Image Information and Visual Quality

Hamid Rahim Sheikh, *Member, IEEE,* Alan C. Bovik, *Fellow, IEEE,*

Abstract

Measurement of visual quality is of fundamental importance to numerous image and video processing applications. The goal of quality assessment (QA) research is to design algorithms that can automatically assess the quality of images or videos in a perceptually consistent manner. Image QA algorithms generally interpret image quality as fidelity or similarity with a 'reference' or 'perfect' image in some perceptual space. Such 'Full-Reference' QA methods attempt to achieve consistency in quality prediction by modeling salient physiological and psychovisual features of the human visual system (HVS), or by signal fidelity measures. In this paper we approach the image QA problem as an information fidelity problem. Specifically, we propose to quantify the loss of *image information* to the distortion process, and explore the relationship between image information and visual quality. QA systems are invariably involved with judging the visual quality of 'natural' images and videos that are meant for 'human consumption.' Researchers have developed sophisticated models to capture the statistics of such natural signals. Using these models, we previously presented an information fidelity criterion for image quality assessment that related image quality with the amount of information shared between a reference and a distorted image [1]. In this paper, we propose an image information measure that quantifies the information that is present in the reference image, and also quantify how much of this reference information can be extracted from the distorted image. Combining these two quantities, we propose a *visual information fidelity* measure for image quality assessment. We validate the performance of our algorithm with an extensive subjective study involving 779 images, and show that our method outperforms recent state-of-the-art image quality assessment algorithms by a sizeable margin in our simulations. The code and the data from the subjective study are available at [2].

Index Terms

Image Quality Assessment, Natural Scene Statistics, Image Information, Information Fidelity.

I. Introduction

The field of digital image and video processing deals, in large part, with signals that are meant to convey reproductions of visual information for human consumption, and many image and video processing systems, such as those for acquisition, compression, restoration, enhancement and reproduction etc., operate solely on these visual

H. R. Sheikh was previously affiliated with the Laboratory for Image and Video Engineering, Department of Electrical & Computer Engineering, The University of Texas at Austin, Austin, USA. He is currently affiliated with Texas Instruments Inc., Dallas, TX, USA. Phone: (214) 480-3186, email: hamid.sheikh@ieee.org

A. C. Bovik is affiliated with the Department of Electrical & Computer Engineering, The University of Texas at Austin, Austin, TX 78712-1084USA, Phone: (512) 471-5370, email:bovik@ece.utexas.edu

This work was supported by a grant from the National Science Foundation.

reproductions. These systems typically involve tradeoffs between resources and the visual quality of the output. In order to make these tradeoffs we need a way of measuring the quality of images or videos that come from a system running under a given configuration. The obvious way of measuring quality is to solicit the opinion of human observers. However, such subjective evaluations are not only cumbersome and expensive, but they also cannot be incorporated into automatic systems that adjust themselves in real-time based on the feedback of output quality. The goal of quality assessment research is, therefore, to design algorithms for *objective* evaluation of quality in a way that is consistent with subjective human evaluation. Such QA methods would prove invaluable for testing, optimizing, bench-marking, and monitoring applications.

Traditionally, researchers have focussed on measuring signal fidelity as a means of assessing visual quality. Signal fidelity is measured with respect to a reference signal that is assumed to have 'perfect' quality. During the design or evaluation of a system, the reference signal is typically processed to yield a distorted (or test) image, which can then be compared against the reference using so-called *full reference* (FR) QA methods. Typically this comparison involves measuring the 'distance' between the two signals in a perceptually meaningful way. This paper presents a FR QA method for images.

A simple and widely used fidelity measure is the Peak Signal to Noise Ratio (PSNR), or the corresponding distortion metric, the Mean Squared Error (MSE). The MSE is the L_2 norm of the arithmetic difference between the reference and the test signals. It is an attractive measure for the (loss of) image quality due to its simplicity and mathematical convenience. However, the correlation between MSE/PSNR and human judgement of quality is not tight enough for most applications, and the goal of QA research over the past three decades has been to improve upon the PSNR.

For FR QA methods, modeling of the human visual system has been regarded as the most suitable paradigm for achieving better quality predictions. The underlying premise is that the sensitivities of the visual system are different for different aspects of the visual signal that it perceives, such as brightness, contrast, frequency content, and the interaction between different signal components, and it makes sense to compute the strength of the error between the test and the reference signals once the different sensitivities of the HVS have been accurately accounted for. Other researchers have explored signal fidelity criteria that are not based on assumptions about HVS models, but are motivated instead by the need to capture the loss of *structure* in the signal, structure that the HVS hypothetically extracts for cognitive understanding.

In [1], we presented a novel information theoretic criterion for image fidelity measurement that was based on natural scene statistics (NSS). Images and videos of the three dimensional visual environment come from a common class: the class of natural scenes. Natural scenes form a tiny subspace in the space of all possible signals, and researchers have developed sophisticated models to characterize these statistics. Most real-world distortion processes disturb these statistics and make the image or video signals *unnatural*. In [1], we proposed using NSS models in conjunction with a distortion (channel) model to quantify the information shared between the test and the reference images, and showed that this shared information is an aspect of fidelity that relates well with visual quality. In contrast to the HVS error-sensitivity and the structural approaches, the *statistical* approach, used in

an information-theoretic setting, yielded an FR QA method that did not rely on any HVS or viewing geometry parameter, nor any constant requiring optimization, and yet was competitive with state of the art QA methods.

In this paper we extend the concept of information fidelity measurement for image quality assessment by proposing an image information measure. This measure quantifies the information that could ideally be extracted by the brain from the reference image. We then quantify the loss of this information to the distortion using NSS, HVS and an image distortion (channel) model in an information-theoretic framework. We demonstrate that visual quality of images is strongly related to relative image information present in the distorted image, and that this approach outperforms state-of-the-art quality assessment algorithms by a sizeable margin in our simulations. Another salient feature of our algorithm is that it is characterized by only one HVS parameter that is easy to train and optimize for improved performance.

Section II presents some background work in the field of FR QA algorithms as well as an introduction to natural scene statistics models. Section III presents our development of the image information measure and the proposed visual information fidelity criterion. Implementation and subjective validation details are provided in Sections IV and V, while the results are discussed in Section VI. We conclude the paper in Section VII.

II. Background

Full reference quality assessment techniques proposed in the literature can be divided into two major groups: those based on the HVS and those based on arbitrary signal fidelity criteria. (A detailed review of the research on FR QA methods can be found in [3], [4], [5], [6]).

A. HVS Error Based QA methods

HVS based QA methods come in different flavors based on tradeoffs between accuracy in modeling the HVS and computational feasibility. A detailed discussion of these methods can be found in [4], [5], [6]. A number of HVS based methods have been proposed in the literature. Some representative methods include [7], [8], [9], [10], [11], [12], [13], [14].

B. Arbitrary Signal Fidelity Criteria

Researchers have also attempted to use arbitrary signal fidelity criteria in a hope that they would correlate well with perceptual quality. In [15] and [16], a number of these were evaluated for the purpose of quality assessment. In [17] a *structural similarity metric* (SSIM) was proposed to capture the loss of image structure. SSIM was derived by considering hypothetically what constitutes a loss in signal structure. It was hypothesized that distortions in an image that come from variations in lighting, such as contrast or brightness changes, are non-structural distortions, and that these should be treated differently from structural ones, and that one could capture image quality with three aspects of information loss that are complementary to each other: correlation distortion, contrast distortion, and luminance distortion.

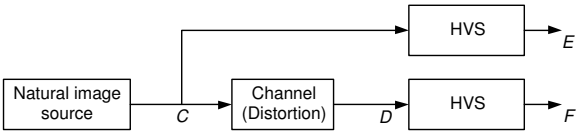


Fig. 1. Mutual information between $\mathcal{C}$ and $\mathcal{E}$ quantifies the information that the brain could ideally extract from the reference image, whereas the mutual information between $\mathcal{C}$ and $\mathcal{F}$ quantifies the corresponding information that could be extracted from the test image.

C. Limitations

A number of limitations of HVS based methods are discussed in [17]. In summary, these have to do with the extrapolation of the vision models that have been proposed in the visual psychology literature to image processing problems. In [17], it was claimed that structural QA methods avoid some of the limitations of HVS based methods since they are not based on threshold psychophysics or the HVS models derived thereof. However they have some limitations of their own. Specifically, although the structural paradigm for QA is an ambitious paradigm, there is no widely accepted way of defining structure and structural distortion in a perceptually meaningful manner. Most structural methods are constructed by *hypothesizing* the functional forms of structural and non-structural distortions and the interaction between them.

In [1], we proposed a new approach to the quality assessment problem where we quantified the information that was shared between the test and the reference images, and demonstrated that this quantification relates well with visual quality. In this paper we further explore the connections between image information and visual quality. Specifically, we will model the reference image as being the output of a stochastic 'natural' source that passes through the HVS channel and is processed later by the brain. We quantify the information content of the reference image as being the mutual information between the input and output of the HVS channel. This is the information that the brain could ideally extract from the output of the HVS. We then quantify the same measure in the presence of an image distortion channel that distorts the output of the natural source before it passes through the HVS channel, thereby measuring the information that the brain could ideally extract from the test image. This is shown pictorially in Figure 1. We then combine the two information measures to form a visual information fidelity measure that relates visual quality to *relative* image information [18].

D. Natural Scene Statistics

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 non-visual stimuli such as Radar and Sonar, X-Rays, ultra-sounds etc. Natural scenes form an extremely tiny subset of the set of all possible images. Many researchers have attempted to understand the structure of this subspace of natural images by studying their statistics (a review on natural scene models could be found in [19]). Researchers believe that the visual stimulus emanating from the natural environment drove the evolution of the HVS, and that

modeling natural scenes and the HVS are essentially dual problems [20]. While many aspects of the HVS have been studied and incorporated into quality assessment algorithms, a usefully comprehensive (and feasible) understanding is still lacking. NSS modeling may serve to fill this gap.

Natural scene statistics have been explicitly incorporated into a number of image processing algorithms: in compression algorithms [21], [22], [23], [24], denoising algorithms [25], [26], [27], image modeling[28], image segmentation [29], and texture analysis and synthesis [30]. While the characteristics of the distortion processes have been incorporated into some quality assessment algorithms (such as those designed for the blocking artifact), the assumptions about the statistics of the images that they afflict are usually quite simplistic. Specifically, most QA algorithms assume that the input images are smooth and low-pass in nature. In [31], an NSS model was used to design a no-reference image quality assessment method for images distorted with the JPEG2000 compression artifacts. In this paper we use NSS models for FR QA, and model natural images in the wavelet domain using Gaussian Scale Mixtures (GSM) [27]. Scale-space-orientation analysis (loosely referred to as wavelet analysis in this paper) of images has been found to be useful for natural image modeling. It is well known that the coefficients of a subband in a wavelet decomposition are neither independent nor identically distributed, though they may be approximately second-order uncorrelated [32]. A coefficient is likely to have a large variance if its neighborhood has a large variance. The marginal densities are sharply peaked around zero with heavy tails, which are typically modeled as Laplacian density functions, while the localized statistics are highly space-varying. Researchers have characterized this behavior of natural images in the wavelet domain by using GSMs [27], a more detailed introduction to which will be given in the next section.

III. Visual Information Fidelity for Image Quality Assessment

Natural images of perfect quality can be modeled as the output of a stochastic source. In the absence of any distortions, this signal passes through the HVS channel of a human observer before entering the brain, which extracts cognitive information from it. For distorted images, we assume that the reference signal has passed through another 'distortion channel' before entering the HVS. This is shown pictorially in Figure 1. The visual information fidelity (VIF) measure that we propose in this paper is derived from a quantification of two mutual information quantities: the mutual information between the input and the output of the HVS channel when no distortion channel is present (we call this the *reference image information*) and the mutual information between the input of the distortion channel and the output of the HVS channel for the test image. We discuss the components of the proposed method in this section.

A. The Source Model

As mentioned in Section II-D, the NSS model that we use is the GSM model in the wavelet domain. It is convenient to deal with one subband of the wavelet decomposition at this point and later generalize this for multiple subbands. A GSM is a random field (RF) that can be expressed as a product of two independent RFs [27]. That is, a GSM

$\mathcal{C} = \{\overrightarrow{C}_i : \ i \in \mathrm{I}\}$, where I denotes the set of spatial indices for the RF, can be expressed as:

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

where $\mathcal{S} = \{S_i : \ i \in \mathrm{I}\}$ is an RF of positive scalars and $\mathcal{U} = \{\overrightarrow{U}_i : \ i \in \mathrm{I}\}$ is a Gaussian vector RF with mean zero and covariance $\mathbf{C}_U$. $\overrightarrow{C}_i$ and $\overrightarrow{U}_i$ are M dimensional vectors, and we assume that for the RF $\mathcal{U}$, $\overrightarrow{U}_i$ is independent of $\overrightarrow{U}_j$, $\forall i \neq j$,. In this paper we model each subband of a scale-space-orientation wavelet decomposition (such as the steerable pyramid [33]) of an image as a GSM RF. We partition the subband coefficients into non-overlapping blocks of M coefficients each, and model block i as the vector $\overrightarrow{C}_i$.

One could easily make the following observations regarding the above model: $\mathcal{C}$ is normally distributed given $\mathcal{S}$ (with mean zero and covariance $S_i^2\mathbf{C}_U$), that given S_i, $\overrightarrow{C}_i$ are independent of S_j for all $j \neq i$, and that given $\mathcal{S}$, $\overrightarrow{C}_i$ are conditionally independent of $\overrightarrow{C}_j$, $\forall i \neq j$ [27]. The GSM model has been shown to capture key statistical features of natural images. In particular, researchers have shown that 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 that the non-linear dependencies between the wavelet coefficients of natural images can be captured by modeling the field $\mathcal{S}$ as being highly self-correlated [27], [34].

B. 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:

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

where $\mathcal{C}$ denotes the RF from a subband in the reference signal, $\mathcal{D} = \{\overrightarrow{D}_i : \ i \in \mathrm{I}\}$ 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 $\mathcal{V} = \{\overrightarrow{V}_i : \ i \in \mathrm{I}\}$ 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 is 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 blur and additive noise. 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 components separately. Figures 2 and 3 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 (Figure 2(e)), the model fails to reproduce the perceptual annoyance adequately (Figure 2(f)), but it still performs much better than the additive-only noise model shown in Figure 4(f).

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.

C. The Human Visual System Model

The HVS model that we use is also described in the wavelet domain. Since HVS models are the dual of NSS models [20], 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 [1]. 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 paper 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 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 importantly 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 $\mathcal{N} = \{\overrightarrow{N}_i : \ i \in \mathrm{I}\}$ and $\mathcal{N}' = \{\overrightarrow{N'}_i : \ i \in \mathrm{I}\}$, where $\overrightarrow{N}_i$ and $\overrightarrow{N'}_i$ are zero-mean uncorrelated multivariate Gaussian with the same dimensionality as $\overrightarrow{C}_i$:

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

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

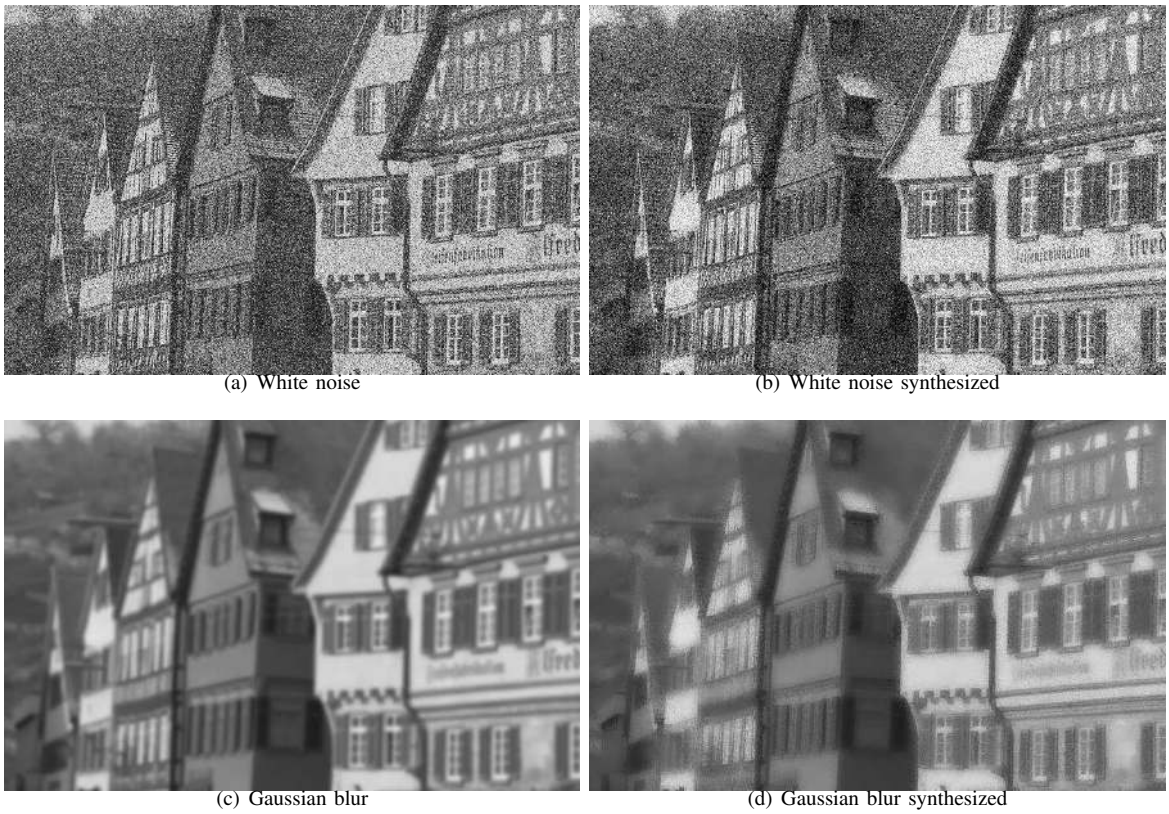
where $\mathcal{E}$ and $\mathcal{F}$ denote the visual signal at the output of the HVS model from the reference and the test images in one subband respectively, from which the brain extracts cognitive information (Figure 1). The RFs $\mathcal{N}$ and $\mathcal{N}'$ are

(a) JPEG2000 (b) JPEG2000 synthesized

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

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

Fig. 2. 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 has JPEG at low bit rates, the model's performance worsens, but is still better than the additive-only noise model of Figure 4.



(a) White noise (b) White noise synthesized

(c) Gaussian blur (d) Gaussian blur synthesized

Fig. 3. Distorted images and their synthesized versions for the attenuation/additive noise distortion model.

assumed to be independent of $\mathcal{U}$, $\mathcal{S}$, and $\mathcal{V}$. We model the covariance of $\mathcal{N}$ 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).

D. The Visual Information Fidelity Criterion

With the source, distortion, and HVS models as described above, the visual information fidelity criterion that we propose can be derived. Let $\overrightarrow{C}^N = (\overrightarrow{C}_1, \overrightarrow{C}_2, \ldots, \overrightarrow{C}_N)$ denote N elements from $\mathcal{C}$. Let S^N, $\overrightarrow{D}^N$, $\overrightarrow{E}^N$ and $\overrightarrow{F}^N$ be correspondingly defined. In this section we will assume that the model parameters $\mathcal{G}$, σ_v^2 and σ_n^2 are known.

The mutual information $I(\overrightarrow{C}^N; \overrightarrow{E}^N)$ quantifies the amount of information that can be extracted from the output of the HVS by the brain when the test image is being viewed. 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 $I(\overrightarrow{C}^N; \overrightarrow{D}^N | S^N = s^N)$ instead of $I(\overrightarrow{C}^N; \overrightarrow{D}^N)$, where s^N denotes a realization of S^N for a particular

[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.

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 visual neurons [1], and lends the VIF to analytical tractability as well.

For the reference image, we can analyze $I(\overrightarrow{C}^N; \overrightarrow{E}^N | S^N = s^N)$, where s^N denotes a *realization* of S^N. In this paper we will denote $I(\overrightarrow{C}^N; \overrightarrow{E}^N | \overrightarrow{S}^N = s^N)$ as $I(\overrightarrow{C}^N; \overrightarrow{E}^N | s^N)$. With the stated assumptions on $\mathcal{C}$ and the distortion model (2), we get:

$$\begin{aligned} I(\overrightarrow{C}^N; \overrightarrow{E}^N | s^N) &= \sum_{j=1}^{N} \sum_{i=1}^{N} I(\overrightarrow{C}_i; \overrightarrow{E}_j | \overrightarrow{C}^{i-1}, \overrightarrow{E}^{j-1}, s^N) && (6) \\ &= \sum_{i=1}^{N} I(\overrightarrow{C}_i; \overrightarrow{E}_i | s_i) && (7) \\ &= \sum_{i=1}^{N} (h(\overrightarrow{C}_i + \overrightarrow{N}_i | s_i) - h(\overrightarrow{N}_i | s_i)) && (8) \\ &= \frac{1}{2} \sum_{i=1}^{N} \log_2 \left(\frac{|s_i^2 \mathbf{C}_U + \sigma_n^2 \mathbf{I}|}{|\sigma_n^2 \mathbf{I}|} \right) && (9) \end{aligned}$$

where we get (6) from chain rule [35], and (7) from the conditional independence of $\mathcal{C}$ and $\mathcal{N}$ given $\mathcal{S}$, and $|.|$ denotes the determinant. Similarly we can show that for the test image

$$\begin{aligned} & I(\overrightarrow{C}^N; \overrightarrow{F}^N | s^N) \\ &= \sum_{i=1}^{N} (h(g_i \overrightarrow{C}_i + \overrightarrow{V}_i + \overrightarrow{N}_i | s_i) - h(\overrightarrow{V}_i + \overrightarrow{N}_i | s_i)) && (10) \\ &= \frac{1}{2} \sum_{i=1}^{N} \log_2 \left(\frac{|g_i^2 s_i^2 \mathbf{C}_U + (\sigma_v^2 + \sigma_n^2) \mathbf{I}|}{|(\sigma_v^2 + \sigma_n^2) \mathbf{I}|} \right) && (11) \end{aligned}$$

Since $\mathbf{C}_U$ is symmetric, it can be factored as $\mathbf{C}_U = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^{\mathbf{T}}$, where $\mathbf{Q}$ is an orthonormal matrix, and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues λ_k. One can use this matrix factorization to show:

$$\begin{aligned} I(\overrightarrow{C}^N; \overrightarrow{E}^N | s^N) &= \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{M} \log_2 \left(1 + \frac{s_i^2 \lambda_k}{\sigma_n^2} \right) && (12) \\ I(\overrightarrow{C}^N; \overrightarrow{F}^N | s^N) &= \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{M} \log_2 \left(1 + \frac{g_i^2 s_i^2 \lambda_k}{\sigma_v^2 + \sigma_n^2} \right) && (13) \end{aligned}$$

$I(\overrightarrow{C}^N; \overrightarrow{E}^N | s^N)$ and $I(\overrightarrow{C}^N; \overrightarrow{F}^N | s^N)$ represent the information that could ideally be extracted by the brain from a particular subband in the reference and the test images respectively. We call $I(\overrightarrow{C}^N; \overrightarrow{E}^N | s^N)$ 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 information that could be extracted from the test image is 2.0 bits per pixel, and if the information that could be extracted from the corresponding reference image is 2.1 bits per pixel, then the brain can recover most of the information content of the reference image from the test image. By contrast, if the corresponding reference image information were, say, 5.0 bits per pixel, then we have lost 3.0 bits of information to the distortion channel, and the visual quality of the test image should be inferior.

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

Also we have only dealt with one subband so far. 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 VIF that we propose in this paper is given by:

$$\text{VIF} \quad = \quad \frac{\sum_{j \in \text{subbands}} I(\overrightarrow{C}^{N,j}; \overrightarrow{F}^{N,j} | s^{N,j})}{\sum_{j \in \text{subbands}} I(\overrightarrow{C}^{N,j}; \overrightarrow{E}^{N,j} | s^{N,j})} \tag{14}$$

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

The VIF given in (14) is computed for a collection of $N \times M$ wavelet coefficients from each subband that could either represent an entire subband of an image, or a spatially localized region of subband coefficients. In the former case, the VIF is one 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.

E. Properties of VIF

The VIF has a number of interesting features. Firstly, note that VIF is bounded below by zero (such as when $I(\overrightarrow{C}^N; \overrightarrow{F}^N | s^N) = 0$ and $I(\overrightarrow{C}^N; \overrightarrow{E}^N | s^N) \neq 0$), which indicates that all information about the reference image has been lost in the distortion channel. Secondly, in case the image is not distorted at all, and VIF is calculated between the reference image and its copy, VIF is *exactly* unity. This is because $g_i = 1 \ \forall i$, and $\sigma_v^2 = 0$, and therefore $I(\overrightarrow{C}^N; \overrightarrow{F}^N | s^N) = I(\overrightarrow{C}^N; \overrightarrow{E}^N | s^N)$. Thus for all practical distortion types, VIF will lie in the interval $[0, 1]$. Thirdly, and this is where we feel that VIF has a distinction over traditional quality assessment methods, a linear contrast enhancement of the reference image that does not add noise to it will result in a VIF value *larger* than unity, thereby signifying that the enhanced image has a *superior* visual quality than the reference image! It is common observation that contrast enhancement of images increases their perceptual quality unless quantization, clipping, or display non-linearities add additional distortion. Theoretically, contrast enhancement results in a higher signal-to-noise ratio at the output of the HVS neurons, thereby allowing the brain to have a greater ability to discriminate objects present in the visual signal. The VIF is able to capture this improvement in visual quality.

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 increases 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 point-wise contrast improvement on a computer is therefore more an artifact of limited machine precision and display nonlinearities.

To the best of our knowledge, no other quality assessment algorithm has the ability to predict if the visual image quality has been enhanced by a contrast enhancement operation. We envision extending the notion of quantifying improvement in visual quality of images by image enhancement operations using a similar information-theoretic paradigm.

It is interesting to see a few test cases that illustrate these properties of VIF visually. The implementation details of VIF are given in the next section; here we only wish to illustrate the above discussion pictorially. Figure 6 shows a reference image that has been distorted with 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 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 captures the 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 for both.

Figure 7 illustrates the behavior of VIF with spatial quality maps. Figure 7(a) shows a reference image and Figure 7(b) the corresponding JPEG2000 compressed image. Note that the distortions are clearly visible. Figure 7(c) shows the reference image information map in the same location. The information map shows the spread of statistical information in the reference image. In flat image regions, the information content of the image is low, whereas in textured regions and regions containing strong edges, the image information is high. The quality map in Figure 7(d) shows the proportion of the image information that has been lost to JPEG2000 compression.

F. Similarities of VIF with HVS Based Methods

It was shown previously that the information fidelity criterion (IFC) presented in [1] is functionally equivalent to HVS based methods under certain conditions. For VIF, the numerator in (14) is basically IFC (apart from the visual noise source) and hence is functionally similar to HVS based methods as discussed in detail in [1]. We feel that the normalization by reference image information in (14) can be thought of as being a *content dependent adjustment* of HVS based methods. Specifically, after the HVS based methods compute the perceptual error strength, the annoyance factor of a particular perceptual error strength may be different for different images, and thus may give a different impression of quality. We feel that the normalization by reference image information adjusts for this variation in image content.

IV. Implementation Issues

In order to implement VIF criterion in (14) a number of assumptions are needed about the source, distortion, and HVS models. We outline them in this section.

A. Assumptions about the source model

Note that mutual information (and hence VIF) can only be calculated between RF's and not their *realizations*, that is, a particular reference and the test image under consideration. We will assume ergodicity of the RF's and that reasonable estimates for the statistics of the RF's can be obtained from their realizations. We then quantify the mutual information between the RF's 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 [36]:

$$\widehat{s}_i^2 = \frac{\overrightarrow{C}_i^T \mathbf{C}_U^{-1} \overrightarrow{C}_i}{M} \tag{15}$$

Estimation of the covariance matrix $\mathbf{C}_U$ is also straightforward from the reference image wavelet coefficients [36]:

$$\widehat{\mathbf{C}}_U = \frac{1}{N} \sum_{i=1}^{N} \overrightarrow{C}_i \overrightarrow{C}_i^T \tag{16}$$

In (15) and (16), $\frac{1}{N}\sum_{i=1}^{N} s_i^2$ is assumed to be unity without loss of generality [36].

B. Assumptions about the distortion model

In order for the assumptions on the distortion operator to hold, we estimate the parameters of the distortion channel *locally*. Hence we will use a $B \times B$ window centered at coefficient i 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 well as the output (the test signal) of the system (2) are available:

$$\widehat{g}_i = \widehat{\mathrm{Cov}}(C,D)\widehat{\mathrm{Cov}}(C,C)^{-1} \tag{17}$$

$$\widehat{\sigma}_{v,i}^2 = \widehat{\mathrm{Cov}}(D,D) - \widehat{g}_i \widehat{\mathrm{Cov}}(C,D) \tag{18}$$

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.

C. Assumptions about the HVS model

The HVS model is parameterized by only one parameter: the variance of visual noise σ_n^2. It is easy to hand-optimize the value of the parameter σ_n^2 by running the algorithm over a range of values and observing its performance. While the performance is affected by the choice of σ_n^2, the algorithm's overall performance continues to be highly competitive with other methods for a wide range of values.

Further specifics of the estimation methods used in our testing are given in Section VI.

V. Subjective Experiments for Validation

In order to calibrate and test the algorithm, an extensive psychometric study was conducted. In these experiments, a number of 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 (FF) channel model. A database was derived from the 29 images such that each image had test versions with each distortion type, and for each distortion type the perceptual quality roughly covered the entire quality range. Observers were asked to provide their perception of quality on a continuous linear scale that was divided into five equal regions marked with adjectives "Bad", "Poor", "Fair", "Good" and "Excellent". About 20-25 human observers rated each image. Each distortion type was evaluated by different subjects in different experiments using the same equipment and viewing conditions. In this way a total of 982 images, out of which 203 were the reference images, were evaluated by human subjects in seven experiments. The raw scores were converted to difference scores (between the test and the reference) [37] and then converted to Z-scores [38], scaled back to $1-100$ range, and finally a Difference Mean Opinion Score (DMOS) for each distorted image. The average RMSE for the DMOS was 5.92 with an average 95% confidence interval of width 5.48. The database is available at [2].

VI. Results

In this section we present results on validation of VIF on the database presented in Section V, and present comparisons with other quality assessment algorithms. Specifically, we compare the performance of VIF against PSNR, SSIM [17], and the well known Sarnoff model (Sarnoff JND-Metrix 8.0 [39]). We present results for two versions of VIF: VIF using the finest resolution at all orientations, and using the horizontal and vertical orientations only. Table I summarizes the results for the quality assessment methods, which are discussed in Section VI-C.

A. Simulation Details

Some additional simulation details are as follows. Although full color images were distorted in the subjective evaluation, the QA algorithms (except Sarnoff's) operated upon the luminance component only. GSM vectors were constructed from non-overlapping 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 (14). MSSIM (Mean SSIM) was calculated on the luminance component after decimating (filtering and downsampling) it by a factor of 4 (see [17]).

B. Calibration of the Objective Score

It is generally acceptable for a QA method to stably predict subjective quality within a non-linear mapping, since the mapping can be compensated for easily. Moreover, since the mapping is likely to depend upon the subjective validation/application scope and methodology, it is best to leave it to the final application, and not to make it part

Validation against DMOS					
Model	CC	MAE	RMS	OR	SROCC
PSNR	0.826	7.272	9.087	0.114	0.820
Sarnoff	0.901	5.252	6.992	0.046	0.902
MSSIM	0.912	4.980	6.616	0.035	0.910
VIF	0.949	3.878	5.083	0.013	0.949
VIF (hv)	0.950	3.820	5.025	0.013	0.950

TABLE I

VALIDATION SCORES FOR DIFFERENT QUALITY ASSESSMENT METHODS. THE METHODS TESTED WERE PSNR, SARNOFF JND-METRIX 8.0 [39], MSSIM [17], VIF, AND VIF USING HORIZONTAL AND VERTICAL ORIENTATIONS ONLY. THE METHODS WERE TESTED AGAINST DMOS FROM THE SUBJECTIVE STUDY AFTER A NON-LINEAR MAPPING. THE VALIDATION CRITERIA ARE: CORRELATION COEFFICIENT (CC), MEAN ABSOLUTE ERROR (MAE), ROOT MEAN SQUARED ERROR (RMS), OUTLIER RATIO (OR) AND SPEARMAN RANK-ORDER CORRELATION COEFFICIENT (SROCC).

of the QA algorithm. Thus, in both the VQEG Phase-I and Phase-II testing and validation, a non-linear mapping between the objective and the subjective scores was allowed, and all the performance validation metrics were computed *after* compensating for it [37]. This is true for the results in Table I, where a five-parameter non-linearity (a logistic function with additive linear term constrained to be monotonic) is used for all methods except for VIF, for which we used the mapping on the logarithm of VIF. The fitting of the logistic curve to some of the methods tested is shown in Figure 8, while the quality predictions after compensating for the mapping are shown in Figure 9. The mapping function used is given in (19), while the fitting was done using MATLAB's *fminunc*.

$$\text{Quality}(x) = \beta_1 \text{logistic}\left(\beta_2, (x - \beta_3)\right) + \beta_4 x + \beta_5 \tag{19}$$

$$\text{logistic}(\tau, x) = \frac{1}{2} - \frac{1}{1 + \exp(\tau x)} \tag{20}$$

C. Discussion

1) Overall performance: Table I shows that VIF is competitive with all state-of-the-art FR QA methods presented in this paper and outperforms them in our simulations by a sizeable margin. Also note that VIF and MSSIM use only the luminance components of the images to make quality predictions, whereas the JND-Metrix uses all color information. Extending VIF to incorporate color could further improve performance.

As noted in [1], the performance of VIF improves slightly when only the horizontal and vertical orientations are used in the summation in (14), although the improvement is less marked than in [1]. Nevertheless, the reduced computational complexity makes this a much more attractive implementation option.

2) Cross-distortion performance: It is interesting to study the performance of VIF on specific distortion types. Many image QA methods perform well on single distortion types, but their limitations show up on a broader validation study involving different distortion types. Nevertheless, it is sometimes interesting from an application

RMSE performance on specific distortions.				
Distortion	PSNR	Sarnoff	MSSIM	VIF
JPEG2000	7.187	5.028	4.693	4.745
JPEG	8.173	5.451	5.511	5.309
White noise	2.588	3.967	2.709	2.494
Gaussian blur	9.774	5.104	5.159	3.399
FF	7.517	6.713	6.990	3.921

TABLE II

RMSE PERFORMANCE OF THE QA METHODS ON INDIVIDUAL DISTORTION TYPES.

perspective to restrict the quality measures to a single distortion type. Table II shows the performance of VIF and other measures on each of the five distortion types. Note that while the JND-Metrix, MSSIM and VIF perform quite well on individual distortion types, their performance worsens in cross-distortion validation, with VIF's worsening the least. Note that VIF performs better than (or at par with) JND-Metrix and MSSIM in cross-distortion validation (Table I) as well as individual distortion types.

Figure 10 shows graphically why is it important for a QA measure to perform well across distortions. Figure 10 shows the predicted DMOS calibration curves for each of the five distortion types present in the database [2]. Ideally for a QA method, these curves should lie on top of each other. If this were the case, then the QA measure could stably predict quality across distortion types. For the PSNR scale for example, we see that the good quality images (where DMOS is around 20), have PSNR values that lie in the approximate interval from 40 to 50 dBs, which is roughly 25% of the entire range of values that the PSNR takes. In contrast, we see that for good to medium quality images (DMOS values between 20 and 40), VIF curves are very close to each other, signifying that the mapping of VIF to visual quality is more stable, and has a smaller dependence on the underlying distortion type. Note that the distortion types present in the database are quite diverse, including linear blur, blocking, white noise as well as blurring/ringing from JPEG2000 compression, and transmission error in JPEG2000 bit stream.

At poorer quality ranges, the calibration curves for all four methods diverge, as shown in Figure 10 (one could note by visual inspection that the curves for VIF diverge far less than those for PSNR). One reason for this could be the lack of proper judgement scales in human observers for bad quality images, or psychometric scale warping effects at the lower end of quality.

3) Dependence on the HVS parameter: It was mentioned in Section IV that the value of the internal neural noise varaince, σ_n^2 was hand-optimized. It is instructional to study the dependence of the performance of VIF on σ_n^2. Ideally, σ_n^2 should depend on the dynamic range of the input, and a multiplicative constant should instead be tuned, as was done in [17], but here we only wish to show that the performance of VIF is relatively robust to small changes in the value of the parameter σ_n^2. Figure 11 shows how the RMSE in the quality prediction error varies

[2]The non-linearity used for MSSIM is different from the one used in Figure 8 and Table II for illustrative purposes.

with σ_n^2. It can be seen that VIF performs better than all the methods compared against in this paper for the entire range of values of σ_n^2 shown in Figure 11 (see Table I), with an approximate minimum occurring at 0.10.

4) Computational Complexity: The VIF has one disadvantage when compared against PSNR or MSSIM: it has a higher computational complexity. Most of this complexity comes from computing the wavelet decomposition, and the parameters of the distortion model. In [1], one version of the fidelity criterion using downsampling was presented, which has the potential to substantially reduce the computational complexity of the algorithm. Also, many estimation methods presented in the paper could be simplified greatly at the cost of slight reduction in performance. Nevertheless, even without these optimizations, VIF using the horizontal and vertical subbands with unoptimized MATLAB implementation takes about 13 seconds to run on 512×768 images on a Pentium IV, 2.6 GHz laptop. The bulk of this complexity comes from the highly overcomplete steerable pyramid decomposition. We are developing a lower complexity version of VIF in the pixel domain. For comparison, MSSIM takes about 2 seconds on 512×768 images.

VII. Conclusions and Future Work

In this paper we explored the relationship between image information and visual quality, and presented a visual information fidelity criterion for full-reference image quality assessment. The VIF, which was derived from a statistical model for natural scenes, a model for image distortions, and a human visual system model in an information-theoretic setting, outperformed traditional image QA methods in our simulations by a sizeable margin. The VIF was demonstrated to be better than a state-of-the-art HVS based method, the Sarnoff's JND-Metrix, as well as a state-of-the-art structural fidelity criterion, the structural similarity (SSIM) index, in our testing. We demonstrated that VIF performs well in single-distortion as well as in cross-distortion scenarios.

We are continuing efforts into extending VIF for video quality assessment using spatiotemporal natural scene models as well as by using inter-subband correlations. We are hopeful that this new paradigm will give new understanding into the relationship between image information and visual perception of quality.

References

[1] 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 Processing*, Apr. 2004, accepted.

[2] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE image quality assessment database release 2," 2005, available at *http://live.ece.utexas.edu/research/quality*.

[3] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Processing*, vol. 70, no. 3, pp. 177–200, Nov. 1998.

[4] T. N. Pappas and R. J. Safranek, "Perceptual criteria for image quality evaluation," in *Handbook of Image & Video Proc.*, A. Bovik, Ed. Academic Press, 2000.

[5] S. Winkler, "Issues in vision modeling for perceptual video quality assessment," *Signal Processing*, vol. 78, pp. 231–252, 1999.

[6] Z. Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Applications*, B. Furht and O. Marques, Eds. CRC Press, 2003.

[7] S. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity," in *Proc. SPIE*, vol. 1616, 1992, pp. 2–15.

[8] J. Lubin, "A visual discrimination model for image system design and evaluation," in *Visual Models for Target Detection and Recognition*, E. Peli, Ed. Singapore: World Scientific Publishers, 1995, pp. 207–220.

[9] A. B. Watson, "DCTune: A technique for visual optimization of DCT quantization matrices for individual images," in *Society for Information Display Digest of Technical Papers*, vol. XXIV, 1993, pp. 946–949.

[10] A. P. Bradley, "A wavelet visible difference predictor," *IEEE Trans. Image Processing*, vol. 5, no. 8, pp. 717–730, May 1999.

[11] Y. K. Lai and C.-C. J. Kuo, "A Haar wavelet approach to compressed image quality measurement," *Journal of Visual Communication and Image Representation*, vol. 11, pp. 17–40, Mar. 2000.

[12] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. SPIE*, vol. 2179, 1994, pp. 127–141.

[13] D. J. Heeger and P. C. Teo, "A model of perceptual image fidelity," in *Proc. IEEE Int. Conf. Image Proc.*, 1995, pp. 343–345.

[14] A. M. Pons, J. Malo, J. M. Artigas, and P. Capilla, "Image quality metric based on multidimensional contrast perception models," *Displays*, vol. 20, pp. 93–110, 1999.

[15] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Communications*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.

[16] I. Avcibaş, B. Sankur, and K. Sayood, "Statistical evaluation of image quality measures," *Journal of Electronic Imaging*, vol. 11, no. 2, pp. 206–23, Apr. 2002.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error measurement to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, Apr. 2004.

[18] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, May 2004.

[19] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *Journal of Mathematical Imaging and Vision*, vol. 18, pp. 17–33, 2003.

[20] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, pp. 1193–216, May 2001.

[21] J. M. Shapiro, "Embedded image coding using zerotrees of wavelets coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.

[22] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 6, no. 3, pp. 243–250, June 1996.

[23] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards, and Practice*. Kluwer Academic Publishers, 2001.

[24] 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, Dec. 1999.

[25] M. K. Mihçak, I. Kozintsev, K. Ramachandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Signal Processing Letters*, vol. 6, no. 12, pp. 300–303, Dec. 1999.

[26] J. K. Romberg, H. Choi, and R. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden markov models," *IEEE Trans. Image Processing*, vol. 10, no. 7, pp. 1056–1068, July 2001.

[27] M. J. Wainwright, E. P. Simoncelli, and A. S. Wilsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," *Applied and Computational Harmonic Analysis*, vol. 11, pp. 89–123, 2001.

[28] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Trans. Image Processing*, vol. 9, no. 10, pp. 1661–66, Oct. 2000.

[29] H. Choi and R. G. Baraniuk, "Multiscale image segmentation using wavelet-domain hidden Markov models," *IEEE Trans. Image Processing*, vol. 10, no. 9, pp. 1309–1321, Sept. 2001.

[30] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 49–71, 2000.

[31] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. Image Processing*, 2005, to appear.

[32] E. P. Simoncelli, "Modeling the joint statistics of images in the wavelet domain," in *Proc. SPIE*, vol. 3813, July 1999, pp. 188–195.

[33] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Proc. IEEE Int. Conf. Image Proc.*, Oct. 1995, pp. 444–447.

[34] J. Portilla, M. W. V. Strela, 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, Nov. 2003.

[35] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 1991.

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

[37] P. Corriveau, *et al.*, "Video quality experts group: Current results and future directions," *Proc. SPIE Visual Comm. and Image Processing*, vol. 4067, June 2000.

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

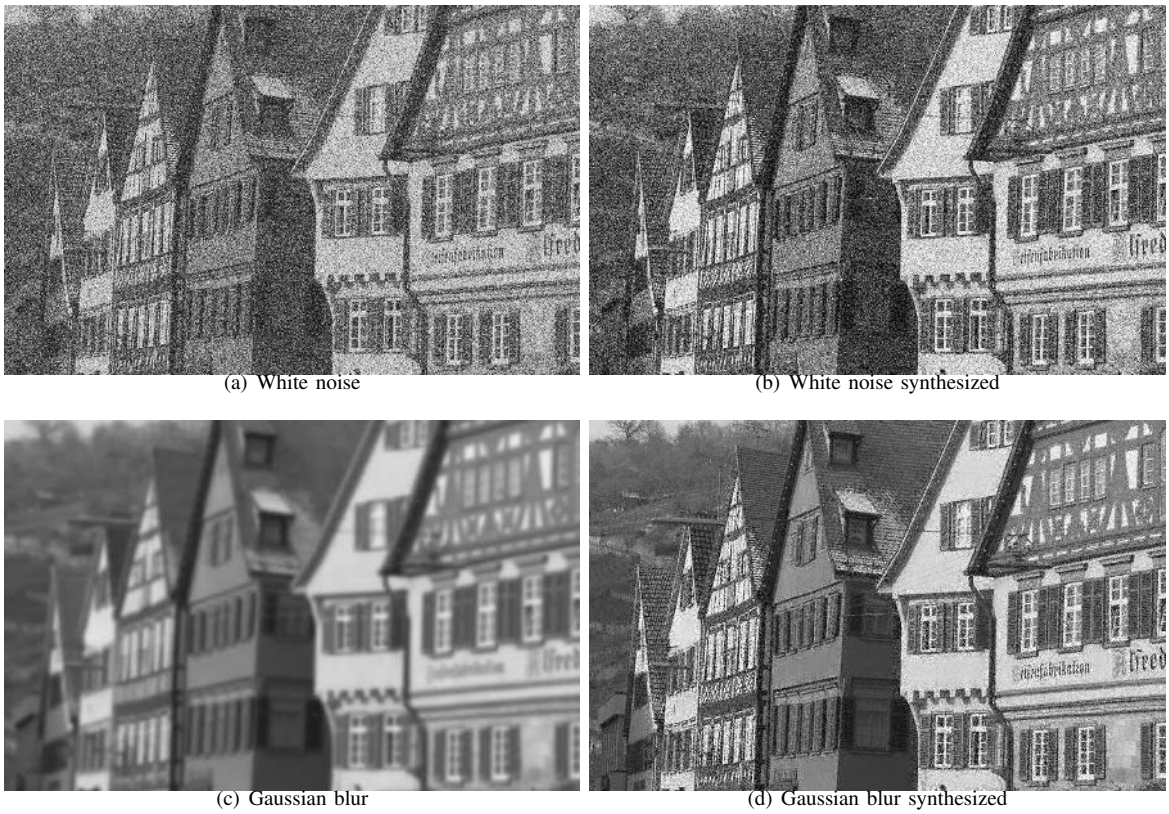
[39] Sarnoff Corporation, "JNDmetrix Technology," 2003, evaluation Version available: *http://www.sarnoff.com/products_services/video_vision/jndmetrix/downloads.asp*.

(a) JPEG2000 (b) JPEG2000 synthesized

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

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

Fig. 4. 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.



(a) White noise (b) White noise synthesized

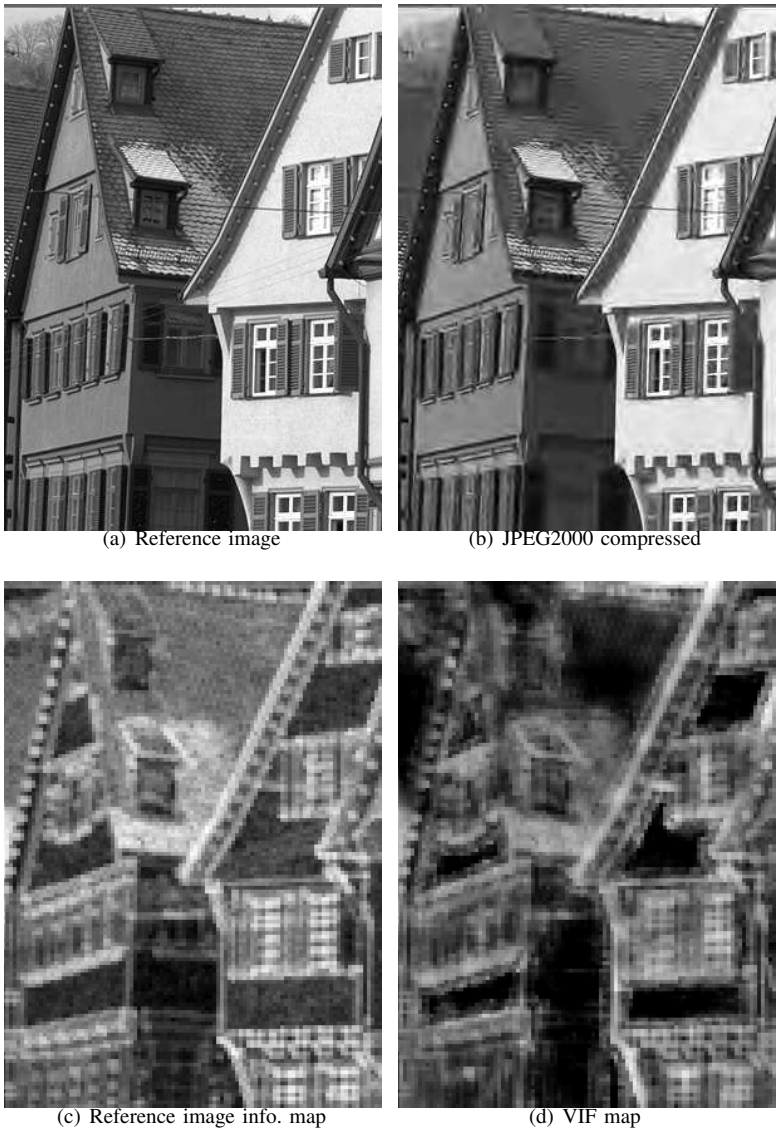
(c) Gaussian blur (d) Gaussian blur synthesized

Fig. 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.

(a) Reference image (b) Contrast enhancement

(c) Blurred (d) JPEG compressed

Fig. 6. The 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 Goldhill image (VIF $= 1$). (b) Contrast stretched Goldhill image (VIF $= 1.10$). (c) Gaussian blur (VIF $= 0.07$) and (d) JPEG compressed (VIF $= 0.10$).



(a) Reference image (b) JPEG2000 compressed

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

Fig. 7. Spatial maps showing how VIF captures spatial information loss. Note that VIF is not the mean of VIF-map.

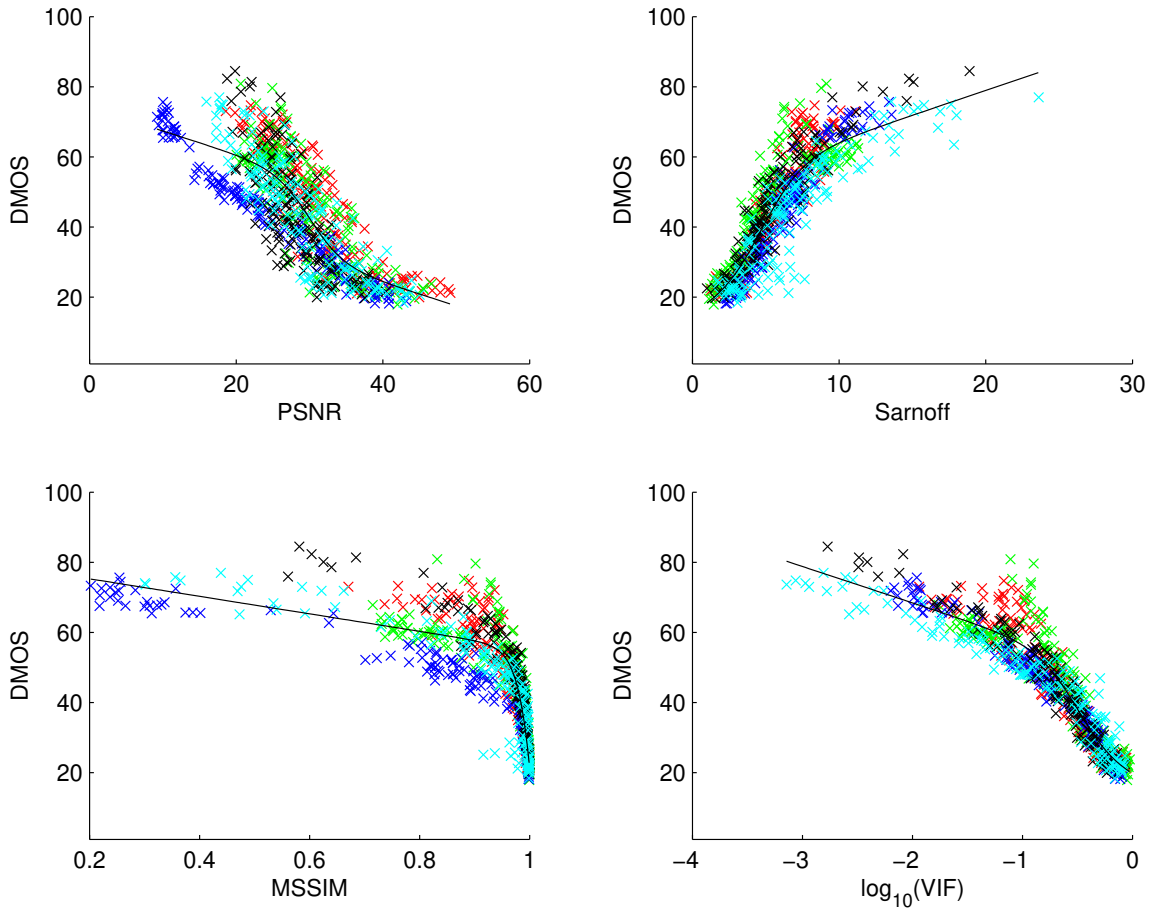


Fig. 8. Scatter plots for the four objective quality criteria: PSNR, Sarnoff's JND-metrix, MSSIM, and log(VIF) for VIF using horizontal/vertical orientations. The distortion types are: JPEG2000 (red), JPEG (green), white noise in RGB space (blue), Gaussian blur (black), and transmission errors in JPEG2000 stream over fast-fading Rayleigh channel (cyan).

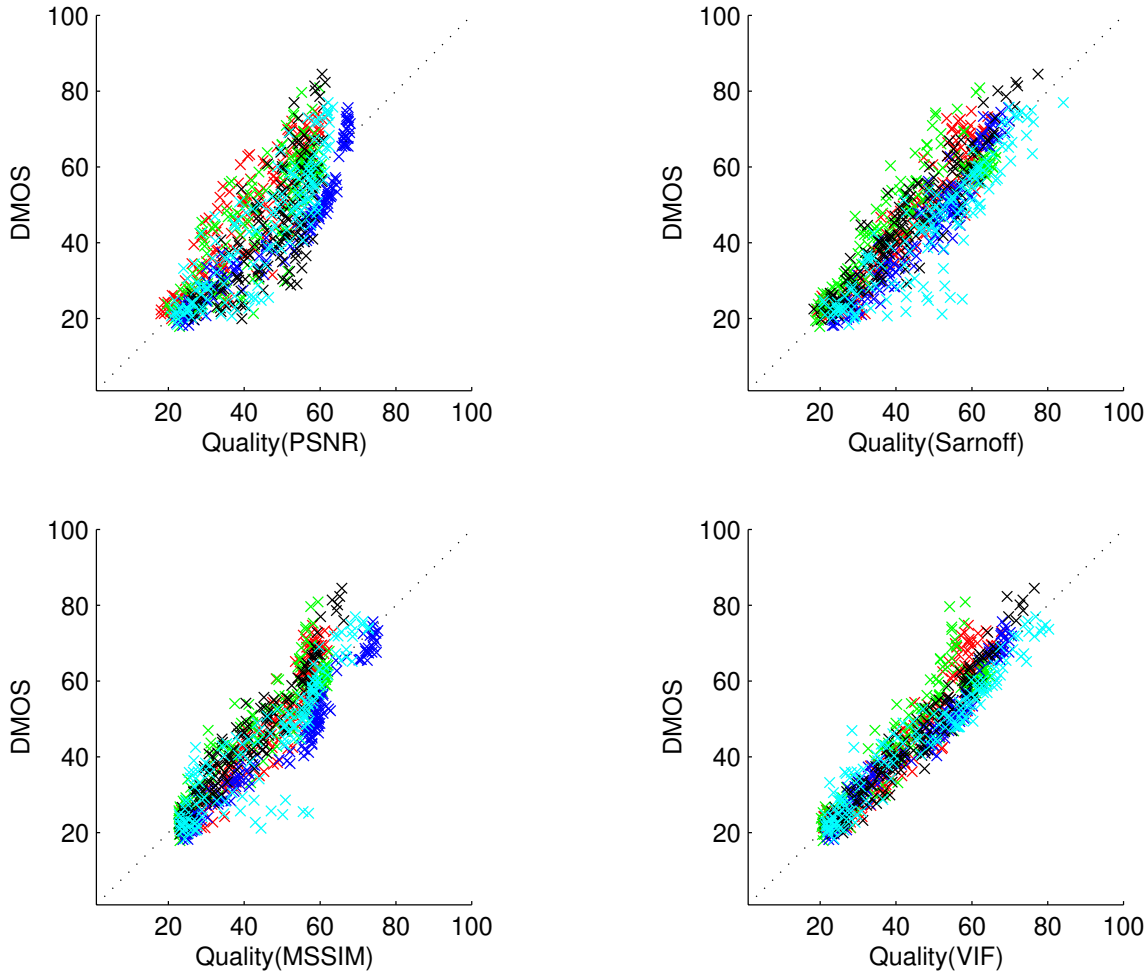


Fig. 9. Scatter plots for the quality predictions by the four methods after compensating for quality calibration: PSNR, Sarnoff's JND-metrix, MSSIM, and VIF using horizontal/vertical orientations. The distortion types are: JPEG2000 (red), JPEG (green), white noise in RGB space (blue), Gaussian blur (black), and transmission errors in JPEG2000 stream over fast-fading Rayleigh channel (cyan).

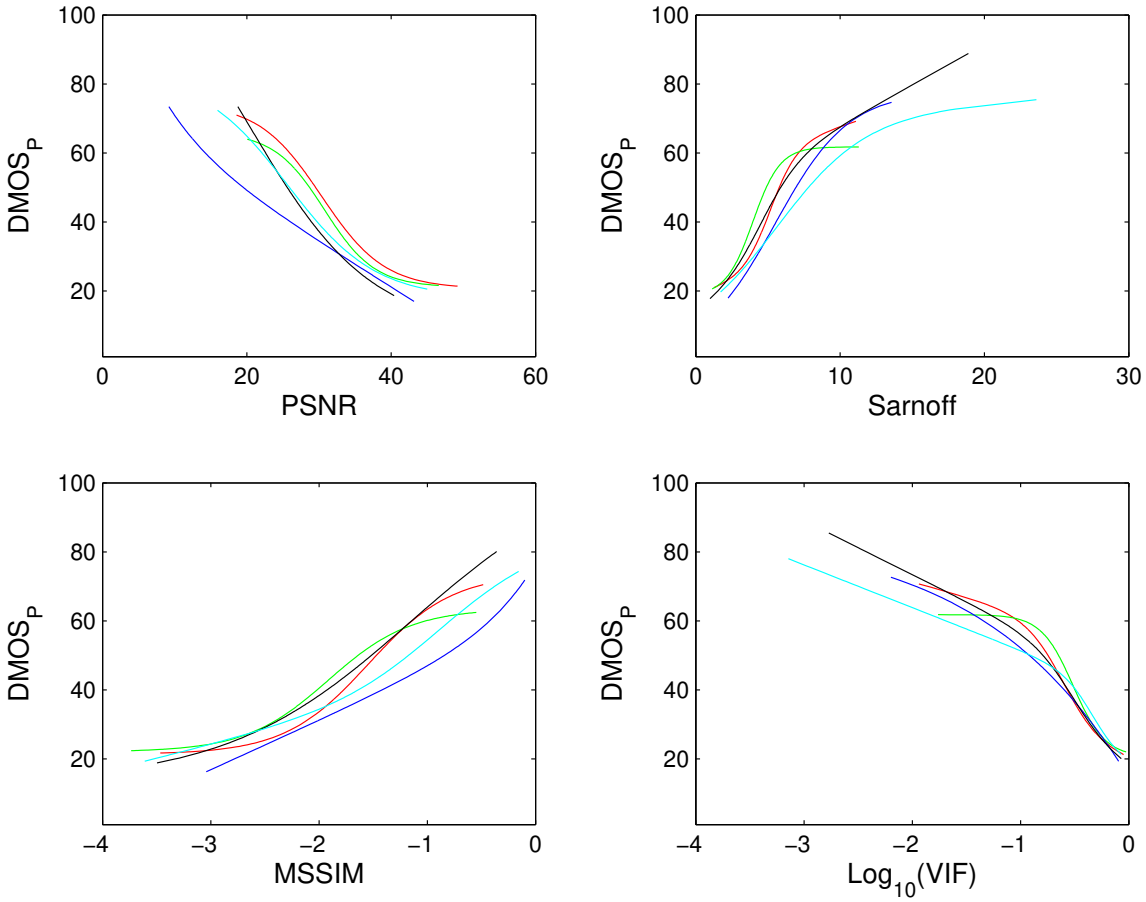


Fig. 10. Calibration curves for the four quality assessment methods for individual distortion types. The distortion types are: JPEG2000 (red), JPEG (green), white noise in RGB space (blue), Gaussian blur (black), and transmission errors in JPEG2000 stream over fast-fading Rayleigh channel (cyan). Note that VIF can be stably calibrated for predicting quality for a wider range of distortion types. The mapping used for MSSIM in this figure is $\log_{10}(1 - \text{MSSIM})$ for illustrative purposes.

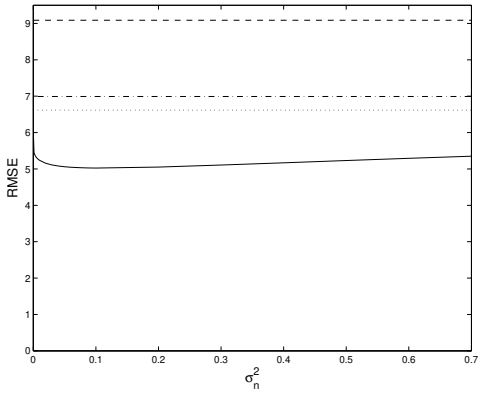


Fig. 11. Dependence of VIF performance on the σ_n^2 parameter. Note that VIF performs better than other methods against which it is compared in this paper for all range of values of σ_n^2 shown this figure: VIF (solid), PSNR (dashed), Sarnoff JNDMetrix 8.0 (dash-dot), and MSSIM (dotted) .