8.4

Information Theoretic Approaches to Image Quality Assessment

Hamid R. Sheikh
and Alan C. Bovik
The University of Texas at Austin

1 Introduction

Digital image and video processing systems deal, in large part, with signals that are meant for "human consumption." These image and video signals are reproductions of visual information that typically come from the three-dimensional (3D) environment. To design such systems to operate efficiently, automatic visual quality assessment algorithms are needed that can evaluate images and videos and report their quality without human involvement. Researchers have approached the problem of quality assessment from many directions, and some state-of-the-art methods were described in Chapters 8.2 and 8.3.

1.1 Full-Reference Quality Assessment

In the so-called full-reference (FR) quality assessment (QA) paradigm, a reference image of perfect quality is assumed to be available with respect to which a test (or distorted) image could be evaluated for quality. In the FRQA literature, researchers have tried to improve on the simple error measure, the mean squared error (MSE) and the corresponding peak signal-to-noise ratio (PSNR), as a measure of image quality, since MSE (and PSNR) are widely known to correlate too poorly with visual quality for most applications [3, 22]. Two broad schools of thoughts could roughly be identified among the researchers in FRQA: those who believe that modeling the salient physiologic and psychological components of the

human visual system (HVS) is the key to attaining accuracy in quality assessment, and those who try to address the problem in a more top-down manner by exploring signal similarity criteria for measuring image or video quality.

The HVS-based approach is the dominant approach in the QA literature. The underlying assumption is that the visual system processes different signal characteristics in different stages with different sensitivities. For example, signal luminance is processed in the retina, while localized frequency, scale, and orientation are processed in the primary visual cortex. Furthermore, many nonlinear effects have been discovered to operate within the HVS [20]. All of these factors make it reasonable to assume that a quantification of the strength of the error between the test and the reference signals, once the different sensitivities of the HVS have been accurately accounted for, would correlate better with visual quality than MSE applied to raw pixel values. Perhaps the earliest work in this direction was done by Mannos and Sakrison [5], and although HVS-based methods have been refined over the last three decades, and have been reasonably successful, there are many questions that arise in their design [21]. To circumvent some of these limitations, other researchers have explored using different signal fidelity criteria in hopes of capturing the hypothetical "visual information and structure" in images. Such methods have also achieved reasonable success.

1.2. Information Fidelity

In this chapter, we present information–theoretic approaches to the QA problem, where the QA problem is viewed as an information-fidelity problem rather than a signal-fidelity problem. An image source communicates to a receiver through a channel that limits the amount of information that could flow through it, thereby introducing distortions. This is pictorially shown in Fig. l. The output of the image source is the reference image, the output of the channel is the test image, and the goal is to relate the visual quality of the test image to the amount of information shared between the test and the reference signals, or more precisely, the *mutual information* between them. Although mutual information is a *statistical* measure of information fidelity, and may only be loosely related with what humans regard as image information, it places fundamental limits on the amount of cognitive information that could be extracted from an image. For example, in cases where the channel is distorting images severely, corresponding to low mutual information between the test and the reference, the ability of human viewers to obtain semantic information by discriminating and identifying objects in images is also hampered. Thus, information fidelity methods exploit the relationship between statistical image information and visual quality.

1.3 Natural Scene Statistics

Statistical models for signal sources and transmission channels are at the core of information theoretic analysis techniques. A fundamental component of information fidelity–based QA methods is a model for image sources, the so-called natural scene statistics (NSS) model. Images and videos whose quality needs to be assessed are usually optical images of the 3D visual environment, or *natural scenes*. Natural scenes form a very tiny subspace in the space of all possible image signals. One could imagine the sparseness of this subspace by hypothetically considering how long it would take for a random image generator to generate anything that looks like a natural scene. Researchers have devoted immense effort in developing sophisticated models that capture key statistical features of natural images. A review of these models has been presented in Chapter 4.7.

In contrast to HVS-based methods or signal fidelity measures that do not make any assumptions regarding the signals they would operate on, information fidelity–based methods depend *critically* on accurate models for the statistics of natural image sources. Most real-world distortion processes disturb these statistics and make the image or video signals *unnatural*. The information fidelity paradigm attempts to quantify this unnaturalness in terms of a well-known statistical fidelity measure: the mutual information.

In this chapter, we present two full-reference QA methods based on the information fidelity paradigm. Both methods share a common mathematic framework. The first method, the information fidelity criterion (IFC) [11], uses a distortion channel model as depicted in Fig. 1. The IFC quantifies the information shared between the test image and the distorted image. It has the advantage of being parameterless like MSE, in that there are no tuning parameters to optimize for. The other method that we will present in this chapter is the visual information fidelity (VIF) measure [9, 10], which uses an additional HVS channel model and two aspects of image information for quantifying perceptual quality: the information shared between the test and the reference image, and the information content of the reference image itself. This is depicted pictorially in Fig. 2.

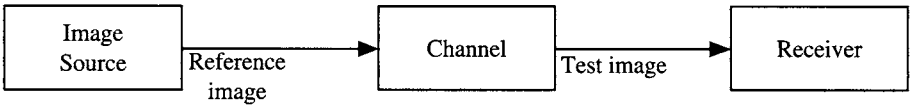


FIGURE 1 The information–fidelity problem: A channel distorts images and limits the amount of information that could flow from the source to the receiver. Quality should relate to the amount of information about the reference image that could be extracted from the test image.

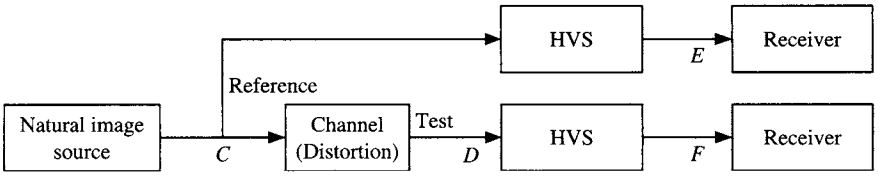


FIGURE 2 An information–theoretic setup for quantifying visual quality using a distortion channel model as well as a human visual system (HVS) model. The HVS also acts as a channel that limits the flow of information from the source to the receiver. Image quality could also be quantified using a relative comparison of the information in the upper path of the figure and the information in the lower path.

2 Mathematic Preliminaries

In this section, we will present some mathematic preliminaries and notational issues that are necessary for the development of information fidelity–based QA methods presented in this chapter. Specifically, we introduce vector Gaussian scale mixture (GSM) models for modeling natural image sources. We will also describe a simple, yet effective, channel model and a simple HVS model. We will then briefly introduce mutual information, and using the models presented in this section, we present the information fidelity–based quality assessment methods in Section 3.

2.1 The Source Model

Images and videos of the visual environment captured using high-quality capture devices operating in the visual spectrum are broadly classified as natural scenes. This differentiates them from text, computer-generated graphics scenes, cartoons and animations, paintings and drawings, random noise, or images and videos captured from nonvisual stimuli such as radar, sonar, x-rays, ultrasounds, and so forth. Natural scenes constitute an extremely tiny subset of the set of all possible images [4, 7]. Many researchers have discovered statistical models that describe the salient features of this subspace of natural images, and some of these models were reviewed in Chapter 4.7. Recently, scale–space–orientation analysis tools (loosely referred to as "wavelet" analysis in this chapter), for example the steerable pyramid [14, 15], have been successfully used for modeling natural images [1, 12, 19]. In these analysis methods, each subband of wavelet coefficients represents image features having similar scale and orientation. While wavelet analysis may usually be sufficient for approximately linearly decorrelating natural images, the nonlinear redundancies present in them cannot be modeled using any linear transform [13]. The GSM framework has been shown to be suitable for describing such dependencies between wavelet coefficients of natural images and their marginal statistics [18, 19].

Thus, the model for natural images that we use in this chapter is the GSM model in the wavelet domain, more specifically a vector GSM model. We will describe the model for only one subband of the wavelet decomposition at this point, and later generalize the results for multiple subbands. A GSM is a random field (RF) that can be expressed as a product of two independent RFs [19]. That is, a GSM $\mathcal{C} = \{\vec{C}_i : i \in \mathrm{I}\}$, where I denotes the set of spatial indices for the RF, can be expressed as:

$$\mathcal{C} = \mathcal{S} \cdot \mathcal{U} = \{S_i \cdot \vec{U}_i : i \in \mathrm{I}\} \tag{1}$$

where $\mathcal{S} = \{S_i : i \in \mathrm{I}\}$ is an RF of positive scalars and $\mathcal{U} = \{\vec{U}_i : i \in \mathrm{I}\}$ is a Gaussian vector RF with mean zero and covariance $\mathbf{C}_U$. $\vec{C}_i$ and $\vec{U}_i$ are M dimensional vectors, and we assume that for the RF $\mathcal{U}$, $\vec{U}_i$ is independent of $\vec{U}_j$, $\forall i \neq j$. In this chapter, we model each subband of a scale–space–orientation wavelet decomposition (such as the steerable pyramid [14]) of an image as a GSM. We partition the subband coefficients into nonoverlapping blocks of M coefficients each, and model block i as the vector $\vec{C}_i$. Thus image blocks are assumed to be uncorrelated with each other, and any linear correlations between wavelet coefficients are modeled only through the covariance matrix $\mathbf{C}_U$. Note that a "blocking" approach for capturing linear dependence between wavelet coefficients is not perfect, and blocks may continue to have residual dependence among each other. This means that the assumption of $\vec{U}_i$ being independent of $\vec{U}_j$, $\forall i \neq j$ would hold only approximately. The blocking approach, however, works fine for QA purposes.

One could easily make the following observations regarding this model: $\mathcal{C}$ is normally distributed given $\mathcal{S}$ (with mean zero, and covariance of $\vec{C}_i$ being $S_i^2\mathbf{C}_U$), that given S_i, $\vec{C}_i$ are independent of S_j for all $j \neq i$, and that given $\mathcal{S}$, $\vec{C}_i$ are conditionally independent of $\vec{C}_j, \forall i \neq j$ [19]. These properties of the GSM model make analytic treatment of information fidelity possible.

Researchers have shown that the GSM model can capture key statistical features of natural images: Linear dependencies in natural images can be captured by the GSM framework using a wavelet decomposition and the covariance matrix $\mathbf{C}_U$, the heavy-tailed marginal distributions of the wavelet coefficients can be modeled by using an appropriate distribution for $\mathcal{S}$, and the nonlinear dependencies between the wavelet coefficients of natural images can be captured by constraining the field $\mathcal{S}$ to be self-correlated [6, 19].

Researchers have explored models for describing the statistical properties of the field $\mathcal{S}$ for natural images as well

[6, 19]. However, we will not delve into a discussion about these models since they are not used in the methods that we present in this chapter.

2.2 The Distortion Model

The purpose of a distortion model is to describe how the statistics of an image are disturbed by a generic distortion operator. The distortion model that we have chosen provides important functionality while being mathematically tractable and computationally simple. It is a signal attenuation and additive noise model in the wavelet domain:

$$\mathcal{D} = \mathcal{GC} + \mathcal{V} = \{g_i \vec{C}_i + \vec{V}_i : i \in \mathrm{I}\} \tag{2}$$

where $\mathcal{C}$ denotes the RF from a subband in the reference signal, $\mathcal{D} = \{\vec{D}_i : i \in \mathrm{I}\}$ denotes the RF from the corresponding subband from the test (distorted) signal, $\mathcal{G} = \{g_i : i \in \mathrm{I}\}$ is a deterministic scalar gain field, and $\mathcal{V} = \{\vec{V}_i : i \in \mathrm{I}\}$ is a stationary additive zero-mean Gaussian noise RF with variance $\mathbf{C}_V = \sigma_V^2 \mathbf{I}$. The RF $\mathcal{V}$ is white and independent of $\mathcal{S}$ and $\mathcal{U}$. We constrain the field $\mathcal{G}$ to be slowly varying.

This model captures important, and complementary, distortion types: blur, additive noise, and global or local contrast changes. The underlying premise in the choice of this model is that in terms of their *perceptual annoyance*, distortion types that are prevalent in real-world systems could roughly be approximated *locally* as a combination of these distortions. The attenuation factors g_i would capture the loss of signal energy in a subband due to blur distortion, and the process $\mathcal{V}$ would capture the additive noise component separately. Figures 3 and 4 show some real-world distortions and the synthesized images from the corresponding distortion channel. The synthesized images were generated from the reference image and the estimated distortion channel for two types of channels: a signal attenuation with additive noise channel and an additive noise–only channel. A good distortion model is one where the distorted image and the synthesized image look equally *perceptually annoying*, and the goal of the distortion model is not to model image artifacts, but the perceptual annoyance of the artifacts. Thus, even though the distortion model may not be able to capture distortions such as ringing or blocking exactly, it may still be able to capture their perceptual annoyance. Notice that the signal attenuation and additive noise model can capture the effects of real-world distortions adequately in terms of the perceptual annoyance, whereas the additive-only distortion model performs quite poorly. For distortion types that are significantly different from blur and white noise, such as JPEG compression at very low bit rates (Fig. 3(E)), the model fails to reproduce the perceptual annoyance adequately (Fig. 3(F)), but it still performs much better than the additive only noise model shown in Figs. 5 and 6.

Moreover, changes in image contrast, such as those resulting from variations in ambient lighting or contrast-enhancement operations, are not modeled as noise, since they too could be incorporated into the attenuation field $\mathcal{G}$. For practical distortion types that could be described locally as a combination of blur and noise, g_i would be less than unity, while they could be larger than unity for some "distortion types" such as contrast enhancements.

2.3 The Human Visual System Model

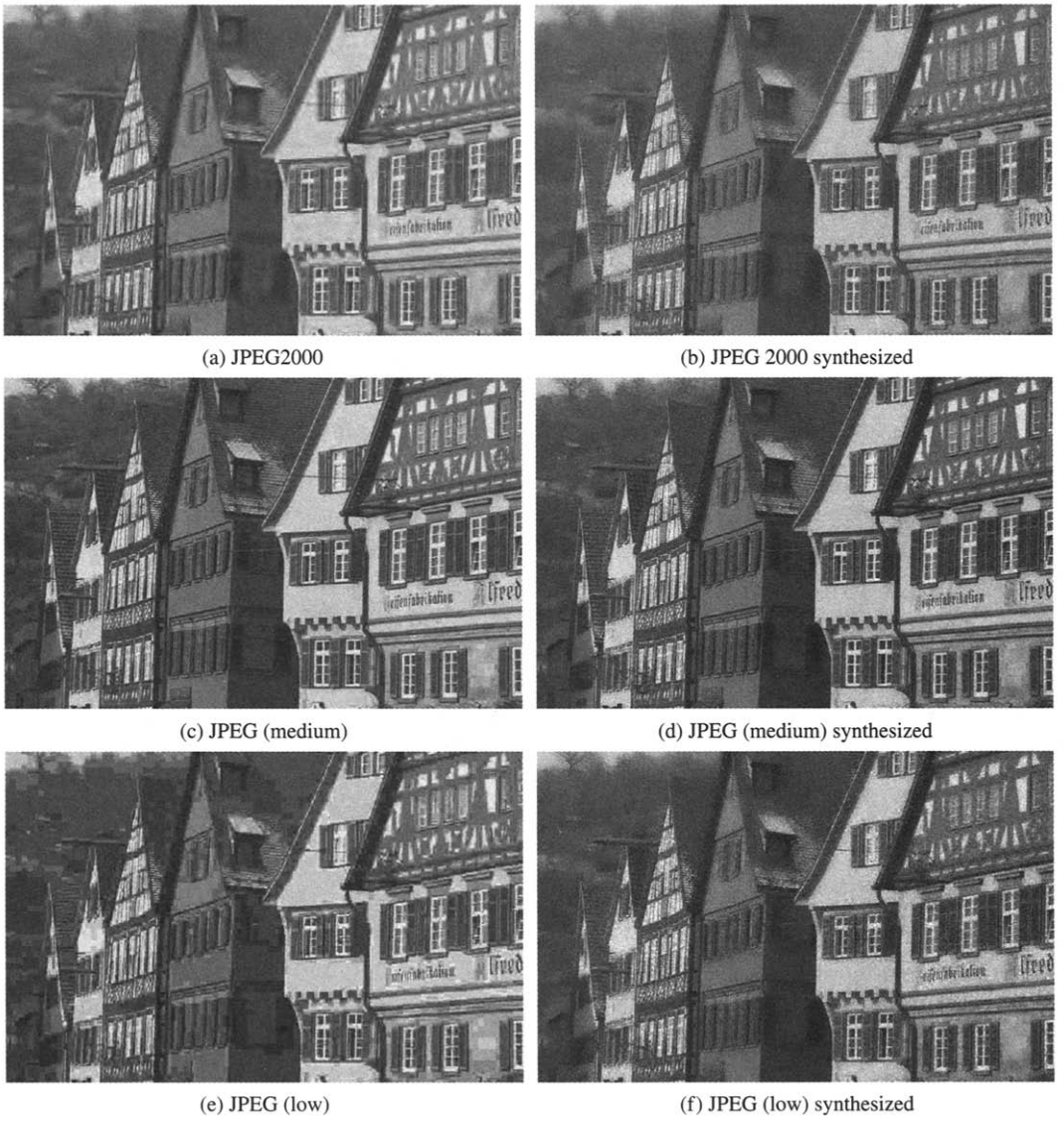
A HVS model is only required in the second of the two information fidelity–based QA methods that we will present in this chapter, that is, the VIF measure. The HVS model that we use is also described in the wavelet domain. Since HVS models are the dual of NSS models [16], many aspects of the HVS are already modeled in the NSS description, such as a scale-space-orientation channel decomposition, response exponent, and masking effect modeling [11]. The components that are missing include, among others, the optical point spread function (PSF), luminance masking, the contrast sensitivity function (CSF), and internal neural noise sources. Incidentally, it is the modeling of these components that is heavily dependent on viewing configuration, display calibration, and ambient lighting conditions.

In this chapter, we approach the HVS as a "distortion channel" that imposes limits on how much information could flow through it. Although one could model different components of the HVS using psychophysical data, the purpose of the HVS model in the information fidelity setup is to quantify the uncertainty that the HVS adds to the signal that flows through it. As a matter of analytical and computational simplicity, and more important to ease the dependency of the overall algorithm on viewing configuration information, we lump all sources of HVS uncertainty into one additive noise component that serves as a *distortion baseline* in comparison to which the distortion added by the distortion channel could be evaluated. We call this lumped HVS distortion *visual noise*, and model it as a stationary, zero mean, additive white Gaussian noise model in the wavelet domain. Thus, we model the HVS noise in the wavelet domain as stationary RFs $\mathcal{N} = \{\vec{N}_i : i \in \mathrm{I}\}$, and $\mathcal{N}' = \{\vec{N}'_i : i \in \mathrm{I}\}$ where $\vec{N}_i$ and $\vec{N}'_i$ are zero-mean uncorrelated multivariate Gaussian with the same dimensionality as $\vec{C}_i$:

$$\mathcal{E} = \mathcal{C} + \mathcal{N} \text{ (reference image)} \tag{3}$$

$$\mathcal{F} = \mathcal{D} + \mathcal{N}' \text{ (test image)} \tag{4}$$

where $\mathcal{E}$ and $\mathcal{F}$ denote the visual signal at the output of the HVS model from the reference and the test images in one subband respectively, from which the brain extracts cognitive information (Fig. 2). The RFs $\mathcal{N}$ and $\mathcal{N}'$ are assumed to be



(a) JPEG2000 (b) JPEG 2000 synthesized

(c) JPEG (medium) (d) JPEG (medium) synthesized

(e) JPEG (low) (f) JPEG (low) synthesized

FIGURE 3 Distorted images and their synthesized versions for the attenuation/additive noise distortion model. The images have been synthesized using two-band image decompositions. A good distortion model should be able to synthesize images whose perceptual annoyance is similar to the actual distortion. Note that the attenuation with additive noise model adequately captures the perceptual annoyance of real-world distortions. For distortions that deviate significantly from blur+noise, such as JPEG at low bit rates, the model's performance worsens, but is still better than the additive-only noise model of Fig. 5.

independent of $\mathcal{U}$, $\mathcal{S}$, and $\mathcal{V}$. We model the covariance of N and $\mathcal{N}'$ as:

$$\mathbf{C}_N = \mathbf{C}_{N'} = \sigma_N^2 \mathbf{I} \tag{5}$$

where σ_N^2 is an HVS model parameter (variance of the visual noise).

2.4 Differential Entropy and Mutual Information

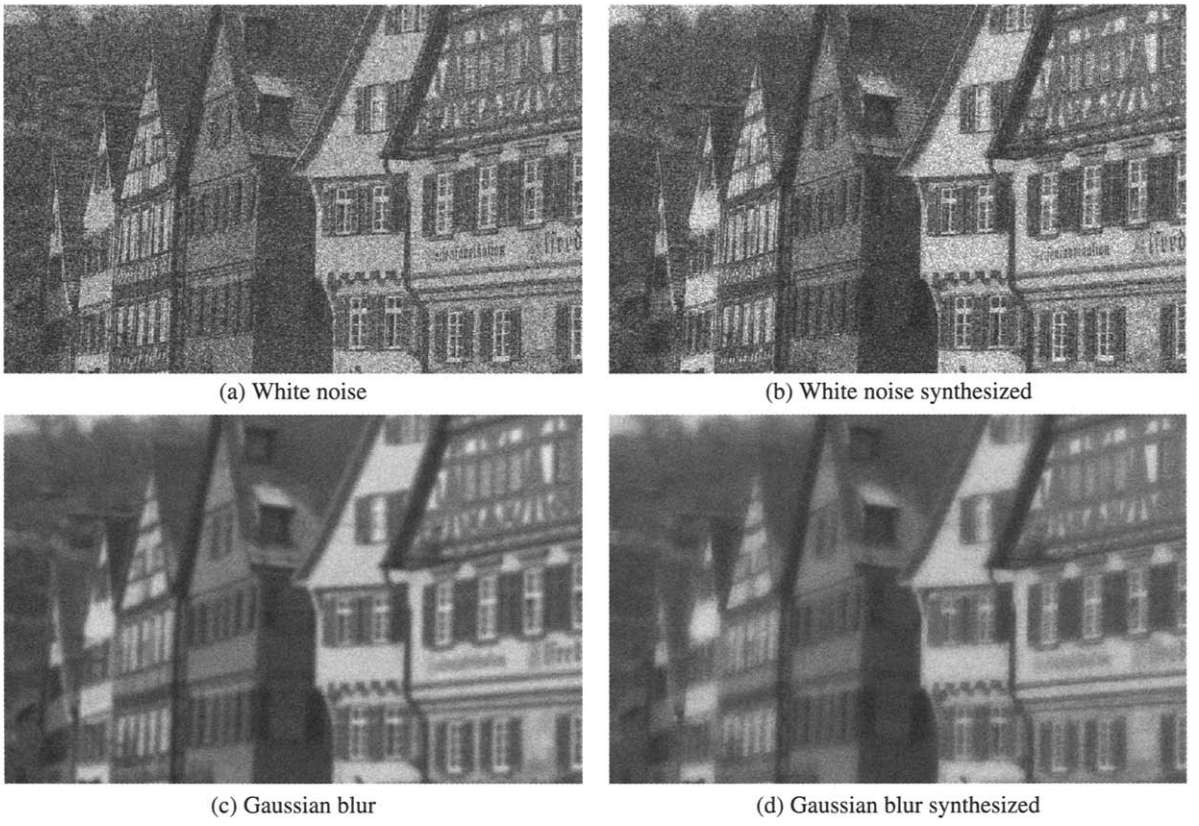
In this section, we will briefly review some definitions and results from information theory that are used later in this chapter. Readers are referred to [2] for a more detailed treatment.

The *differential entropy* of a continuous random variable X with a density $f(X)$ is defined as:

$$h(X) = -\int f(x) \log_2 f(x)\, \mathrm{d}x \tag{6}$$

The conditional differential entropy of a variable X given Y could similarly be defined:

$$h(X|Y) = -\iint f(x,y) \log_2 f(x|y)\, \mathrm{d}x\mathrm{d}y \tag{7}$$



(a) White noise (b) White noise synthesized

(c) Gaussian blur (d) Gaussian blur synthesized

FIGURE 4 Distorted images and their synthesized versions for the attenuation/additive noise distortion model.

For a normally distributed random vector $\vec{X}$ of dimensionality n, with mean $\vec{\mu}$ and covariance $\mathbf{K}$, the differential entropy can be shown to be:

$$h(\vec{X}) = \frac{1}{2}\log_2(2\pi e)^n|\mathbf{K}| \text{ bits} \tag{8}$$

where $|\mathbf{K}|$ denotes the determinant of $\mathbf{K}$.

The *mutual information* between two random variables X and Y with joint density $f(x, y)$ is defined as:

$$I(X; Y) = \int f(x,y)\log_2\frac{f(x,y)}{f(x)f(y)}\mathrm{d}x\mathrm{d}y \tag{9}$$

From the definitions it could be easily shown that

$$I(X; Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) \tag{10}$$

In case X and Y are independent, then $I(X; Y) = 0$.

From the definitions of differential entropy, mutual information, and the chain rule for probability density functions, it is easy to derive the following chain rules for a collection of N random variables $\{X_1, X_2, \ldots, X_N\}$ (denoted as X^N):

$$h(X^N) = \sum_{i=1}^{N} h(X_i|X^{i-1}) \tag{11}$$

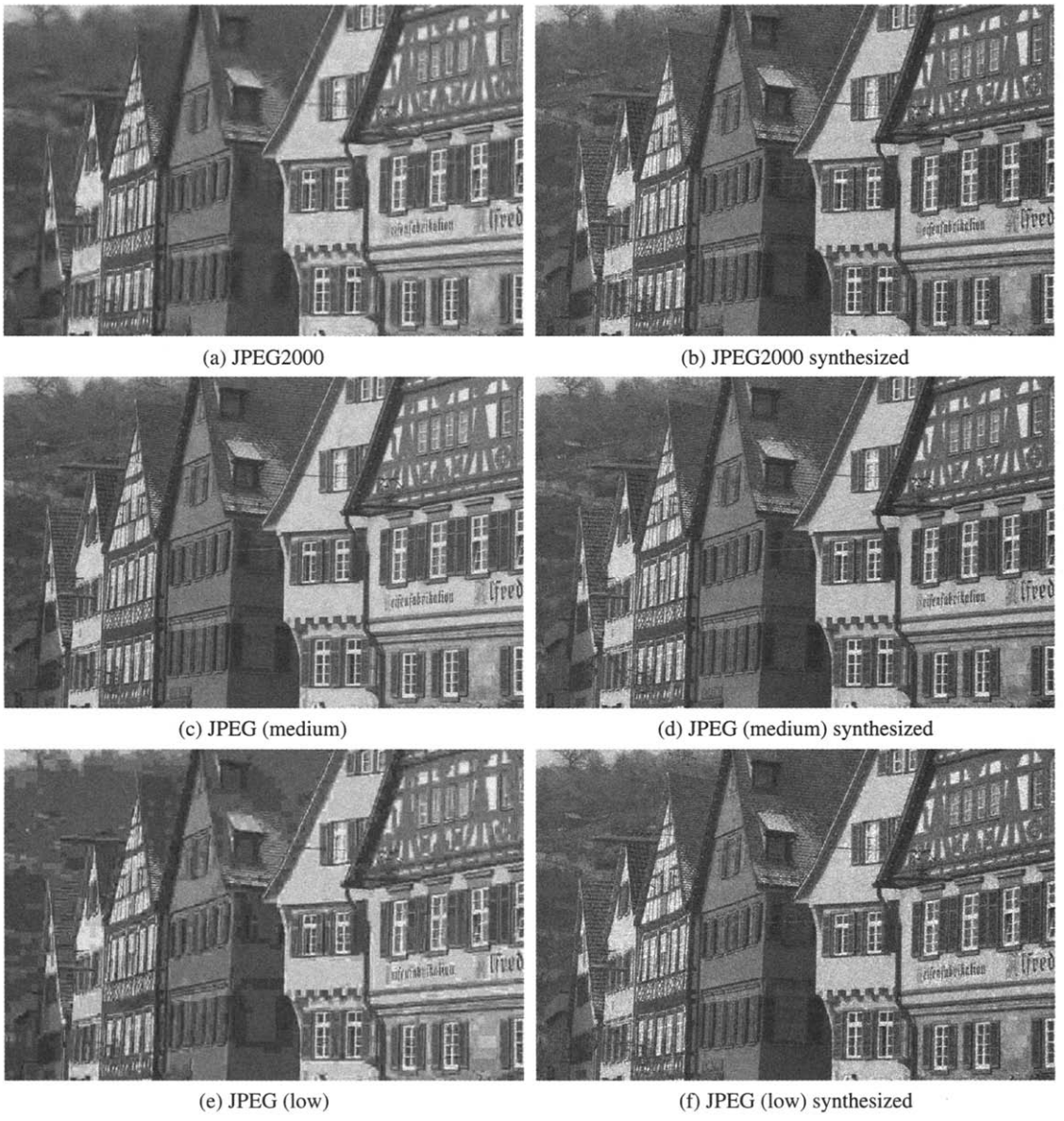
$$I(X^N; Y) = \sum_{i=1}^{N} I(X_i; Y|X^{i-1}) \tag{12}$$

$$I(X^N; Y^N) = \sum_{i=1}^{N}\sum_{j=1}^{N} I(X_i; Y_j|X^{i-1}, Y^{j-1}) \tag{13}$$

The notion of mutual information and the above equalities could easily be extended to conditional differential entropy and conditional mutual information. Thus, for example, if X and Y are conditionally independent given Z, then $I(X; Y|Z) = 0$.

3. Information Fidelity for Quality Assessment of Natural Images

In this section, we will describe two recently proposed methods for image QA based on the information fidelity



(a) JPEG2000 (b) JPEG2000 synthesized

(c) JPEG (medium) (d) JPEG (medium) synthesized

(e) JPEG (low) (f) JPEG (low) synthesized

FIGURE 5 Distorted images and their synthesized versions for the additive noise distortion model. Note that the model fails to capture blurring adequately, and the synthesized images have a much different perceptual quality.

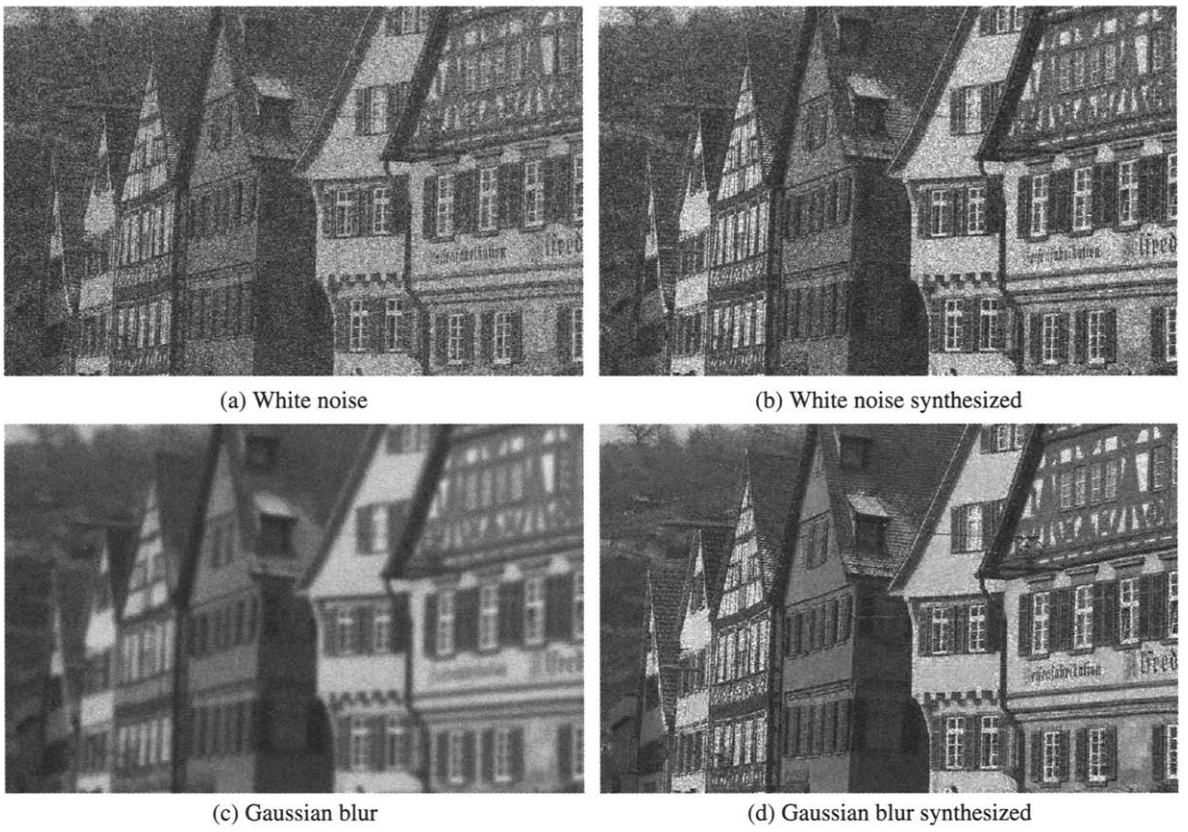
paradigm. Both of these methods quantify various aspects of image information. Results are presented in Section 5.

3.1 The Information Fidelity Criterion

The IFC quantifies the information shared between a test and the reference image. The reference image is assumed to pass through a channel yielding the test image, and the mutual information between the reference and the test images is used for predicting visual quality.

Let $\vec{C}^N = \{\vec{C}_1, \vec{C}_2, \ldots, \vec{C}_N\}$ denote N elements from C. Let S^N and $\vec{D}^N$ be correspondingly defined. The mutual information between the test and the reference wavelet coefficients $I(\vec{C}^N; \vec{D}^N)$ quantifies the amount of information shared between them. However, we are interested in the quality of a particular reference-test image pair, and not the average quality of the ensemble of images as they pass through the distortion channel.[1] It is therefore reasonable to *tune* the natural scene model to a specific reference image by treating $I(\vec{C}^N; \vec{D}^N | S^N = s^N)$ instead of $I(\vec{C}^N; \vec{D}^N)$, where s^N denotes a realization of S^N for a particular reference image. The realization s^N could be thought of as 'model parameters' for the associated reference image. The conditioning on $\mathcal{S}$ is intuitively in line with divisive normalization models for the

[1]For some design applications where the distortion channel is being designed to maximize visual quality, it would make more sense to optimize the design for the ensemble of images instead.



(a) White noise (b) White noise synthesized

(c) Gaussian blur (d) Gaussian blur synthesized

FIGURE 6 Distorted images and their synthesized versions for the additive noise distortion model. Note that the model fails to capture blurring adequately, and the synthesized images have a much different perceptual quality.

visual neurons [11], and lends the IFC to analytical tractability as well. In this chapter we will denote $I(\vec{C}^N; \vec{E}^N|\vec{S}^N = s^N$ as $I(\vec{C}^N; \vec{E}^N|s^N)$. With the stated assumptions on $\mathcal{C}$ and the distortion model, it can easily be shown that [11]:

$$I(\vec{C}^N; \vec{D}^N|s^N) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{M}\log_2\left(1 + \frac{g_i^2 s_i^2 \lambda_k}{\sigma_V^2}\right) \qquad (14)$$

where λ_k are the eigenvalues of $\mathbf{C}_U$.

Note that in the above treatment it is assumed that the model parameters s^N, $\mathcal{G}$ and σ_V^2 are known. Details of practical estimation of these parameters are given in Section 4.

In the development of the IFC, we have so far only dealt with one subband. One could easily incorporate multiple subbands by assuming that each subband is completely independent of others in terms of the RFs as well as the distortion model parameters. Thus, the IFC is given by:

$$\text{IFC} = \sum_{j\in\text{subbands}} I(\vec{C}^{N,j}; \vec{D}^{N,j}|s^{N,j}) \qquad (15)$$

where the summation is carried over the subbands of interest, and $\vec{C}^{N,j}$ represent N_j elements of the RF $\mathcal{C}_j$ that describes the coefficients from subband j, and so on.

3.2 Properties of Information Fidelity Criterion

The IFC has a number of interesting properties. One interesting feature of the IFC is that like the MSE, it does not depend on model parameters such as those associated with display device physics, data from visual psychology experiments, viewing configuration information, or stabilizing constants. The IFC does not require training data either. In the implementation of the IFC, however, a number of simulation parameters need to be introduced, such as those required for model parameter estimation. We will discuss parameter estimation in Section 4. Yet, the minimal reliance of the IFC on parameters that could be "tuned" for performance, and no reliance on psychovisual data or viewing geometry data, is a desirable feature.

3.3 Similarities Between Information Fidelity Criterion and Human Visual System–Based Methods

Interestingly, it has been shown that the IFC is functionally similar to HVS-based FRQA algorithms [11]. Figure 7 shows an HVS-based FRQA method that computes the error signal between the processed reference and test signals, and then

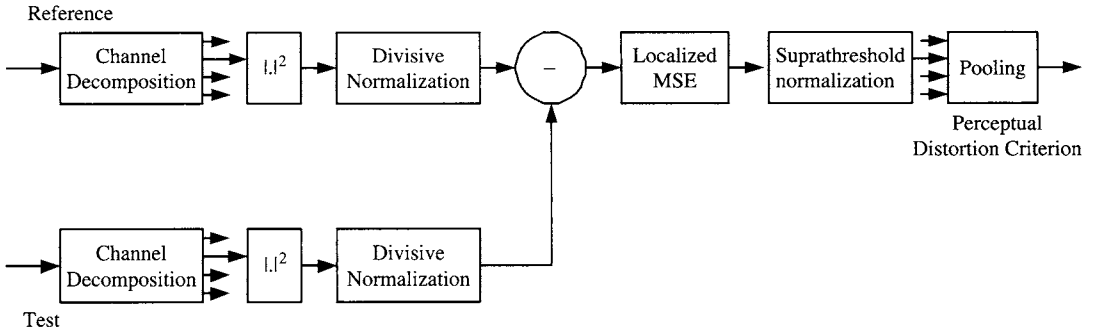


FIGURE 7 Information fidelity criterion using scalar Gaussian scale mixture model has been shown to be functionally equivalent to the HVS based full-reference quality assessment method shown here [11].

processes the error signal before computing the final perceptual distortion measure. A number of key similarities with most HVS QA methods are immediately evident. These include a scale-space-orientation channel decomposition, response exponent, masking effect modeling, localized error pooling, suprathreshold effect modeling and a final pooling into a quality score. It has been shown that the method shown in Fig. 7 is functionally equivalent to IFC using a scalar GSM model. The differences between Fig. 7 and IFC have to do with improved source and distortion modeling, and the interplay of pooling and suprthreshold effect modeling. These have been discussed in detail in [11].

3.4 The Visual Information Fidelity Measure

The IFC presented in Section 3.3 did not take into consideration the fact that the visual system also limits the amount of information that could be extracted from a visual signal that passes through it. In this respect, it too is a "distortion channel." The visual information fidelity (VIF) measure [9] attempts to quantify the loss of information to the HVS channel relative to the amount of information lost from the signal to the distortion channel for evaluating visual quality.

Let $\vec{C}^N = \{\vec{C}_1, \vec{C}_2, \ldots, \vec{C}_N\}$ denote N elements from $\mathcal{C}$. Let S^N, $\vec{D}^N$, $\vec{E}^N$ and $\vec{F}^N$ be correspondingly defined (figure 2). Once again, in this section we will assume that the model parameters s^N, $\mathcal{G}$, σ_V^2, and σ_N^2 are known, and as before, we will treat conditional mutual information quantities $I(\vec{C}^N; \vec{E}^N|s^N)$ and $I(\vec{C}^N; \vec{F}^N|s^N)$ only.

It can be shown [9] that:

$$I(\vec{C}^N; \vec{E}^N|s^N) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{M}\log_2\left(1+\frac{s_i^2\lambda_k}{\sigma_N^2}\right) \tag{16}$$

$$I(\vec{C}^N; \vec{F}^N|s^N) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{M}\log_2\left(1+\frac{g_i^2 s_i^2\lambda_k}{\sigma_V^2+\sigma_N^2}\right) \tag{17}$$

where λ_k are the eigenvalues of $\mathbf{C}_U$.

$I(\vec{C}^N; \vec{E}^N|s^N)$ and $I(\vec{C}^N; \vec{F}^N|s^N)$ represent the information that could ideally be extracted by the brain from a particular subband of the reference and the test images respectively. Therefore, $I(\vec{C}^N; \vec{E}^N|s^N)$ could be interpreted as the *reference image information*. Intuitively, visual quality should relate to the amount of image information that the brain could extract from the test image relative to the amount of information that the brain could extract from the reference image. For example, if the reference image information is 5.0 bits of information per pixel and the test image has 4.8 bits per pixel, then relatively little information has been lost to the distortion channel. In contrast, if the reference image information were 10.0 bits per pixel, and the test image still has 4.8 bits per pixel, then a relatively large amount of information has been lost to the distortion channel, and the visual quality of the test image should be inferior.

A simple *ratio* of the two information measures relates quite well with visual quality [9]. It is easy to motivate the suitability of this relationship between image information and visual quality. When a human observer sees a distorted image, he or she has an idea of the amount information that he expects to receive in the image (modeled through the known $\mathcal{S}$ field), and it is natural to expect the fraction of the expected information that is actually received from the distorted image to relate well with visual quality.

As before, the VIF could easily be extended to incorporate multiple subbands by assuming that each subband is completely independent of others in terms of the RFs as well as the distortion model parameters. Thus, the VIF is given by:

$$\text{VIF} = \frac{\sum_{j\in\text{subbands}} I(\vec{C}^{N,j}; .\vec{F}^{N,j}|s^{N,j})}{\sum_{j\in\text{subbands}} I(\vec{C}^{N,j}; .\vec{E}^{N,j}|s^{N,j})} \tag{18}$$

where we sum over the subbands of interest, and $\vec{C}^{N,j}$ represent N elements of the RF $\mathcal{C}_j$ that describes the coefficients from subband j, and so on.

The VIF given in (18) is computed for a collection of wavelet coefficients that could either represent an entire subband of an image, or a spatially localized set of subband

coefficients. In the former case, the VIF is a single number that quantifies the information fidelity for the entire image, whereas in the latter case, a sliding-window approach could be used to compute a *quality map* that could visually illustrate how the visual quality of the test image varies over space.

3.5 Properties of Visual Information Fidelity

The VIF has a number of interesting features. First note that VIF is bounded below by zero (such as when $I(\vec{C}^N; \vec{F}^N|s^N) = 0$ and $I(\vec{C}^N; \vec{E}^N|s^N) \neq 0$), which indicates that all information about the reference image has been lost in the distortion channel. Second, if the test image is an exact copy of the reference image, then VIF is *exactly* unity. Thus, for all practical distortion types, VIF would lie in the interval [0, 1]. Third, and this is where VIF has a distinction over traditional QA methods, a linear contrast enhancement of the reference image that does not add noise would result in a VIF value *larger* than unity, signifying that the contrast-enhanced image has a visual quality *superior* to the reference image! It is a common observation that contrast enhancement of images increases their perceptual quality unless quantization, clipping, or display nonlinearities add additional distortion. Theoretically, contrast enhancement results in a higher SNR at the output of the HVS neurons, which allows the brain to have a greater ability to discriminate objects present in the visual signal. This increase in SNR and the resulting improvement in visual quality is captured by the VIF.

While it is common experience that even linear point-wise contrast enhancement improves quality to a certain extent only, and that the quality starts deteriorating beyond a certain enhancement factor, we believe that in the real world, the perceived quality increases with contrast enhancement over many orders of magnitude. Illumination increase in the environment (which leads to an increase in the contrast of the light signals entering the eye as well, contrast being the signal that is encoded by the retina and sent to the brain) increases our perception of the quality of the perceived image over many orders of magnitude until the HVS neurons are driven to saturation. The effect of limited piece-wise contrast improvement on a computer is therefore more an artifact of limited machine precision and display nonlinearities.

It is reasonable to envision extensions of the notion of quantifying *improvement* in visual quality of images by other image enhancement operations such as deblurring or denoising using a similar information–theoretic paradigm.

Figure 8 shows a reference image and three of its distorted versions that come from three different types of distortion, all of which have been adjusted to have about the same MSE with the reference image. The distortion types illustrated in Fig. 8 are contrast stretch, Gaussian blur, and JPEG compression. In comparison with the reference image, the contrast enhanced image has a better visual quality despite the fact that the "distortion" (in terms of a perceivable difference with the reference image) is clearly visible. A VIF value larger than unity indicates that the perceptual difference in fact constitutes improvement in visual quality. In contrast, both the blurred image and the JPEG-compressed image have clearly visible distortions and poorer visual quality, which is captured by a low VIF measure.

Figure 9 illustrates spatial quality maps generated by VIF. Figure 9(A) shows a reference image and Fig. 9(B) the corresponding JPEG2000-compressed image in which the distortions are clearly visible. Figure 9(C) shows the reference image information map. The information map shows the spread of statistical information in the reference image. The statistical information content of the image is low in flat image regions, whereas in textured regions and regions containing strong edges, it is high. The quality map in Fig. 9(D) shows the proportion of the image information that has been lost to JPEG2000 compression. Note that due to the nonlinear normalization in the denominator of VIF, the scalar VIF value for a reference/test pair is *not* the mean of the corresponding VIF map!

3.6 Visual Information Fidelity and Human Visual System–Based Quality Assessment Methods

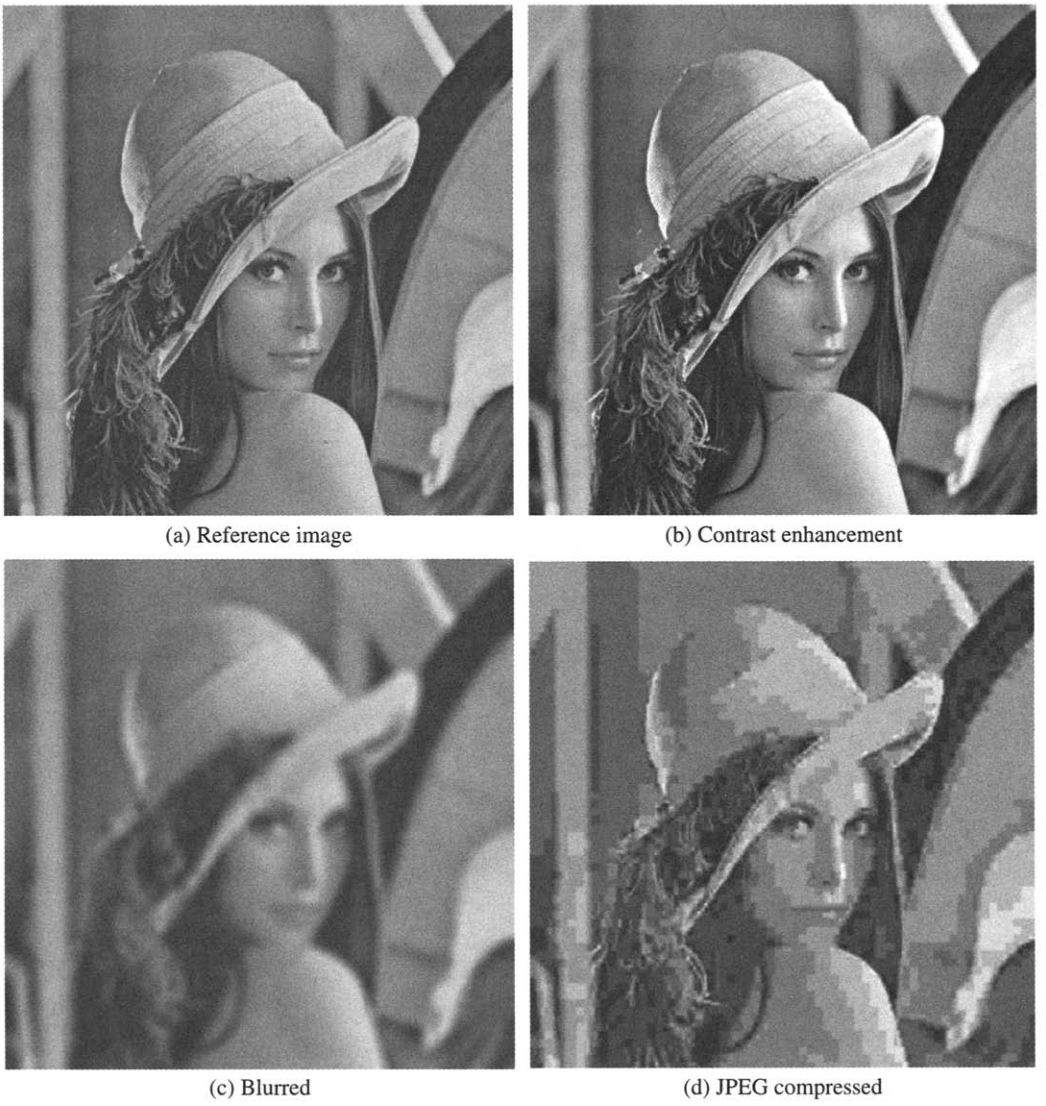
Note that VIF is basically IFC normalized by the reference image information (except for the visual noise component), and most of the discussion of the similarities and contrasts between IFC and HVS-based QA methods carries onto VIF as well. The normalization by reference image information is something that is currently not implemented in HVS-based QA methods, and its connections with the HVS needs further research.

4. Implementation Issues

To implement IFC or VIF a number of assumptions are needed about the source, distortion, and HVS models, as well as estimation methods for the different parameters. We outline them in this section.

4.1 Assumptions about the Source Model

Note that while we model natural image sources as being stochastic in nature, the QA algorithms need to operate on particular *realizations* of these sources. Mutual information (and hence the IFC and VIF) can only be calculated between the RFs and not their realizations, that is, a particular reference and the test image under consideration. We will assume ergodicity of the RFs and that reasonable estimates for the statistics of the RFs can be obtained from their realizations. Mutual information is then quantified between



(a) Reference image (b) Contrast enhancement

(c) Blurred (d) JPEG compressed

FIGURE 8 The visual information fidelity (VIF) has an interesting feature: it can capture the effects of linear contrast enhancements on images, and quantify the improvement in visual quality. A VIF value greater than unity indicates this improvement, while a VIF value less than unity signifies a loss of visual quality. (A) Reference Lena image (VIF = 1.0). (B) Contrast-stretched Lena image (VIF = 1.17). (C) Gaussian blur (VIF = 0.05) and (D) JPEG-compressed (VIF = 0.05).

the RFs having the same statistics as those obtained from particular realizations.

The source model parameters that need to be estimated from the data consist of the field $\mathcal{S}$. For the vector GSM model, the maximum-likelihood estimate of s_i^2 can be found as follows [17]:

$$\hat{s}_i^2 = \frac{\vec{C}_i^T \mathbf{C}_U^{-1} \vec{C}_i}{M} \tag{19}$$

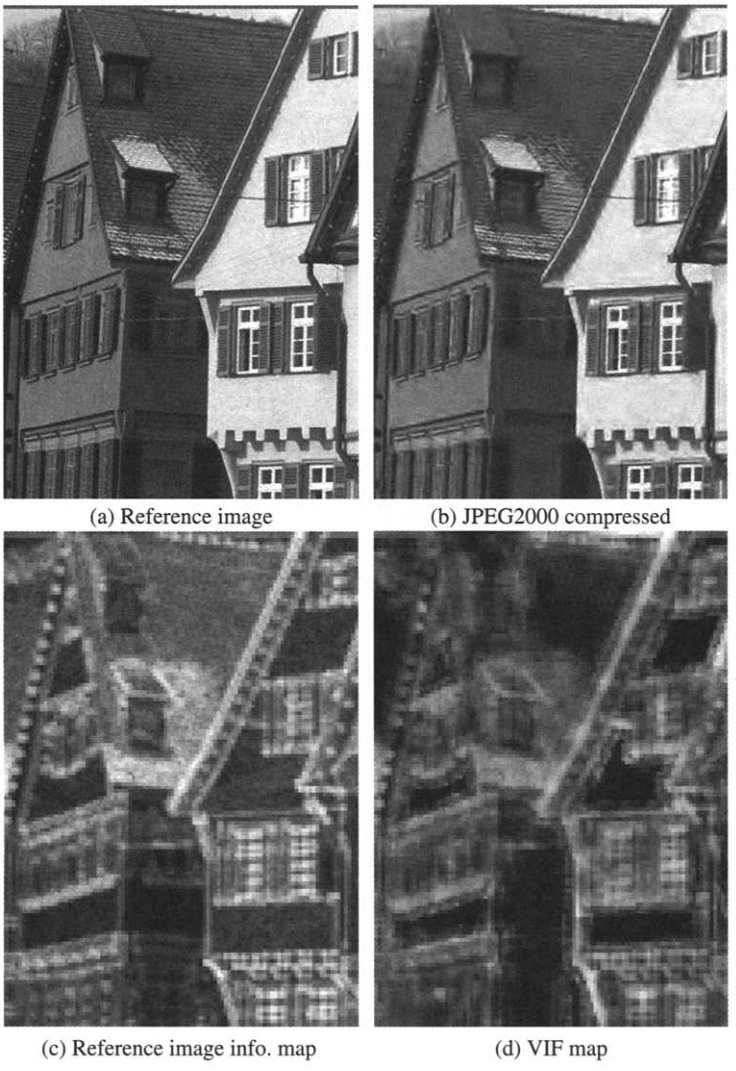
Estimation of the covariance matrix $\mathbf{C}_U$ is also straightforward from the reference image wavelet coefficients [17]:

$$\hat{\mathbf{C}}_U = \frac{1}{N} \sum_{i=1}^{N} \vec{C}_i \vec{C}_i^T \tag{20}$$

In (19) and (20), $\frac{1}{N}\sum_{i=1}^{N} s_i^2$ is assumed to be unity without loss of generality [17].

4.2 Assumptions about the Distortion Model

For the assumptions on the distortion channel to approximately hold, the parameters of the distortion channel are estimated locally. A spatially localized block-window centered at coefficient i could be used to estimate g_i and σ_V^2 at i. The value of the field $\mathcal{G}$ over the block centered at coefficient i, which we denote as g_i, and the variance of the RF $\mathcal{V}$, which we denote as $\sigma_{V,i}^2$, are fairly easy to estimate (by linear regression) since both the input (the reference signal) as



(a) Reference image (b) JPEG2000 compressed

(c) Reference image info. map (d) VIF map

FIGURE 9 Spatial maps showing how visual information fidelity (VIF) captures spatial information loss.

well as the output (the test signal) of the system (2) are available:

$$\hat{g}_i = \widehat{\mathrm{COV}}(C, D)\widehat{\mathrm{COV}}(C, C)^{-1} \tag{21}$$

$$\sigma^2_{V,i} = \widehat{\mathrm{COV}}(D, D) - \hat{g}_i\widehat{\mathrm{COV}}(C, D) \tag{22}$$

where the covariances are approximated by sample estimates using sample points from the corresponding blocks centered at coefficient i in the reference and the test signals.

4.3 Assumptions about the Human Visual System Model

Recall that the HVS model is needed for VIF only, and IFC does not use any HVS modeling. For VIF, the HVS model is parameterized by only one parameter: the variance of visual noise σ^2_N. It is easy to hand-optimize the value of the parameter σ^2_N by running the algorithm over a range of values and observing its performance. While the performance is affected by the choice of σ^2_N, the algorithm's overall performance continues to be highly competitive with other methods for a wide range of values. We will discuss the dependence of the performance of VIF later in Section 5.

5 Results

In this section, we present results on the validation of IFC and VIF and comparisons with PSNR and the well-known Sarnoff model (Sarnoff JND-Metrix 8.0 [8]). The validation is done using subjective quality scores obtained from a group of human observers, and the performance of the QA algorithms is evaluated by comparing the quality predictions of the algorithms against subjective scores. Since all images were not the same size, the IFC in (15) was normalized by the number of pixels to yield information in bits per pixel.

5.1 Subjective Experiments for Validation

For the data used for the results presented in this section, 20 to 28 human subjects were asked to assign each image with a score indicating their assessment of the quality of that image, defined as the extent to which the artifacts were visible and annoying. Twenty-nine high-resolution 24-bits/pixel RGB color images (typically 768×512) were distorted using five distortion types: JPEG2000, JPEG, white noise in the RGB components, Gaussian blur, and transmission errors in the JPEG2000 bit stream using a fast-fading Rayleigh channel model. A database was derived from the 29 images to yield a total of 779 distorted images, which, together with the undistorted images, were then evaluated by human subjects. The raw scores were processed to yield difference mean opinion scores (DMOS) scores for validation and testing.

5.2 Simulation Details

Some additional simulation details are as follows. Although full-color images were distorted in the subjective evaluation, the IFC and VIF algorithms operated on the luminance component only. GSM vectors were constructed from nonoverlapping 3×3 neighborhoods, and the distortion model was estimated with an 18×18 sliding window. Only the subbands at the finest level were used in the summation of (15) and (18).

5.3 Calibration of the Objective Score

It is generally acceptable for a QA method to stably predict subjective quality within a nonlinear mapping, since the mapping can be compensated for easily. Moreover, since the mapping is likely to depend on the subjective validation/application scope and methodology, it is best to leave it to the final application, and not to make it part of the QA algorithm. Logistic functions are typically used for nonlinear fitting in QA applications. For the results presented here, a five-parameter nonlinearity (a logistic function with additive linear term) was used on the logarithm of the IFC and VIF. The mapping function used is given in (23), while the fitting was done using numeric methods.

$$\text{Quality}(x) = \beta_1 \text{logistic}(\beta_2, (x - \beta_3)) + \beta_4 x + \beta_5 \quad (23)$$

$$\text{logistic}(\tau, x) = \frac{1}{2} - \frac{1}{1 + \exp(\tau x)} \quad (24)$$

The fitting of the logistic curve for the methods tested is shown in Fig. 10, while the quality predictions after

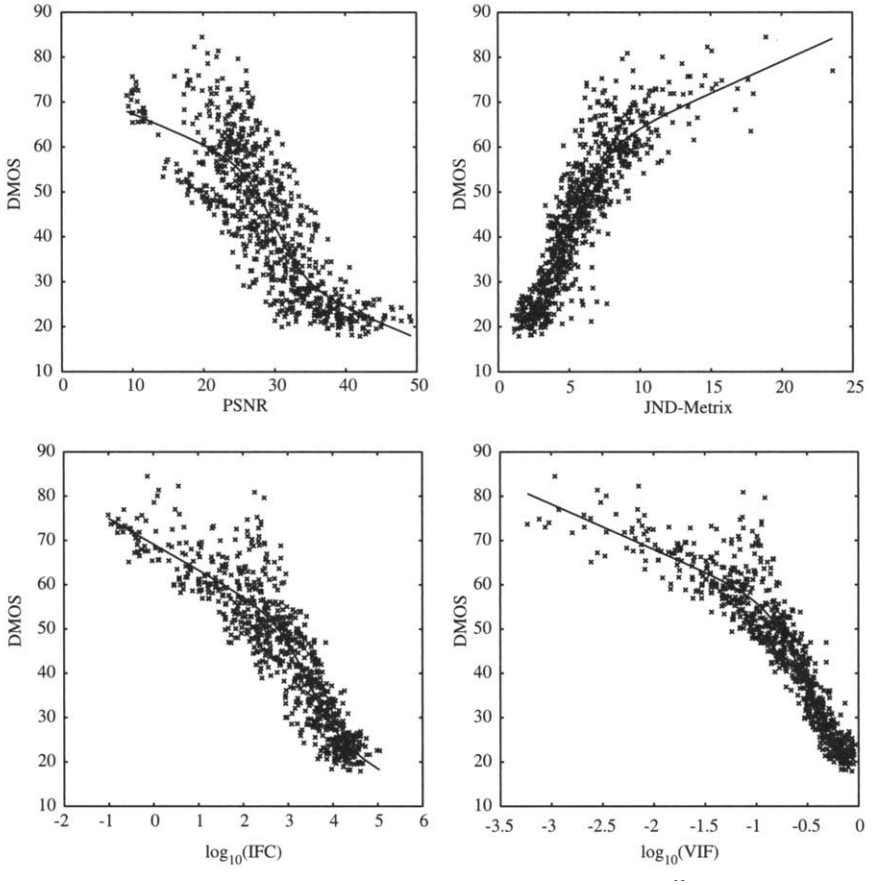


FIGURE 10 Scatter plots for the objective quality criteria: PSNR, Sarnoff's JND-Metrix 8.0 [8], IFC, and VIF.

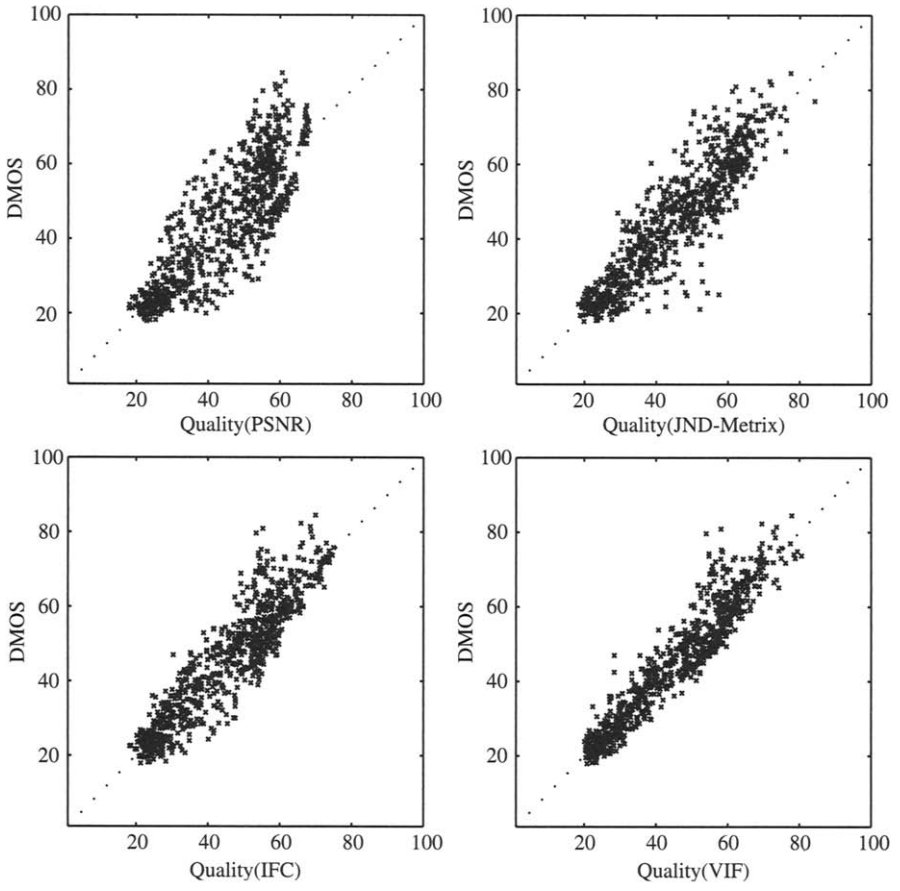


FIGURE 11 Scatter plots for the objective quality criteria: PSNR, Sarnoff's JND-Metrix 8.0 [8], IFC, and VIF after linearization.

TABLE 1 Validation scores for different quality assessment methods

	Validation Against DMOS	
Model	CC	RMS
PSNR	0.826	9.087
JND-Metrix	0.901	6.992
IFC	0.918	6.403
VIF	0.949	5.083

The methods tested were peak signal-to-noise ratio (PSNR), Sarnoff's JND-Metrix 8.0 [8], information fidelity criterion (IFC) and visual information fidelity (VIF). The methods were tested against DMOS from the subjective study after a nonlinear mapping. The validation criteria are linear correlation coefficient (CC) and root mean squared error (RMSE).

compensating for the mapping are shown in Fig. 11. Table 1 quantifies the performance of the various methods in terms of well known validation quantities: the linear correlation coefficient between predicted quality and subjective quality, and the root-mean-squared error (RMSE) between them.

5.4 Discussion

It is evident from Table 1 and Figs. 10 and 11 that the IFC and VIF are both able to predict image quality much more accurately than PSNR and Sarnoff's JND-Metrix 8.0. VIF exhibits a superior performance relative to IFC. However this performance comes at the cost of introducing a model parameter σ_N^2. Figure 12 shows how the performance of the VIF varies with σ_N^2. Note that while the performance of VIF varies with σ_N^2, VIF is superior to PSNR for all range of values shown in Fig. 12.

6 Conclusions and Future Work

In this chapter we presented a new framework for doing full-reference image quality assessment based on information fidelity, which is an information theoretic setup using natural scene statistics. We explored the relationship between image information and visual quality, and presented two methods for full-reference image quality assessment. The information fidelity criterion (IFC) quantified the information shared between the reference and the distorted images, while

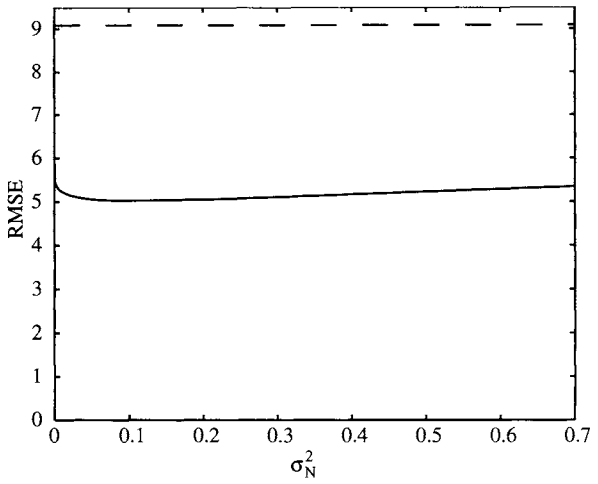


FIGURE 12 Dependence of visual information fidelity (VIF) performance on the σ_N^2 parameter. The performance of VIF is relatively stable to changes in the value of σ_N^2. VIF (solid) and PSNR (dashed).

the visual information fidelity (VIF) measure used the reference image information as well to quantify relative information fidelity. The IFC and VIF were derived from a statistical model for natural scenes, a model for image distortions, and (for VIF) a human visual system model in an information–theoretic setting. Both the VIF and the IFC are capable of accurately predicting image quality.

It would be interesting to see possible extensions of the concepts presented in this chapter for video quality assessment using spatiotemporal natural scene models and other properties of natural scenes, such as color and local phase. This new paradigm has the potential to give new understanding into the relationship between image information and visual perception of quality.

References

[1] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans. Image Process.* 8, 1688–1701 (1999).

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 1991).

[3] M. P. Eckert and A. P. Bradley. "Perceptual quality metrics applied to still image compression." *Signal Processing*, 70, 177–200 (1998).

[4] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am.* 4, 2379–2394 (1987).

[5] J. L. Mannos and D. J. Sakrison. "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Inform. Theory*, 4, 525–536 (1974).

[6] J. Portilla, M. Wainwright, V. Strela, and E. P. Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Trans. Image Process.* 12, 1338–1351 (2003).

[7] D. L. Ruderman, "The statistics of natural images," *Network: Computation in Neural Sys.* 5, 517–548 (1994).

[8] Sarnoff Corporation, "JND-Metrix Technology," Evaluation version available at: *http://www.sarnoff.com/products_services/video_vision/jndmetrix/downloads.asp*, 2003.

[9] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.* (2003). 2006 to appear.

[10] H. R. Sheikh and A. C. Bovik, "Image information and visual quality", In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (2004).

[11] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.* (2004) 2006 to appear.

[12] E. P. Simoncelli, "Statistical models for images: compression, restoration and synthesis", *In Proc. IEEE Asilomar Conf. Sign. Sys. Comput.* (1997).

[13] E. P. Simoncelli, "Modeling the joint statistics of images in the wavelet domain," *Proc. SPIE*, 3813, 188–195 (1999).

[14] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: a flexible architecture for multi-scale derivative computation," *Proc. IEEE Int. Conf. Image Proc.* 444–447 (1995).

[15] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, Shiftable multi-scale transforms, *IEEE Trans. Inform. Theory*, 38, 587–607 (1992).

[16] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Ann. Rev. Neurosci.*, 24, 1193–216 (2001).

[17] V. Strela, J. Portilla, and E. Simoncelli, "Image denoising using a local Gaussian Scale Mixture model in the wavelet domain," *Proc. SPIE* 4119, 363–371 (2000).

[18] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," *Adv. Neural Inform. Process. Syst.* 12, 855–861 (2000).

[19] M. J. Wainwright, E. P. Simoncelli, and A. S. Wilsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," *Appl. Comput. Harmonic Anal.* 11, 89–123 (2001).

[20] B. A. Wandell, *Foundations of Vision* (Sinauer Associates, Sunderland, MA, 1995).

[21] Z. Wang, A. C. Bovik, and L. Lu. "Why is image quality assessment so difficult?" In *Proc. IEEE Int. Conf. Acoust., Speech Sign. Process.* (Orlando, May 2002).

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