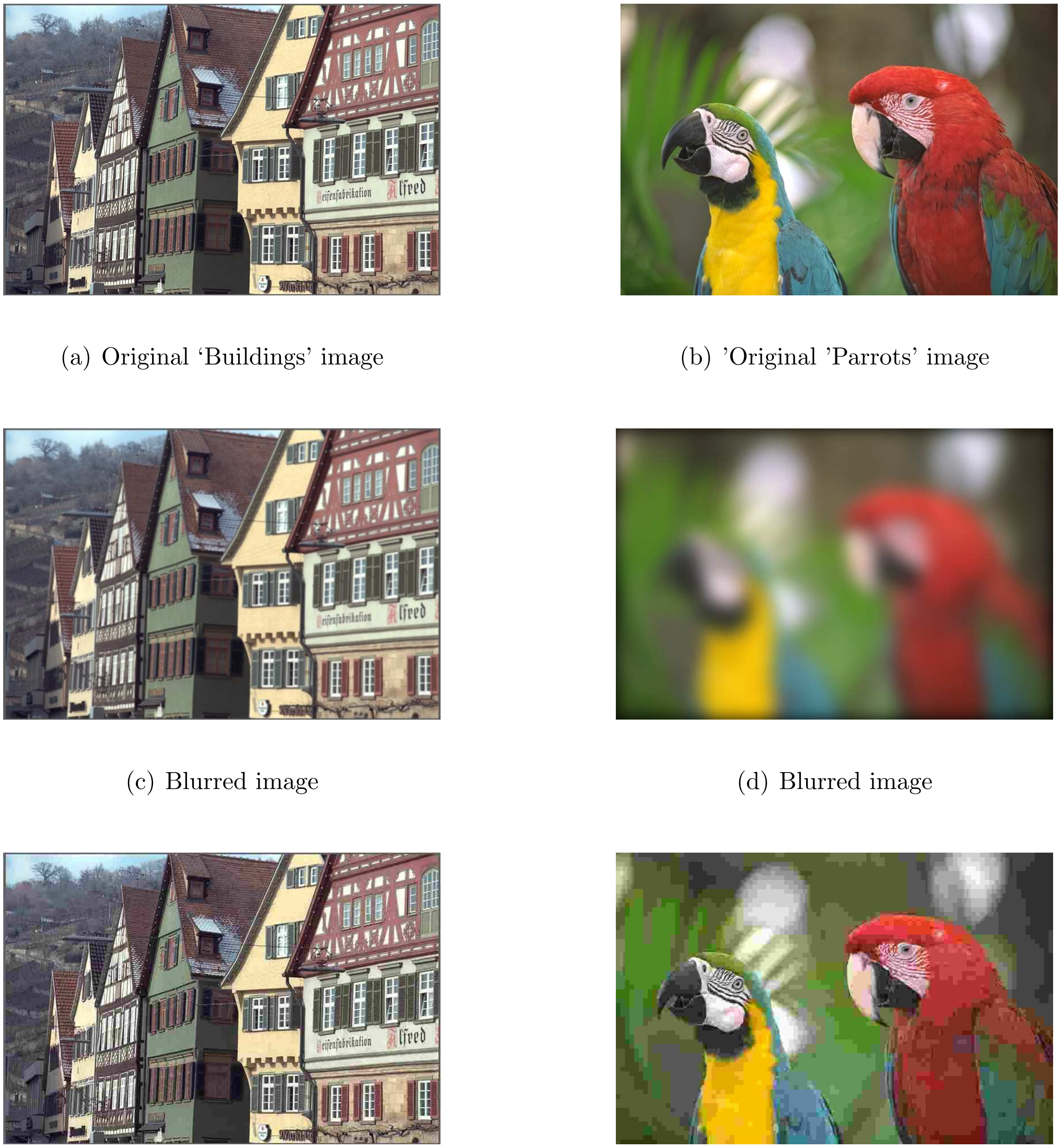
The HVS is very good at evaluating the quality of an image blindly, i.e., without a reference "perfect" image to compare it against. It is however rather difficult to perform this task automatically using a computer. Although the term *image quality* is used, what we are actually referring to is *image fidelity*. We assume that a reference image is available and the quality of the test image is determined by how close it is to the reference image perceptually.

Objective measures for image quality play an important role in evaluating the performance of image fusion algorithms [1]. In many applications, the end receiver of the fused images are humans. Several fusion algorithms presented in the literature have been evaluated using subjective criteria [2], as well as measures such as Mean Square Error (MSE) described below [3]. Measures to evaluate objective quality are hence useful to evaluate the effectiveness of a fusion algorithm, as well as to optimize for various parameters of the algorithm to improve performance.

MSE between the reference and test images is a commonly used metric for image quality. Let $\vec{x} = \{x_i, i = 1, 2, \ldots, N\}$ denote a vector containing the reference image pixel values, where i denotes a spatial index. Similarly, let $\vec{y} = \{y_i, i = 1, 2, \ldots, N\}$ denote the test image. Then the MSE between $\vec{x}$ and $\vec{y}$ is defined as

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

This metric is known to correlate very poorly with visual quality, but is still widely used due to its simplicity. The failure of MSE as a metric for quality is illustrated here using a number of examples where all distortions have been adjusted in strength such that the MSE between the reference and distorted image is 50. The visual quality of the images are, however, drastically different. The MSE is a function of just the difference between corresponding pixel values and implicitly assumes that the perceived distortion is independent of both the actual value of the pixel in the reference image and neighboring values. This directly contradicts the luminance masking and contrast masking properties of the HVS [4]. In reality, the perception of distortion varies with the actual image at hand, as illustrated in Fig. 1. Fig. 1(a) shows the original "Buildings" image. Fig. 1(c) shows a blurred version of the same image obtained by convolving the image with a Gaussian window. Fig. 1(e) shows a JPEG compressed version of the image. Fig. 1(b), 1(d) and 1(f) show the original, Gaussian blurred and JPEG compressed "Parrots" image. The "Parrots" image has relatively large smooth areas, corresponding to low frequencies, and the blurring distortion is very pronounced. The blocking artifacts of JPEG compression are also visibly annoying in this image. This is, however, not the case in the "Buildings" image. Additionally, just adding a constant to every pixel in the image leads to a large MSE, but almost insignificant loss in visual quality. This is illustrated in Fig. 2(a). Salt and Pepper noise is a form of extremely impulsive noise where randomly selected pixels in the image are turned black or white. This kind of noise is usually visible to the eye even at very low strengths and the image has extremely poor quality at a MSE of 50, as illustrated in Fig. 2(b). The reference images shown here are available from the database in [5]. These examples clearly illustrate the failure of MSE as a good metric for visual quality.



(a) Original ‘Buildings’ image

(b) ’Original ’Parrots’ image

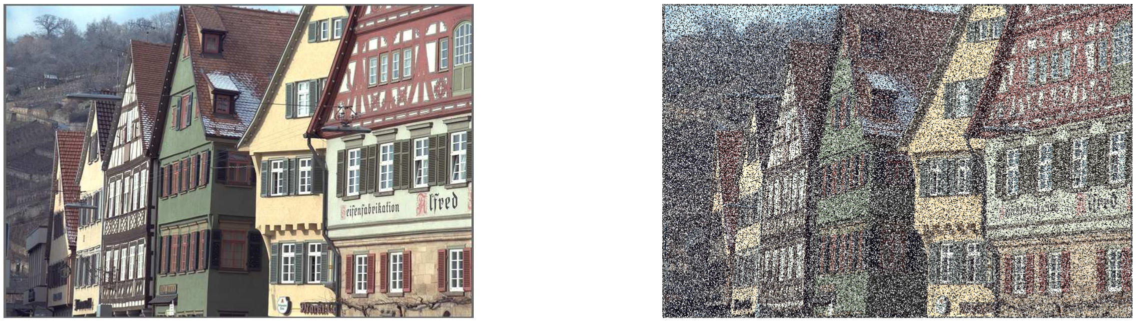
(c) Blurred image

(d) Blurred image

(e) JPEG compressed image

(f) JPEG compressed image

Figure 1: Examples of Distorted Images (MSE = 50)



(a) Mean shifted image

(b) Salt and Pepper Noise

Figure 2: Examples of Additive Distortions (MSE = 50)

Traditional approaches to image quality assessment use a bottom-up approach, where models of the Human Visual System (HVS) are used to derive quality metrics. Bottom-up HVS based approaches are those that combine models for different properties of the HVS in defining a quality metric. The response of the HVS to variations in luminance is a nonlinear function of the local mean luminance and this is commonly referred to as luminance masking. It is called masking because the variations in the distorted signal are masked by the base luminance of the reference image. Secondly, the HVS has a bandpass characteristic and its frequency response can be characterized by the Contrast Sensitivity Function (CSF). Experiments are conducted to detect the threshold of visibility of sine waves of different frequencies to determine the CSF. Contrast masking refers to the masking of certain frequency and orientation information due to the presence of other components that have similar frequency and orientation. Bottom-up HVS based quality metrics use different models to account for the luminance masking, contrast sensitivity and contrast masking features of the HVS and normalize the error terms by corresponding thresholds. The final step involves combining these normalized error terms to obtain either a quality map for the image at every pixel or a single number representing the overall visual quality of an image. A commonly used metric for error pooling is the Minkowski error. A detailed description of HVS based quality measures can be found in [6].

Recent approaches to quality assessment, however, follow a top-down approach where the hypothesized functionality of the HVS is modeled. In this chapter, we describe two such approaches that have been shown to be competitive with bottom-up HVS based approaches in predicting image quality. These methods additionally demonstrate several advantages over bottom-up HVS based measures in several aspects [7, 8].

Structural similarity approaches to image quality assume that the HVS has evolved to extract structural information from an image [8]. The quality of the image as perceived by the human eye is hence related to the loss of structural information in the image. The error metrics used here correspond to measures to quantify structural distortions, which are more meaningful than simple signal similarity criteria like MSE. A detailed description and the intuition behind this approach is presented in Section 2.

A related recent top-down approach hypothesizes that the test image is the output of a communication channel through which the reference image passes and image quality is related to the *mutual information* between these images [7]. Statistical models that accurately characterize the source and the channel are the key to the success of this approach in relating statistical information measures to perceived distortion. Natural images are images obtained from the real world and form a small subspace of the space of all possible signals [9]. A computer that generates images randomly is unlikely to produce anything that even contains objects resembling those in natural images. The statistical properties of the class of natural images have been studied by various researchers and these natural scene models are used in the information theoretic development here. The details of this approach and the derivation of the quality metric are presented in Section 3.

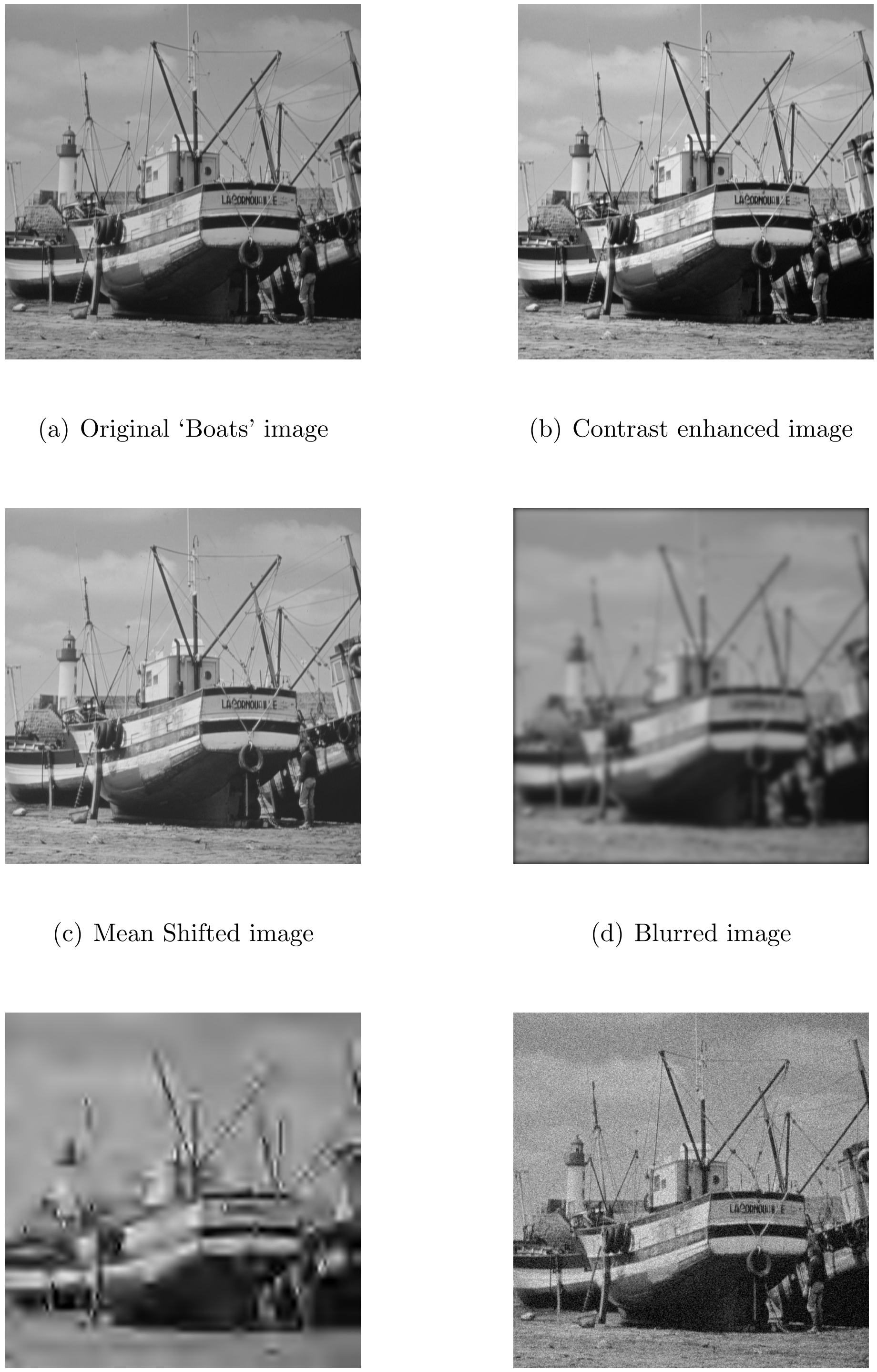
We briefly introduce the notation used throughout this chapter. $\vec{x}$ represents a vector x and bold face character $\mathbf{X}$, represent a matrix X. Capital letters are used to denote random variables and the $\hat{X}$ symbol is used to represent the estimated value of the random variable X. Greek characters are used to denote constants.

Section 4 presents experimental results that demonstrate the success of structural and information-theoretic approaches in image quality assessment. Finally, we conclude this chapter in Section 5 with a brief summary of the two paradigms of quality assessment presented here.

2 The Structural Similarity Paradigm

Traditional bottom-up HVS based measures of image quality have several limitations [8]. The working of the HVS is not yet fully understood and it is not clear how accurate the models for the HVS that are used in quality assessment are. Models of the frequency response of the HVS, for example, are typically obtained by showing human subjects relatively simple patterns like global sinusoidal gratings. Masking behavior of the HVS is modeled using data obtained by showing human subjects superposition of two or three of these sinusoidal patterns. Images of the real world, however, are quite complex and contain several structures that are the superposition of hundreds of simple sinusoidal patterns. It is difficult to justify the generalization of the models obtained from these simple experiments to characterize the HVS. Also, typical experiments performed to understand the properties of the HVS operate at the threshold of visual perception. Quality assessment, however, deals with images that are perceptibly distorted which is known as suprathreshold image distortion. It is again not clear how models developed for near visibility generalize to models that quantify perceived distortion. Finally, an implementation involving accurate models for the HVS might be too complex for most practical applications. A top down approach could lead to a simplified algorithm that works acceptably, as long as the underlying hypothesis characterizes the *primary features* of the distortion that the HVS perceives as loss of quality.

To overcome these limitations, a new framework for image quality assessment has been proposed that assumes that the HVS has evolved to extract *structural information* from images. Hence, a measure of the structural information change can be used to quantify perceived distortions [8]. This is illustrated in Fig. 3. All images shown here have approximately the same MSE with respect to the reference image. Clearly, the mean shifted and contrast enhanced image have very high perceptual quality despite the large MSE. However, this is not the case in the blurred, JPEG2000 compressed and Additive White Gaussian Noise (AWGN) corrupted images. This can be attributed to the fact that there is no loss of structural information in the former case, but this is not true in the latter case. The mean shifting and contrast stretching operations are invertible (except at the points where the luminance saturates) and the original image can be fully recovered. However, the blurring and compression are not easily invertible transformations. Blurring can be inverted, in some cases, by de-convolution when none of the frequency components are zeroed out. The HVS, however, is unable to invert the transformation easily and extract the structural information in the image. In this sense, there is no loss of structural information in the mean-shifted and contrast stretched images. Furthermore, the mean shifted and contrast stretched images have only luminance and contrast changes, as opposed to the blurred images that have severe structural distortions. The luminance and contrast of an image depend on the illumination, which does not affect the structural information in the image. The good visual quality of the mean shifted and contrast stretched images, despite the large MSE, can be attributed in the structural framework to the fact that there is almost no loss of structural information in these images.



(a) Original 'Boats' image

(b) Contrast enhanced image

(c) Mean Shifted image

(d) Blurred image

(e) JPEG 2000 compressed image

(f) AWGN Corrupted image

Figure 3: Illustrative Examples of Structural Distortions

The mathematical formulation of the Structural SIMilarity (SSIM) index is given in Section 2.1. Use of this index to predict image quality and illustrative examples of the performance of this algorithm are presented in Section 2.2.

2.1 The Structural Similarity Index

Fig. 4 shows a block diagram of the SSIM quality assessment algorithm. The luminance of an object is the product of the reflectance and illumination of the object and is independent of the structure of the object. The structural information in an image is defined as those attributes that are independent of the illumination. Hence, to quantify the loss of structural information in an image, the effects of luminance and contrast are first canceled out. The structure comparison is then carried out between the luminance and contrast normalized signals. The final quality score is a function of the luminance, contrast and structure comparisons, as shown in Fig. 4.

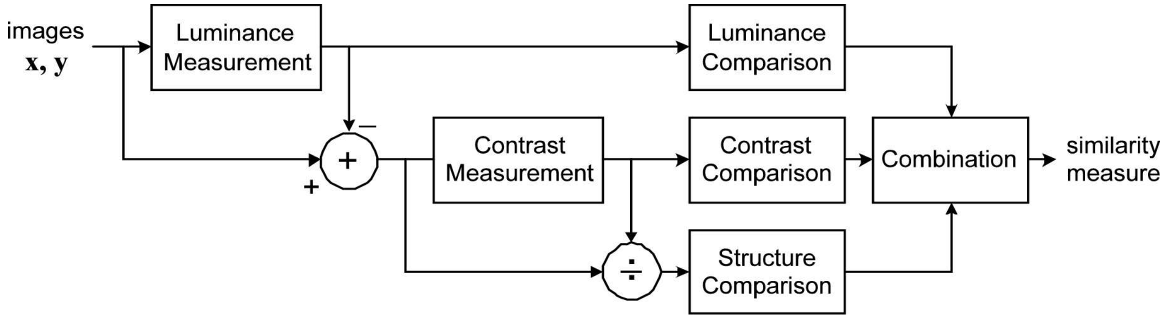


Figure 4: Block Diagram of the SSIM Quality Assessment System [Reproduced from [10]]

Let $\vec{x}$ and $\vec{y}$ represent vectors containing pixels from the reference and distorted images respectively. The reference image is assumed to have "perfect" quality. The SSIM algorithm operates in three stages given by luminance, contrast and structure comparison.

First, the luminance of the two signals is compared. The luminance of an image is estimated using its mean intensity and is given by

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2}$$

where N represents the number of pixels in $\vec{x}$. The luminance comparison function $l(\vec{x},\vec{y})$ is then a function of the luminance of the reference and test images, μ_x and μ_y respectively. Then, the luminance of the images are normalized by subtracting out the mean luminance. The resulting signal, given by $\vec{x} - \mu_x$ can then be thought of as the projection of the image $\vec{x}$ onto an N-dimensional hyperplane defined by

$$\sum_{i=1}^{N} x_i = 0 \tag{3}$$

The second stage is to compare and normalize the contrasts of the two images. The contrast is defined as the estimate of the standard deviation of the image intensities and is given by

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

The factor of $N - 1$ is used in the denominator to obtain an unbiased estimate of the standard deviation. The contrast comparison function $c(\vec{x},\vec{y})$ is then a function of the contrasts of the reference and test signals, σ_x and σ_y respectively. The contrasts of $\vec{x}$ and $\vec{y}$ are then normalized by dividing them by their own standard deviations. The structure comparison is then performed on these normalized signals given by $\frac{\vec{x}-\mu_x}{\sigma_x}$ and $\frac{\vec{y}-\mu_y}{\sigma_y}$ to obtain the index $s(\vec{x},\vec{y})$.

Finally, the three components are combined to obtain the overall Structural SIMilarity index (SSIM) given by

$$\mathrm{SSIM}(\vec{x}, \vec{y}) = f\left(l(\vec{x}, \vec{y}), c(\vec{x}, \vec{y}), s(\vec{x}, \vec{y})\right) \tag{5}$$

The three components used to predict image quality are relatively independent as we cancel out the effect of each one of them by normalization before computing the next component.

We now define the three functions for luminance, contrast and structure comparisons. These functions are designed to satisfy the following properties:

1. Symmetry: $S(\vec{x}, \vec{y}) = S(\vec{y}, \vec{x})$
2. Bounded: $S(\vec{x}, \vec{y}) \leq 1$
3. Unique maximum: $S(\vec{x}, \vec{y}) = 1$ if and only if $\vec{x} = \vec{y}$

For luminance, $l(\vec{x}, \vec{y})$ is defined by

$$l(\vec{x}, \vec{y}) = \frac{2\mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{6}$$

where the constant C_1 is added to increase stability when the denominator becomes very small. One choice for C_1 given in [8] is $C_1 = (K_1 L)^2$, where L is the range of the pixel values and $K_1 << 1$ is a small constant. (6) is qualitatively consistent with Weber's law which is widely used to model luminance masking in the HVS. The Weber's law states that the just noticeable difference in luminance is directly proportional to the background luminance. The HVS is hence sensitive to relative luminance changes and not the absolute change. Letting

R represent the fractional change in luminance, the luminance of the distorted signal can be written as $\mu_y = (1 + R)\mu_x$. Then, we have

$$l(\vec{x}, \vec{y}) = \frac{2(1 + R)}{1 + (1 + R)^2 + \frac{C_1}{\mu_x^2}} \tag{7}$$

If C_1 is small enough compared to μ_x^2, then $l(\vec{x}, \vec{y})$ is a function of just R and this is consistent with Weber's law.

The contrast comparison function is given by

$$c(\vec{x}, \vec{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{8}$$

where $C_2 = (K_2L)^2$ is chosen as earlier and again, $K_2 << 1$. This measure is less sensitive to high base contrast than low base contrast, for the same absolute change in contrast. This is also qualitatively consistent with the contrast masking feature of the HVS.

The structure comparison is then performed between the luminance and contrast normalized images, $\frac{\vec{x}-\mu_x}{\sigma_x}$ and $\frac{\vec{y}-\mu_y}{\sigma_y}$. These lie in the hyperplane defined by $\sum_{i=1}^{N} x_i = 0$. The correlation between these vectors is used as the measure to quantify structural similarity between the images. The correlation is defined by

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{9}$$

where $\sigma_x y$ is given by

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) \tag{10}$$

Geometrically, $s(\vec{x}, \vec{y})$ corresponds to the cosine of the angle between these vectors in the hyperplane.

Finally, these three components are combined to obtain the SSIM index between $\vec{x}$ and $\vec{y}$ using

$$\text{SSIM}(\vec{x}, \vec{y}) = l(\vec{x}, \vec{y})^{\alpha} c(\vec{x}, \vec{y})^{\beta} s(\vec{x}, \vec{y})^{\gamma} \tag{11}$$

where $\alpha, \beta, \gamma > 0$ are parameters to adjust the relative importance of these parameters. Specific values for these constants given by $\alpha = \beta = \gamma = 1$ and $C_3 = \frac{C_2}{2}$ have been shown to be effective in [8].

2.2 SSIM Index in Image Quality Assessment

For image quality assessment, the SSIM index is applied locally rather than globally. This is because image features are highly non-stationary. Additionally, using local windows provides a quality map of an image, as opposed to a single index for the entire image, and can provide valuable information about local quality.

The quantities μ_x, σ_x, μ_y, σ_y and σ_{xy} are computed in a local sliding window, that is moved pixel by pixel over the entire image. To avoid blocking artifacts, the resulting values are weighted using a circularly symmetric 11×11 Gaussian function. The weighting function, $\vec{w} = \{w_i, i = 1, 2, \ldots N\}$, has a standard deviation of 1.5 samples and is normalized to have unit sum ($\sum_{i=1}^{N} w_i = 1$). The estimates of μ_x, σ_x and σ_{xy} are then modified accordingly as

$$\mu_x = \sum_{i=1}^{N} w_i x_i \tag{12}$$

$$\sigma_x = \left(\sum_{i=1}^{N} w_i (x_i - \mu_x)^2 \right)^{\frac{1}{2}} \tag{13}$$

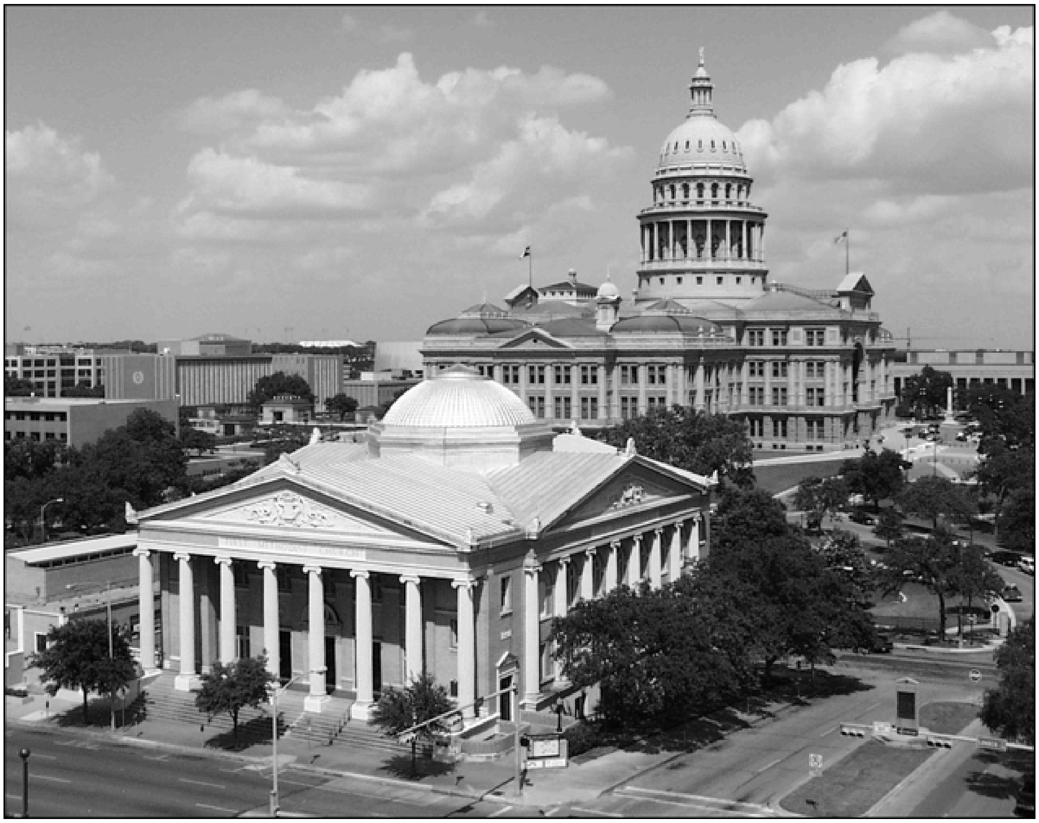
$$\sigma_{xy} = \sum_{i=1}^{N} w_i (x_i - \mu_x)(y_i - \mu_y) \tag{14}$$

The constants K_1 and K_2 used in the definition of $l(\vec{x}, \vec{y})$ and $c(\vec{x}, \vec{y})$ are chosen to be 0.01 and 0.03 experimentally. The overall quality of the entire image is defined to be the Mean SSIM (MSSIM) index and is given by

$$\mathrm{MSSIM}(\vec{X}, \vec{Y}) = \frac{1}{M} \sum_{i=1}^{N} \mathrm{SSIM}(\vec{x_i}, \vec{y_i}) \tag{15}$$

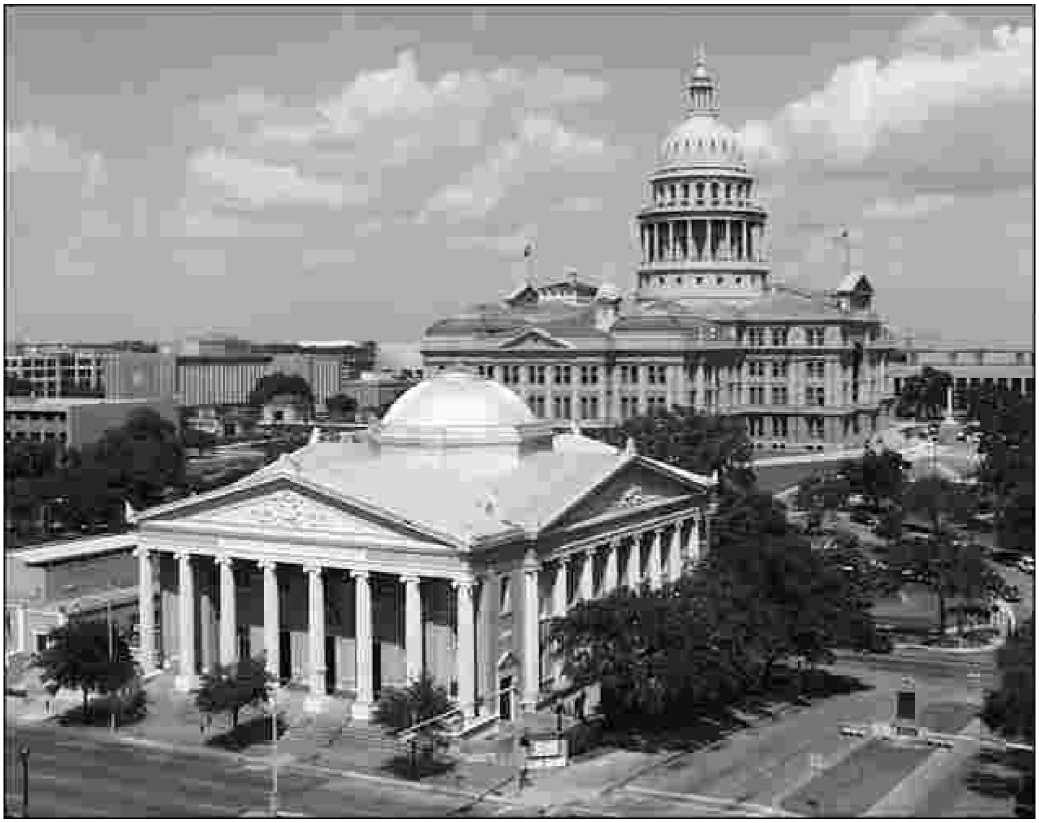
where $\vec{X}$ and $\vec{Y}$ are the reference and test images, $\vec{x_i}$ and $\vec{y_i}$ are the pixels in the i^{th} local window and M is the total number of windows in the image. A MATLAB implementation of the SSIM algorithm is available at [11].

Fig. 5 shows the performance of the SSIM index on an image. Fig. 5(a) and Fig. 5(b) show the original and JPEG compressed "Church and Capitol" images. The characteristic blocking artifacts of JPEG compression are clearly visible in the background of the image, on the roof of the church, in the trees and so on. Also, compression causes loss of high frequency information and the ringing artifacts are clearly visible along the edges of the Capitol dome.



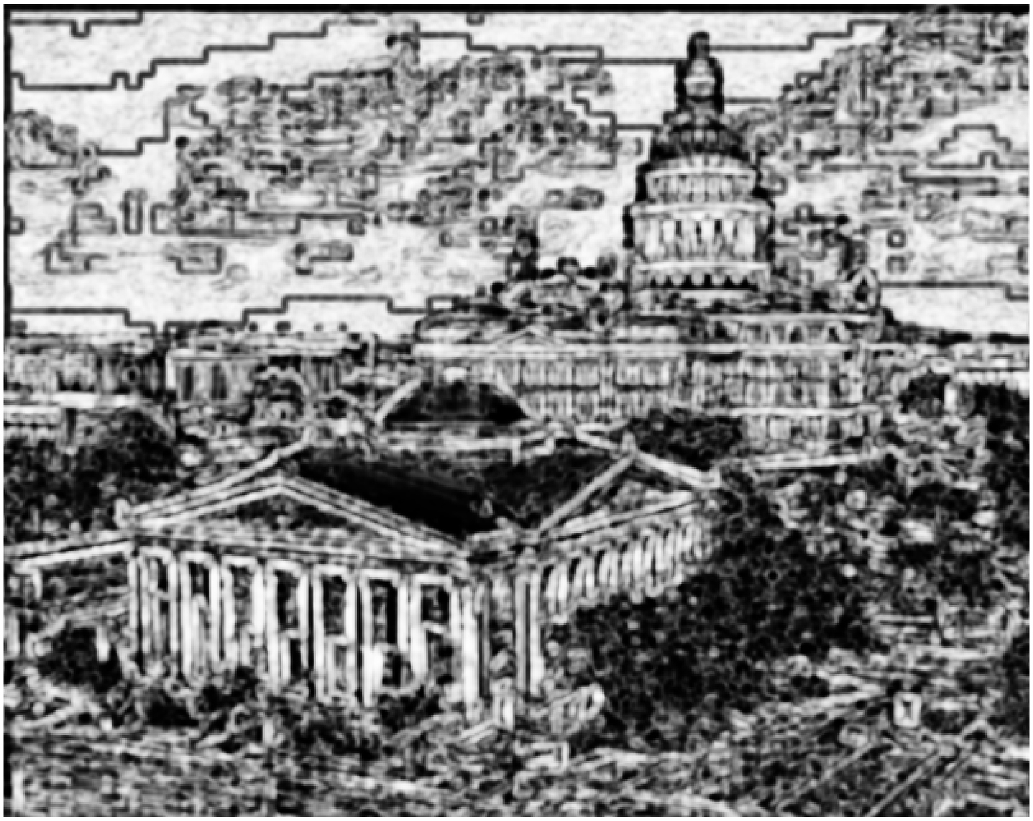
(a) Original ‘Church and Capitol’ image

Figure 5: Illustrative example of SSIM



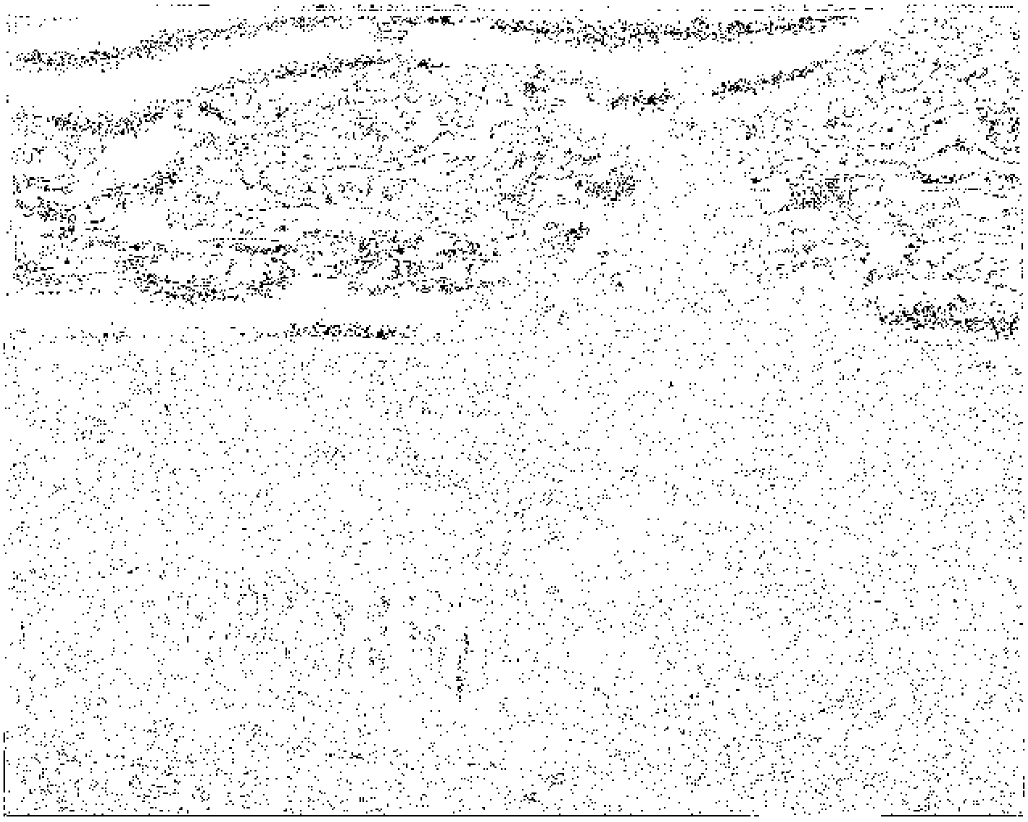
(b) JPEG compressed image

Figure 5: Illustrative example of SSIM



(c) SSIM Quality map

Figure 5: Illustrative example of SSIM



(d) Absolute Error Map

Figure 5: Illustrative example of SSIM

Fig. 5(c) clearly illustrates the effectiveness of SSIM in capturing the loss of quality in these regions. Brighter regions correspond to better visual quality and the map has been scaled for better visibility. The SSIM index clearly captures the loss of quality in the trees and the roof etc., and also captures the ringing artifacts along the edge of the Capitol. Fig. 5(d) shows the absolute error map between the images. This clearly fails to capture the distortion present in different regions of the image adequately.

3 The Information Theoretic Paradigm

The information theoretic paradigm approaches the quality assessment problem as an information fidelity problem, as opposed to a signal fidelity problem. MSE and SSIM are examples of signal fidelity criteria where MSE is a simple mathematical criterion, while SSIM attempts to measure closeness between signals in the perceptual domain. Information fidelity criteria, however, attempt to relate visual quality to the amount of information shared between the reference and test images [7]. This shared information can be quantified by the commonly used statistical measure, namely mutual information.

Here, the test image is assumed to be the output of a communication channel whose source is the reference image. The communication channel consists of a distortion channel as well as the HVS. The distortion channel models various operations like compression, blurring, additive noise, contrast enhancement and so on that lead to loss or enhancement of visual quality of the image. The HVS itself acts as a distortion channel as it limits the amount of information that is extracted from an image that passes through it [12]. This is the consequence of various properties of the HVS like luminance, contrast and texture masking that make certain distortions imperceptible. In fact, image compression algorithms rely on these properties of the HVS to successfully reduce the number of bits used to represent an image, without affecting the visual quality.

Information theoretic analysis requires accurate modeling of the source and the communication channel to quantify the information shared between the source and the output of the communication channel. Source modeling is accomplished using statistical models for *natural images*. Natural images are those that represent images from the real world and not necessarily images of nature. The statistical properties of such images have been studied by numerous researchers in the context of several applications such as compression, de-noising etc [13, 14]. These natural scene models attempt to characterize the distributions of natural images that distinguish them from images generated randomly by a computer.

In Section 3.1, we present the natural scene model that is used in the quality assessment algorithm. Section 3.2 discusses the distortion model and Section 3.3 presents the HVS model used here. The algorithm for quality assessment is presented in Section 3.4. We present certain illustrative examples that describe the properties of this novel quality measure in Section 3.5.

3.1 Natural Scene Model

The semi-parametric class of Gaussian Scale Mixtures (GSM) have been used to model the statistics of the wavelet coefficients of natural images [15]. A random vector $\vec{Y}$ is a GSM if $\vec{Y} \sim Z\vec{U}$ where Z is a scalar random variable, $\vec{U}$ is a zero mean Gaussian random vector and Z and $\vec{U}$ are independent. Z is called the *mixing* density. The GSM density can be represented as the integral of Gaussian density functions weighted by the mixing density; hence the term "mixtures". This class of distributions has heavy-tailed marginal distributions and the joint distributions exhibit certain non-linear dependencies that are characteristic of the wavelet coefficients of natural images [16].

Here, we model a coefficient and a collection of its neighbors in each sub-band of the wavelet decomposition of an image as a GSM. Specifically, we use the steerable pyramid which is a tight frame representation and splits the image into a set of sub-bands at different scales and orientations [17]. We model each sub-band of the wavelet decomposition by a random field $\mathcal{C} = \{\vec{C}_i, i \in \mathscr{I}\}$ given by

$$\mathcal{C} = \mathcal{Z}\mathcal{U} = \{Z_i\vec{U}_i, \quad i \in \mathscr{I}\} \tag{16}$$

where $\mathcal{Z} = \{Z_i, i \in \mathscr{I}\}$ is the mixing field, $\mathcal{U} = \{\vec{U}_i, i \in \mathscr{I}\}$ is an M-dimensional zero-mean Gaussian vector random field with covariance matrix $\mathbf{C_U}$ and $\mathscr{I}$ denotes a set of spatial indices. Also, $\mathcal{U}$ is assumed to be white, i.e., $\vec{U}_i$ is uncorrelated with $\vec{U}_j$ if $i \neq j$. Each sub-band of the wavelet decomposition is partitioned into non-overlapping blocks of M coefficients each and each block is modeled as the vector $\vec{C}_i$.

This model has certain nice properties that make it analytically tractable. Each $\vec{C}_i$ is normally distributed given Z_i. Also, given $\mathcal{Z}$, $\vec{C}_i$ is independent of $\vec{C}_j$ if $i \neq j$. Methods to estimate the multiplier $\mathcal{Z}$ and the covariance matrix $\mathbf{C_U}$ have been described in detail in [15, 14]. This GSM model is used as the source model for natural images in the following discussion.

3.2 Distortion Model

The distortion model that is used here is a signal attenuation and additive noise model in the wavelet domain given by

$$\mathcal{D} = \mathcal{G}\mathcal{C} + \nu = \{g_i \vec{C}_i + \vec{\nu}_i, i \in \mathscr{I}\} \tag{17}$$

where $\mathcal{C}$ denotes the random field that represents one sub-band of the reference image, $\mathcal{D} = \{\vec{D}_i, i \in \mathscr{I}\}$ denotes the random field representing the corresponding sub-band of the test image, $\mathcal{G}$ represents a deterministic scalar attenuation field and ν is a stationary, additive, zero-mean additive Gaussian noise field with covariance $\mathbf{C}_\nu = \sigma_\nu^2 \mathbf{I}$.

This model is both analytically tractable and computationally simple. This model can be used to describe most commonly occurring distortion types *locally*. The deterministic gain field $\mathcal{G}$ captures the loss of signal energy in sub-bands due to various operations like compression and blurring. The additive noise field accounts for local variations in the attenuated signal. Additionally, changes in image contrast can also be described locally by a combination of these two factors. For most practical distortions, g_i would be less than unity, but it could take larger values when the image is contrast enhanced, for instance.

3.3 HVS Model

The HVS model that is used here is also described in the wavelet domain. Natural scene models are in some sense the dual of HVS models, as the HVS has evolved by observing natural images [18]. Hence, many aspects of the HVS have already been incorporated in the natural scene model. It was experimentally determined that just an additive noise model for the HVS gives a marked improvement in the performance of the quality assessment algorithm [7].

The noise added by the HVS is modeled as a stationary, additive noise field $\mathcal{N} = \{\vec{N}_i, i \in \mathscr{I}\}$, where the $\vec{N}_i$ are zero-mean, uncorrelated Gaussian random vectors. We then have

$$\mathcal{E} = \mathcal{C} + \mathcal{N} \tag{18}$$

$$\mathcal{F} = \mathcal{D} + \mathcal{N}' \tag{19}$$

where $\mathcal{E}$ and $\mathcal{F}$ denote the output of the communication channel that is the HVS in this case, when the inputs are the reference and test images respectively. The noise field $\mathcal{N}$ is assumed to be independent of $\mathcal{C}$ and the covariance of $\mathcal{N}$, given by $\mathbf{C_N}$, is modeled using $\mathbf{C_N} = \sigma_N^2\mathbf{I}$. $\mathcal{N}'$ is modeled similarly. σ_N^2 is the variance of the HVS noise and is a parameter of the model that is derived empirically to optimize the performance of the algorithm. Although the performance of the quality assessment algorithm is affected by the choice of σ_N^2, it is quite robust to small changes in the value.

3.4 The Visual Information Fidelity Measure

Let $\vec{C}^N = \{\vec{C}_1, \vec{C}_2, \ldots \vec{C}_N\}$ denote N elements from $\mathcal{C}$. Let $\vec{E}^N$, $\vec{F}^N$, $\vec{D}^N$, Z^N and $\vec{U}^N$ be defined similarly. Also, let $\hat{Z}_i$ and $\hat{g}_i$ denote the estimated value of Z_i and g_i at coefficient i respectively. Similarly, $\hat{\sigma_N}$ and $\hat{\sigma_\nu}$ represent the estimated variances of the HVS noise and the noise in the distortion model respectively. Let the eigen decomposition of the covariance matrix $\mathbf{C_U}$ be given by

$$\mathbf{C_U} = \mathbf{Q \Lambda Q^T} \tag{20}$$

Then, it can be shown [7] using the models described above that

$$I(\vec{C}^N, \vec{E}^N | Z^N) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}\log_2\left(1+\frac{\hat{Z_i}^2\lambda_j}{\hat{\sigma_N}^2}\right) \tag{21}$$

$$I(\vec{D}^N, \vec{F}^N | Z^N) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}\log_2\left(1+\frac{\hat{g_i}^2\hat{Z_i}^2\lambda_j}{\hat{\sigma_N}^2+\hat{\sigma_\nu}^2}\right) \tag{22}$$

where $I(\vec{C}^N, \vec{E}^N | Z^N)$ represents the mutual information between the random field representing the reference image coefficients and the output of the HVS channel, conditioned on the mixing field [7]. Here, λ_i denotes the eigen values of the covariance matrix $\mathbf{C_U}$.

Notice that the form of this equation is very similar to the Shannon capacity of a communication channel. This is not surprising as the capacity of a communication channel is in fact defined by the mutual information between the source and the output of the channel, The quantity in the LHS of (21) can be interpreted as the reference image information, i.e. the amount of information that can be extracted by the HVS from an image that passes through it. Similarly, the quantity in the LHS of (22) can be thought of as the amount of information that can be extracted by the HVS from the reference image after it has passed through the distortion channel. The visual quality of the distorted image should relate to the amount of information that can be extracted by the HVS from the test image relative to the reference image information. If the amount of information that is extracted is very close to the reference image information, then the visual quality of the distorted image is very high as no loss of information occurs in the distortion channel.

The ratio of the two information measures has been shown to relate very well with visual quality [7]. Fig. 6 illustrates the block diagram to compute the VIF measure. Thus,

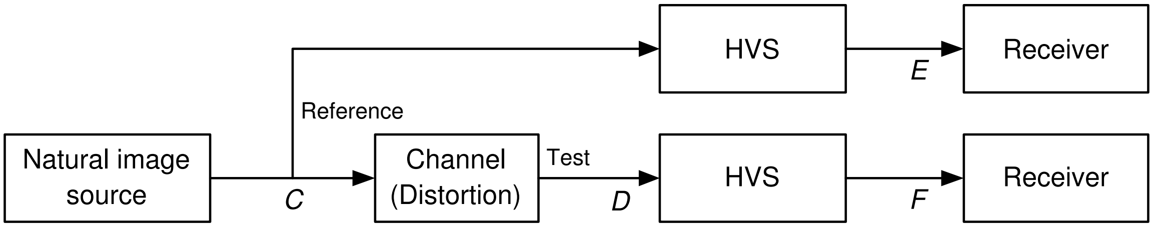


Figure 6: Block Diagram of the VIF Quality Assessment System [Reproduced from [12]]

the Visual Information Fidelity (VIF) criterion is given by

$$\text{VIF} = \frac{\sum_{j \in sub-bands} I(\vec{D}^{N,j}, \vec{F}^{N,j} | Z^{N,j})}{\sum_{j \in sub-bands} I(\vec{C}^{N,j}, \vec{E}^{N,j} | Z^{N,j})} \tag{23}$$

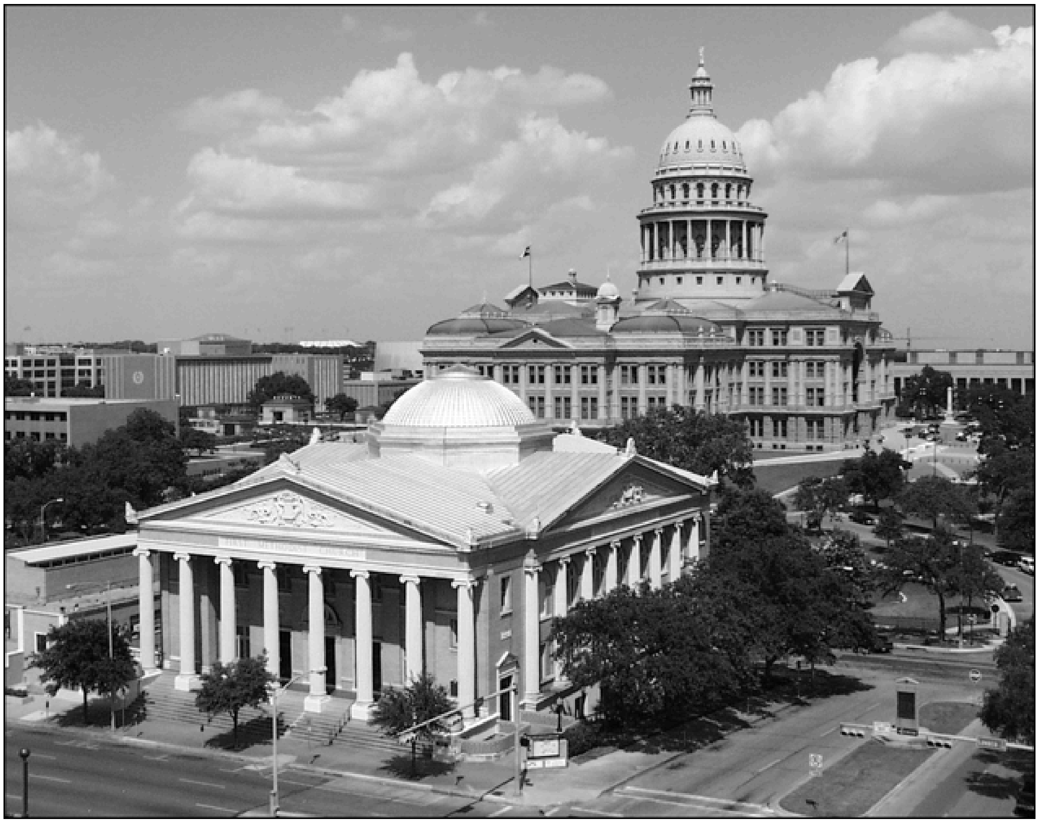
where $\vec{C}^{N,j}$ represents a set of N vectors from the j^{th} sub-bands. The VIF index for the entire image is hence calculated as the sum of this ratio of information measures over all sub–bands of interest, assuming that the random fields representing the sub-bands are all independent of each other. Although this assumption is not strictly true, it considerably simplifies the analysis without adversely affecting prediction accuracy.

Notice that the calculation of the VIF criterion involves the estimation of several parameters in the model. Z^N and $\mathbf{C_U}$ are parameters of the GSM model and ways to obtain the Maximum Likelihood estimates are discussed in [14]. Since we calculate the parameters of the model from the reference image, we are implicitly assuming that the random field $\mathcal{C}$ is ergodic. The parameters of the distortion model, namely $\hat{g}_i$ and $\hat{\sigma_\nu}^2$ can also be obtained easily using linear regression as the reference image is available [7]. The gain field $\mathcal{G}$ is assumed to be constant over small blocks and is estimated using the reference and test image coefficients in these blocks. Finally, as mentioned earlier, the variance of the HVS noise modeled by $\hat{\sigma_N^2}$ was obtained experimentally.

3.5 VIF in Image Quality Assessment

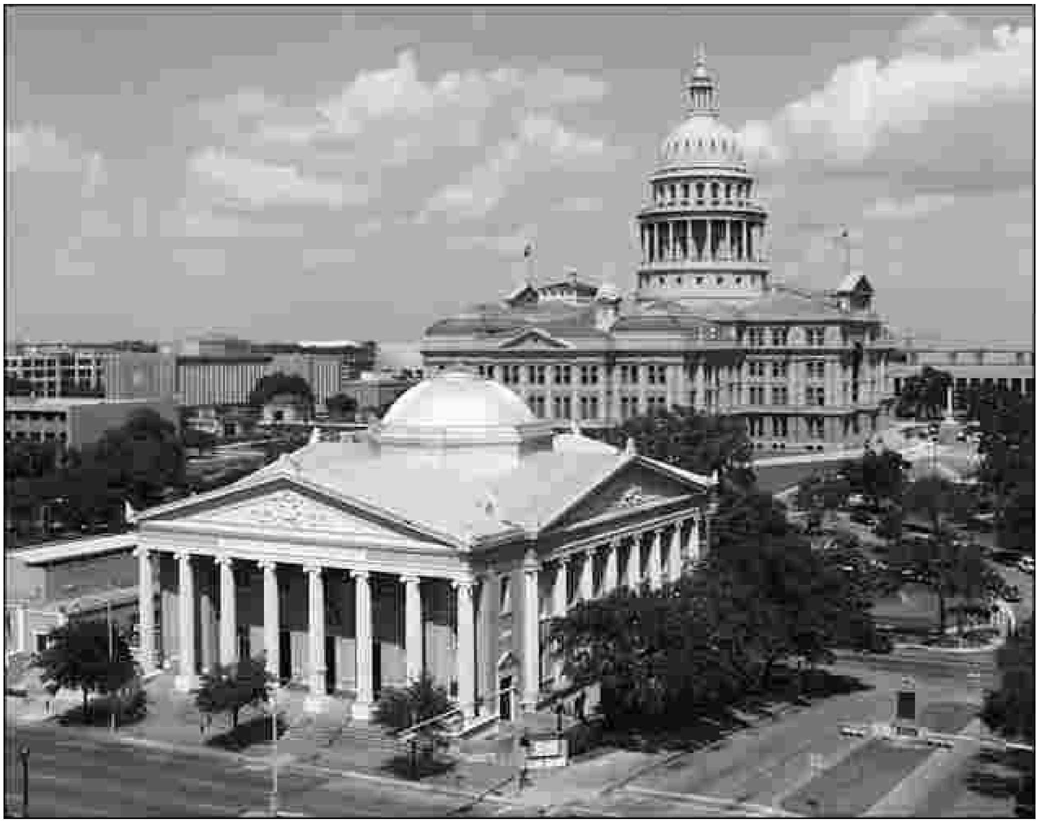
We now briefly discuss the properties of VIF. It is bounded below by zero. Additionally, VIF is exactly unity when the distorted image is identical to the reference image. Note that this was a design criterion in the structural approach as well. For most practical distortions that result in loss of information in the distortion channel, VIF takes values between 0 and 1. Finally, VIF can capture improvements in the quality of the image caused by, for instance, operations like contrast enhancement. In these cases, VIF takes values larger than unity. This is a remarkable property of VIF that distinguishes it from other metrics for image quality. Most other metrics assume that the reference image is of "perfect" quality and quantify only the *loss* in quality of the test image.

Fig. 7 presents an illustrative example of the power of VIF in predicting image quality. Fig. 7(a) shows the reference "Church and Capitol" image and Figure 7(b) shows the JPEG compressed version of the image. These are the same images on which the performance of the SSIM algorithm was illustrated earlier in Section 2.2. Fig. 7(c) shows the information map of the reference image. This corresponds to the denominator of the VIF measure and shows the spread of statistical information in the reference image. It is seen that the information is high in regions of high frequency, but is relatively low in the smooth regions of the image. Fig. 7(d) shows the VIF quality map of the image and illustrates the loss of information due to the distortion. Brighter regions correspond to better quality and the map has been contrast stretched for better visibility. The VIF measure is also successful in predicting the loss of quality in specific regions of the image that we visually noted. This includes the blocking artifacts in the background and roof of the church and the ringing artifacts on the edges of the Capitol.



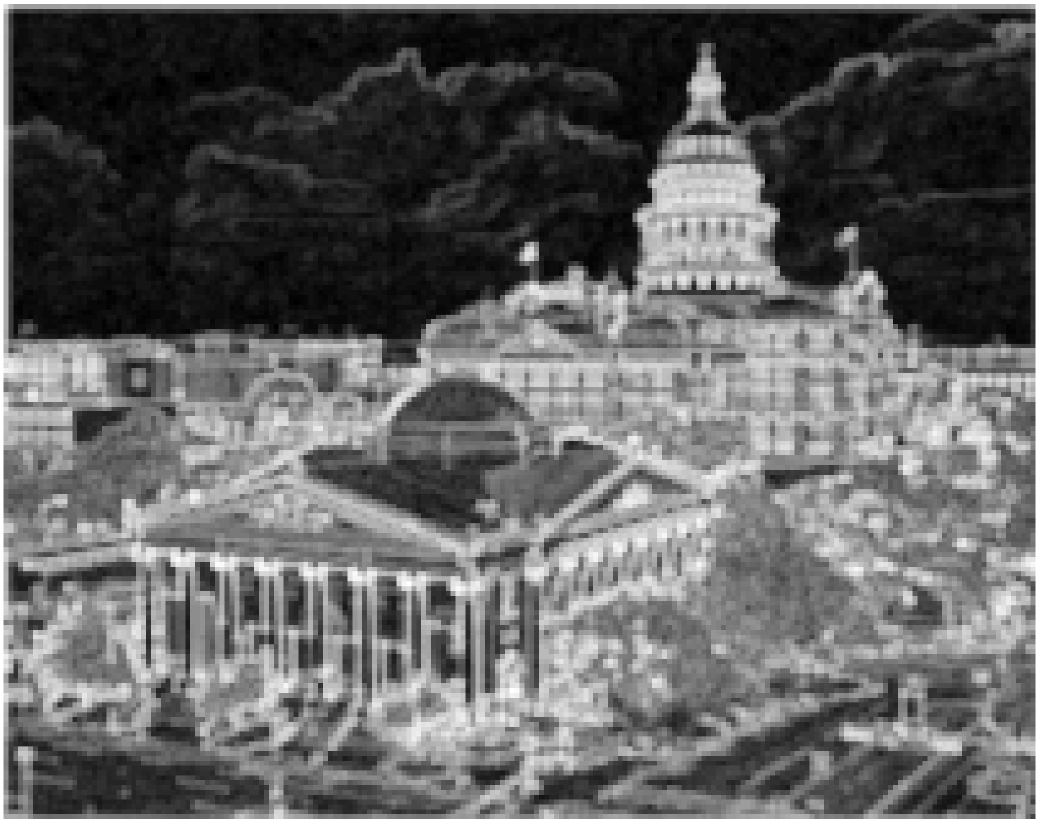
(a) Original ‘Church and Capitol’ image

Figure 7: Illustrative example of VIF



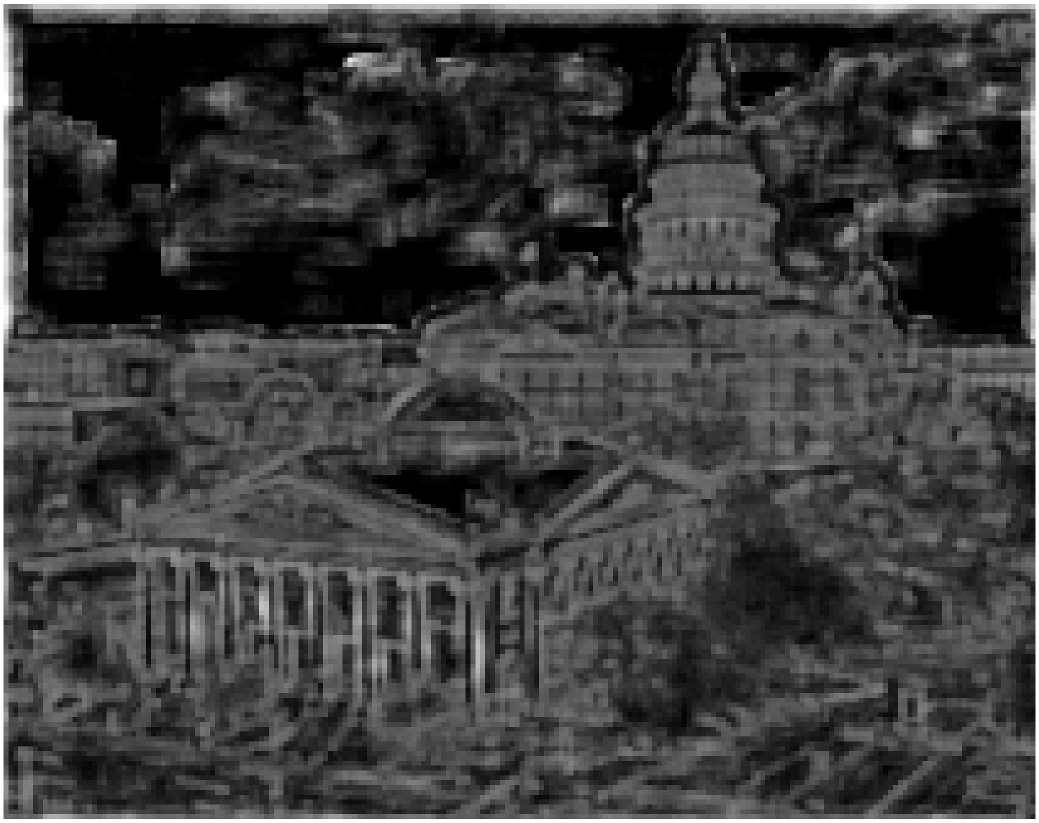
(b) JPEG compressed image

Figure 7: Illustrative example of VIF



(c) Image Information map

Figure 7: Illustrative example of VIF



(d) Absolute Error Map

Figure 7: Illustrative example of VIF

4 Performance of SSIM and VIF

The power of VIF and SSIM in predicting image quality was illustrated in the previous sections using example images. However, to test the performance of the quality assessment algorithm quantitatively, the Video Quality Experts Group (VQEG) Phase I FR-TV specifies four different metrics [19]. First, logistic functions are used in a fitting procedure to provide a non-linear mapping between the objective and subjective scores. The performance of the algorithm is then tested with respect to the following aspects of their ability to predict quality [20]:

1. *Prediction Accuracy:* The ability to predict the subjective score with low error.
2. *Prediction Monotonicity:* The ability to accurately predict relative magnitudes of subjective scores.
3. *Prediction Consistency:* The robustness of the predictor in assigning accurate scores over a range of different images.

The first two metrics used are the correlation coefficient between the subjective and objective scores after variance-weighted and non-linear regression analysis respectively. These metrics characterize the prediction accuracy of the objective measure. The third metric is the Spearman rank-order correlation coefficient between the objective and subjective scores, which characterizes the prediction monotonicity. Finally, the outlier ratio measures the prediction consistency.

Performance of the SSIM algorithm was tested on images that were compressed using JPEG and JPEG2000 at different bit rates. The details of the experiments conducted to obtain subjective quality scores can be found in [8]. Multiple subjects were asked to assign

Model	CC (Variance Weighted Regression)	CC(Non-linear Regression)	OR (Non-linear regression)	SROCC
PSNR	0.903	0.905	0.157	0.901
Sarnoff	0.956	0.956	0.064	0.947
MSSIM	0.967	0.967	0.041	0.963

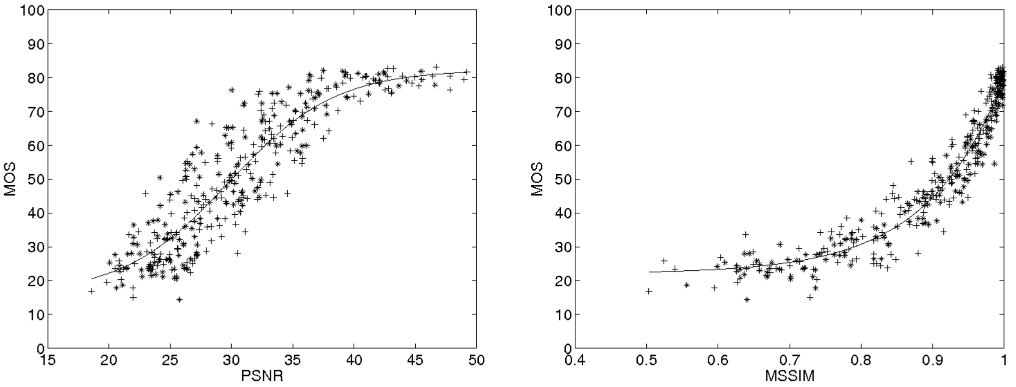
Table 1: Validation of MOS Scores for SSIM: The criteria are Correlation Coefficient (CC), Outlier Ratio (OR), Spearman Rank-Order Correlation Coefficient (SROCC)

quality scores to the same image along a linear scale marked with adjectives ranging from 'bad' to 'good'. However, not all subjects use the entire range of values in the numerical scale and this leads to variability [21]. The raw scores are hence converted to Z-scores. The Z-score z_j of a raw score x_j of a subject X is given by

$$z_j = \frac{x_j - \mu_x}{\sigma_x} \tag{24}$$

where $\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\sigma_x = \frac{1}{(N-1)}\sum_{i=1}^{N}(x_i - \mu_x)^2$, where $x_i, i = 1, \ldots, N$ are the raw scores assigned by subject X to all images. The Z-scores, hence, tell us how many standard deviations from the mean the given score is. The Z-scores are then re-scaled to fit the entire range of values from 1 to 100. The Mean Opinion Score (MOS) for each image is computed as the mean of the Z-scores for that image, after removing any outliers. The scatter plots of the MOS versus the SSIM model prediction are shown in Fig. 8. Each sample point represents one image. The best fitting logistic function is also plotted in the same graph. Also shown here is the scatter plot of the MOS versus Peak Signal to Noise Ratio (PSNR). PSNR is defined by

$$\text{PSNR} = 10\log_{10}\frac{255^2}{\text{MSE}} \tag{25}$$



(a) PSNR

(b) SSIM

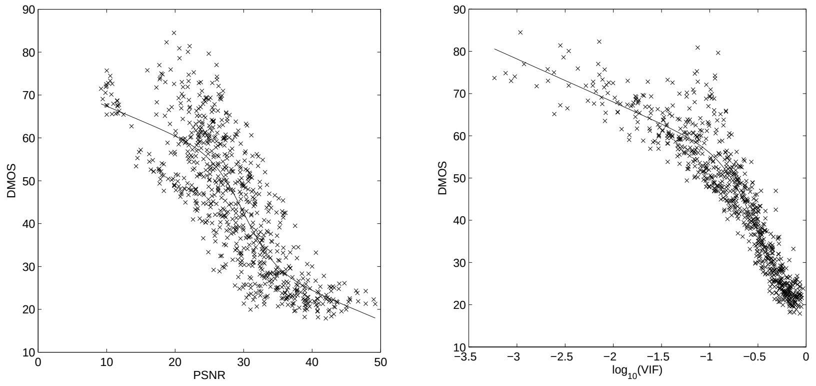
Figure 8: SSIM: Plot of MOS vs. model prediction and the best fitting logistic function to the subjective and objective scores [Reproduced from [10]]

Model	CC	OR	SROCC
PSNR	0.826	0.114	0.820
Sarnoff	0.901	0.046	0.902
VIF	0.949	0.013	0.949

Table 2: Validation of DMOS Scores for VIF: The criteria are Correlation Coefficient (CC), Outlier Ratio (OR), Spearman Rank-Order Correlation Coefficient (SROCC)

for 8-bit images and is just a function of the MSE. The plots clearly illustrate that SSIM performs much better than PSNR in predicting image quality. Table 1 shows the four metrics that were obtained for PSNR, the currently popular Sarnoff model (Sarnoff JND-Metrix 8.0 [22]) and SSIM. Again, SSIM outperforms PSNR for every metric.

Performance of the VIF algorithm was tested on JPEG and JPEG2000 compressed images, blurred images, AWGN corrupted images and images reconstructed after transmission errors in a JPEG2000 bitstream while passing through a fast fading Rayleigh channel. The details of the experiments conducted to obtain subjective quality scores can be found in [7]. Again, the raw scores were converted to Z-scores and then re-scaled to fit the range of values from 1 to 100. The Mean Opinion Score (MOS) for each image is computed. The



(a) PSNR

(b) VIF

Figure 9: VIF: Plot of DMOS vs. model prediction and the best fitting logistic function to the subjective and objective scores [Reproduced from [12]]

raw scores were also converted to difference scores between the reference and test images and then converted to Z-scores and finally, a Difference Mean Opinion Score (DMOS). The scatter plots of the MOS versus the VIF model prediction, as well as PSNR are shown in Fig. 9. The best fitting logistic function is also plotted in the same graph. The plots clearly illustrate that VIF performs much better than PSNR in predicting image quality. Table 2 shows the metrics that were obtained for PSNR, the Sarnoff model and VIF. Again, VIF outperforms PSNR by a sizeable margin for every metric.

Note that the databases used in testing the performance of VIF and SSIM are different, as indicated by the values of the correlation coefficients for both PSNR and the Sarnoff model. The correlation coefficients for both PSNR and SSIM are higher in Table 1 than the corresponding values in Table 2. This indicates that the database used to evaluate the performance of SSIM is in some sense easier than the one used to evaluate VIF. The metrics for SSIM and VIF presented here are, therefore, not comparable. Further comparisons can be found in [7].

5 Conclusions

This chapter presented two different top-down approaches to image quality assessment. Both methods have been shown to out-perform several state-of-the-art quality assessment algorithms. We have presented only some of the results here and further details can be found in [8, 7]. Structural approaches to image quality assessment attempt to measure the closeness of two signals by measuring the amount of structural distortion present in the distorted signal. This approach can be thought of as complementary to traditional bottom-up HVS based measures [10]. Information theoretic approaches, on the other hand, assume that the test image is the output of the channel through which the reference image passes and attempt to relate visual quality to the mutual information between the distorted and reference images. The equivalence of the information-theoretic setting to certain bottom-up HVS based systems has also been shown [23].

The success of both these methods in quality assessment and competitiveness to state-of-the-art methods has been demonstrated beyond doubt. The question as to which role each of them plays in the future of quality assessment research, however, is still unclear. It is even possible that the two paradigms will converge together in building a unified theory of quality assessment. Only further investigation into the structural and information-theoretic framework will answer these questions.

References

[1] G. Piella, "New quality measures for image fusion," in *Proc. Int. Conf. on Information Fusion*, 2004, pp. 542–546.

[2] D. Ryan and R. Tinkler, “Night pilotage assessment of image fusion,” in *Proc. SPIE*, vol. 2465, Orlando, Florida, April 1995, pp. 50–65.

[3] O. Rockinger, “Image sequence fusion using a shift-invariant wavelet transform,” in *Proc. IEEE Int. Conf. Image Processing*, vol. 13, 1997, pp. 288–291.

[4] L. J. Karam, “Lossless coding,” in *Handbook of Image and Video Processing*, A. C. Bovik, Ed. Academic Press, 2000, pp. 461–474.

[5] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. (2003) LIVE image quality assessment database. [Online]. Available: http://live.ece.utexas.edu/research/quality

[6] T. N. Pappas and R. J. Safranek, “Perceptual criteria for image quality evaluation,” in *Handbook of Image and Video Processing*, A. C. Bovik, Ed. Academic Press, 2000, pp. 669–684.

[7] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” *IEEE Trans. Image Processing*, Submitted for publication, 2003.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 1–14, April 2004.

[9] D. L. Ruderman, “The statistics of natural images,” *Network: Computation in Neural Systems*, no. 5, pp. 517–548, 1994.

[10] Z. Wang and A. C. Bovik, “Structural approaches to image quality assessment,” in *Handbook of Image and Video Processing*, A. C. Bovik, Ed. Academic Press, Submitted for Publication.

[11] Z. Wang. The ssim index for image quality assessment. [Online]. Available: http://www.cns.nyu.edu/ lcv/ssim

[12] H. R. Sheikh and A. C. Bovik, "Information theoretic approaches to image quality assessment," in *Handbook of Image and Video Processing*, A. C. Bovik, Ed. Academic Press, Submitted for Publication.

[13] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans. Image Processing*, vol. 8, no. 12, pp. 1688–1701, December 1999.

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

[15] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," in *Adv. Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K. R. Muller, Eds. Cambridge, MA: MIT Press, May 2000, vol. 12, pp. 855–861.

[16] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," *Appl. Comput. Harmon. Anal.*, vol. 11, no. 1, pp. 89–123, July 2001.

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

[18] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neural Science*, vol. 24, pp. 1193–1216, May 2001.

[19] VQEG. (2000, Mar.) Final report from the video quality experts group on the validation of objective models of video quality assessment. [Online]. Available: http://www.vqeg.org

[20] A. Rohaly et. al., "Video quality experts group: Current results and future directions," 2000. [Online]. Available: citeseer.ist.psu.edu/rohaly00video.html

[21] A. van Dijk, J. B. Martens, and A. B. Watson, "Quality assessment of coded images using numerical category scaling," in *Proc. SPIE*, vol. 2451, March 1995, pp. 90–101.

[22] Sarnoff Corporation. (2003) Jndmetrix technology. [Online]. Available: http://www.sarnoff.com/products_services/video_vision/jndmetrix/downloads.asp

[23] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information theoretic criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Processing*, Accepted for publication, 2003.