An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics

Hamid Rahim Sheikh, *Student Member, IEEE,* Alan C. Bovik, *Fellow, IEEE,* Gustavo de Veciana, *Senior Member, 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. Traditionally, image QA algorithms 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 arbitrary signal fidelity criteria. In this paper we approach the problem of image QA by proposing a novel information fidelity criterion that is based on natural scene statistics. QA systems are invariably involved with judging the visual quality of images and videos that are meant for 'human consumption'. Researchers have developed sophisticated models to capture the statistics of natural signals, that is, pictures and videos of the visual environment. Using these statistical models in an information-theoretic setting, we derive a novel QA algorithm that provides clear advantages over the traditional approaches. In particular, it is parameterless and outperforms current methods in our testing. We validate the performance of our algorithm with an extensive subjective study involving 779 images. We also show that although our approach distinctly departs from traditional HVS based methods, it is functionally similar to them under certain conditions, yet it outperforms them due to improved modeling. The code and the data from the subjective study are available at [1].

Index Terms

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

H. R. Sheikh is affiliated with the Laboratory for Image and Video Engineering, Department of Electrical & Computer Engineering, The University of Texas at Austin, Austin, TX 78712-1084 USA, Phone: (512) 471-2887, email: sheikh@ece.utexas.edu

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

G. de Veciana is affiliated with the Department of Electrical & Computer Engineering, The University of Texas at Austin, Austin, TX 78712-1084USA, Phone: (512) 471-1573, email:gustavo@ece.utexas.edu

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 reproductions. These systems typically involve tradeoffs between system resources and the visual quality of the output. In order to make these tradeoffs efficiently, 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 (QA) 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 this paper we explore a novel information theoretic criterion for image fidelity using 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*. We propose to use natural scene models in conjunction with distortion models to quantify the statistical information shared between the test and the reference images, and posit that this shared information is an aspect of fidelity that relates well with visual quality.

The approaches discussed above describe three ways in which one could look at the image quality assessment problem. One viewpoint is *structural*, from the image-content perspective, in which images are considered to be projections of objects in the three dimensional environment that could come from a wide variety of lighting conditions. Such variations constitute *non-structural* distortion that should be treated differently from structural ones, e.g., blurring or blocking that could hamper cognition. The second viewpoint is *psychovisual*, from the human visual receiver perspective, in which researchers simulate the processing of images by the human visual system, and predict the perceptual significance of errors. The third viewpoint, the one that we take in this paper, is the *statistical* viewpoint that considers natural images to be signals with certain statistical properties. These three views are fundamentally connected with each other by the following hypothesis: the physics of image formation of the natural three dimensional visual environment leads to certain statistical properties of the visual stimulus, in response to which the visual system has evolved over eons. However, different aspects of each of these views may have different complexities when it comes to analysis and modeling. In this paper we show that the statistical approach to image quality assessment requires few assumptions, is simple and methodical to derive, and yet it is competitive with the other two approaches in that it outperforms them in our testing. Also, we show that the statistical approach to quality assessment is a *dual* of the psychovisual approach to the same problem; we demonstrate this duality towards the end of this paper.

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 information fidelity criterion. Implementation and subjective validation details are provided in Sections IV and V, while the results are discussed in Section VI. In Section VII we compare and contrast our method with HVS based methods, and conclude the paper in Section VIII.

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 [2]–[5]).

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 [3]–[5]. A number of HVS based methods have been proposed in the literature. Some representative methods include [6]–[13].

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 [14] and [15], a number of these are evaluated for the purpose of quality assessment. In [16] 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. It was hypothesized 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.

C. Limitations

A number of limitations of HVS based methods are discussed in [16]. 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 [16], 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. In [16], the SSIM was constructed by *hypothesizing* the functional forms of structural and non-structural distortions and the interaction between them. In this paper we take a new approach to the quality assessment problem. As mentioned in the Introduction, the third alternative to QA, apart from HVS based and structural approaches, is the statistical approach, which we use in an information theoretic setting. Needless to say, even our approach will make certain assumptions, but once assumptions regarding the source and distortion models and the suitability of mutual information as a valid measure of perceptual information fidelity are made, the components of our algorithm and their interactions fall through without resorting to arbitrary formulations.

Due to the importance of the quality assessment problem to researchers and developers in the image and video processing community, a consortium of experts, the video quality experts group (VQEG), was formed in 1997 to develop, validate, and recommend objective video quality assessment methods [17]. VQEG Phase I testing reported that all of the proponent methods tested, which contained some of the most sophisticated video quality assessment methods of the time, were statistically indistinguishable from PSNR under their testing conditions [18]. The Phase II of testing, which consisted of new proponents under different testing configurations, is also complete and the final report has recommended an FR QA method, although it has been reported that none of the methods tested were comparable to the ‘null model’, a hypothetical model that predicts quality exactly [19], meaning that QA methods need to be improved further.

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, 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 [20]). 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 [21]. 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 [22]–[25], denoising algorithms [26]–[28], image modeling [29], image segmentation [30], and texture analysis and synthesis [31]. 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 [32], 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) [28]. 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 [33]. 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 [28], a more detailed introduction to which will be given in the next section.

III. Information Fidelity Criterion for Image Quality Assessment

In this paper, we propose to approach the quality assessment problem as an information fidelity problem, where a natural image source communicates with a receiver through a channel. The channel imposes fundamental limits on how much information could flow from the source (the reference image), through the channel (the image distortion process) to the receiver (the human observer). Figure 1 shows the scenario graphically. A standard way of dealing with such problems is to analyze them in an information-theoretic framework, in which the mutual information between the input and the output of the channel (the reference and the test images) is quantified using a model for the source and a distortion model. Thus, our assertion in proposing this framework is that the *statistical information* that a test image has of the reference is a good way of quantifying fidelity that could relate well with visual quality.

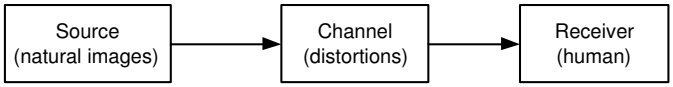

Fig. 1. The quality assessment problem could be analyzed using an information theoretic framework in which a source transmits information through a channel to a receiver. The mutual information between the input of the channel (the reference image) and the output of the channel (the test image) quantifies the amount of information that could ideally be extracted by the receiver (the human observer) from the test image.

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. We model one subband of the wavelet decomposition of an image as a GSM RF, $\mathcal{C} = \{C_i : \ i \in \mathrm{I}\}$, where I denotes the set of spatial indices for the RF. $\mathcal{C}$ is a product of two stationary RF's that are independent of each other [28]:

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

where $\mathcal{S} = \{S_i : \ \ i \in \mathrm{I}\}$ is an RF of positive scalars and $\mathcal{U} = \{U_i : \ \ i \in \mathrm{I}\}$ is a Gaussian scalar RF with mean zero and variance σ_U^2. Note that for the GSM defined in (1), while the marginal distribution of C_i may be sharply-peaked and heavy-tailed, such as those of natural scenes in the wavelet domain, conditioned on S_i, C_i are normally distributed, that is,

$$p_{C_i|S_i}(c_i|s_i) \sim \mathcal{N}(0, s_i^2\sigma_U^2) \tag{2}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian density with mean μ and variance σ^2. Another observation is that given S_i, C_i are independent of $S_j \ \forall \ j \neq i$, meaning that the variance of the coefficient C_i specifies its distribution completely. Additionally, if the RF $\mathcal{U}$ is white, then the elements of $\mathcal{C}$ are conditionally independent given $\mathcal{S}$. The GSM framework can model the marginal statistics of the wavelet coefficients of natural images, the non-linear dependencies that are present between the coefficients, as well as the space-varying localized statistics through appropriate modeling of the RF $\mathcal{S}$ [28].

B. The Distortion Model

The distortion model that we use in this paper is also described in the wavelet domain. It is a simple signal attenuation and additive Gaussian noise model in each subband:

$$\mathcal{D} = \mathcal{G}\mathcal{C} + \mathcal{V} = \{g_i C_i + V_i : \ i \in \mathrm{I}\} \tag{3}$$

where $\mathcal{C}$ denotes the RF from a subband in the reference signal, $\mathcal{D} = \{D_i : \ \ i \in \mathrm{I}\}$ denotes the RF from the corresponding subband from the test (distorted) signal, $\mathcal{G} = \{g_i : \ i \in \mathrm{I}\}$ is a deterministic scalar attenuation field, and $\mathcal{V} = \{V_i : \ \ i \in \mathrm{I}\}$ is a stationary additive zero-mean Gaussian noise RF with variance σ_V^2. The RF $\mathcal{V}$ is white and is independent of $\mathcal{S}$ and $\mathcal{U}$. This model captures two important, and complementary, distortion types: blur and additive noise. We will assume that most distortion types that are prevalent in real world systems can be roughly described *locally* by a combination of these two. In our model, the attenuation factors g_i can capture the

loss of signal energy in a subband to the blur distortion, while the process $\mathcal{V}$ can capture additive noise separately. Additionally, changes in image contrast that result from variations in ambient lighting are not modeled as noise since they too can be incorporated into the attenuation field $\mathcal{G}$.

The choice of a proper distortion model is crucial for image fidelity assessments that are expected to reflect perceptual quality. In essence we want the distortion model to characterize what the HVS perceives as distortion. Based on our experience with different distortion models, we are inclined to hypothesize that the visual system has evolved over time to optimally estimate natural signals embedded in *natural distortions*: blur, white noise, and brightness and contrast stretches due to changes in ambient lighting. The visual stimulus that is encoded by the human eyes is blurred by the optics of the eye as well as the spatially-varying sampling in the retina. It is therefore natural to expect evolution to have worked towards near-optimal processing of blurry signals, say for controlling the focus of the lens, or guiding visual fixations. Similarly, white noise arising due to photon noise or internal neuron noise (especially in low light conditions) affects all visual signals. Adaptation in the HVS to changes in ambient lighting has been known to exist for a long time [34]. Thus HVS signal estimators would have evolved in response to natural signals corrupted by natural distortions, and would be near-optimal for them, but sub-optimal for other distortion types (such as blocking or colored noise) or signal sources. Hence 'over-modeling' the signal source or the distortion process is likely to fail for QA purposes, since it imposes assumptions on the existence of near-optimal estimators in the HVS for the chosen signal and distortion models, which may *not* be true. In essence distortion modeling combined with NSS source modeling is a *dual* of HVS signal estimator modeling.

Another hypothesis is that the field $\mathcal{G}$ could account for the case when the additive noise $\mathcal{V}$ is linearly correlated with $\mathcal{C}$. Previously, researchers have noted that as the correlation of the noise with the reference signal increases, MSE becomes poorer in predicting perceptual quality [35]. While the second hypothesis could be a corollary to the first, we feel that both of these hypotheses (and perhaps more) need to be investigated further with psychovisual experiments so that the exact contribution of a distortion model in the quality prediction problem could be understood properly. For the purpose of image quality assessment presented in this paper, the distortion model of (3) is adequate, and works well in our simulations.

C. The Information Fidelity Criterion

Given a statistical model for the source and the distortion (channel), the obvious information fidelity criterion is the mutual information between the source and the distorted images. We first derive the mutual information for one subband and later generalize for multiple subbands.

Let $C^N = (C_1, C_2, \ldots, C_N)$ denote N elements from $\mathcal{C}$. In this section we will assume that the underlying RF $\mathcal{U}$ is uncorrelated (and hence $\mathcal{C}$ is an RF with conditionally independent elements given $\mathcal{S}$), and that the distortion model parameters $\mathcal{G}$ and σ_V^2 are known *a priori*. Let $D^N = (D_1, D_2, \ldots\ D_N)$ denote the *corresponding* N elements from $\mathcal{D}$. The mutual information between these is denoted as $I(C^N; D^N)$.

Due to the non-linear dependence among the C^N by way of $\mathcal{S}$, it is much easier to analyze the mutual information assuming $\mathcal{S}$ is known. This conditioning 'tunes' the GSM model for the particular reference image, and thus models

the source more specifically. Thus the information fidelity criterion that we propose in this paper is the conditional mutual information $I(C^N; D^N|S^N = s^N)$, where $S^N = (S_1, S_2, \ldots, S_N)$ are the corresponding N elements of $\mathcal{S}$, and s^N denotes a *realization* of S^N. In this paper we will denote $I(C^N; D^N|S^N = s^N)$ as $I(C^N; D^N|s^N)$. With the stated assumptions on $\mathcal{C}$ and the distortion model (3), one can show:

$$
\begin{aligned}
I(C^N; D^N|s^N) &= \sum_{j=1}^{N}\sum_{i=1}^{N} I(C_i; D_j|C^{i-1}, D^{j-1}, s^N) && (4)\\
&= \sum_{i=1}^{N} I(C_i; D_i|C^{i-1}, D^{i-1}, s^N) && (5)\\
&= \sum_{i=1}^{N} I(C_i; D_i|s_i) && (6)
\end{aligned}
$$

where we get (4) by the chain rule [36], and (5) and (6) by conditional independence of $\mathcal{C}$ given $\mathcal{S}$, independence of the noise $\mathcal{V}$, the fact that the distortion model keeps D_i independent of $C_j,\ \forall i \neq j$, and that given S_i, C_i and D_i are independent of $S_j\ \forall\ j \neq i$. Using the fact that C_i are Gaussian given S_i, and V_i are also Gaussian with variance σ_V^2, we get:

$$
\begin{aligned}
I(C^N; D^N|s^N) &= \sum_{i=1}^{N} I(C_i; D_i|s_i) && (7)\\
&= \sum_{i=1}^{N} (h(D_i|s_i) - h(D_i|C_i, s_i)) && (8)\\
&= \sum_{i=1}^{N} (h(g_i C_i + V_i|s_i) - h(V_i)) && (9)\\
&= \frac{1}{2}\sum_{i=1}^{N} \log_2\left(1 + \frac{g_i^2 s_i^2 \sigma_U^2}{\sigma_V^2}\right) && (10)
\end{aligned}
$$

where $h(X)$ denotes the differential entropy of a continuous random variable X, and for X distributed as $\mathcal{N}(\mu, \sigma^2)$, $h(X) = 1/2 \log_2 2\pi e\sigma^2$ [36].

Equation (10) was derived for one subband. It is straightforward to use separate GSM RF's for modeling each subband of interest in the image. We will denote the RF modeling the wavelet coefficients of the reference image in the k-th subband as $\mathcal{C}^k$, and in test (distorted) image as $\mathcal{D}^k$, and assume that $\mathcal{C}^k$ are independent of each other. We will further assume that each subband is distorted independently. Thus, the RF's $\mathcal{V}^k$ are also independent of each other. The information fidelity criterion (IFC) is then obtained by summing over all subbands:

$$
\text{IFC} = \sum_{k \in \text{subbands}} I(C^{N_k,k}; D^{N_k,k}|s^{N_k,k}) \qquad (11)
$$

where $C^{N_k,k}$ denotes N_k coefficients from the RF $\mathcal{C}^k$ of the k-th subband, and similarly for $D^{N_k,k}$ and $s^{N_k,k}$.

Equation (11) is our information fidelity criterion that quantifies the statistical information that is shared between the source and the distorted images. An attractive feature of our criterion is that like MSE and some other mathematical fidelity metrics, it does not involve parameters associated with display device physics, data from visual

psychology experiments, viewing configuration information, or stabilizing constants, which dictate the accuracy of HVS based FR QA methods (and some structural ones too). The IFC does not require training data either. However some implementation parameters will obviously arise once (11) is implemented. We will discuss implementation in the next section.

The IFC is not a distortion metric, but a fidelity criterion. It theoretically ranges from zero (no fidelity) to infinity (perfect fidelity within a non-zero multiplicative constant in the absence of noise[1]). Perfect fidelity within a multiplicative constant is something that is in contrast with the approach in SSIM [16], in which contrast distortion (multiplicative constant) was one of the three attributes of distortion that was regarded as a visual degradation, albeit one that has a different (and 'orthogonal') contribution towards perceptual fidelity than noise and local-luminance distortions. In this paper we view multiplicative constants (contrast stretches) as signal gains or attenuations *interacting* with additive noise. Thus, with this approach, the same noise variance would be perceptually less annoying if it were added to a contrast stretched image than if it were added to a contrast attenuated image. Since each subband has its own multiplicative constant, blur distortion could also be captured by this model as the finer scale subbands would be attenuated more than coarser scale subbands.

IV. Implementation Issues

In order to implement the fidelity criterion in (11) a number of assumptions are required about the source and the distortion models. We outline them in this section.

A. Assumptions about the Source Model

Note that mutual information (and hence the IFC) can only be calculated between RF's and not their *realizations*, that is, a particular reference and 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 statistics obtained from particular realizations.

For the scalar GSM model, estimates of s_i^2 can be obtained by localized sample variance estimation since for natural images $\mathcal{S}$ is known to be a spatially correlated field, and σ_U^2 can be assumed to be unity without loss of generality.

B. Assumptions about the Distortion Model

The IFC assumes that the distortion model parameters $\mathcal{G}$ and σ_V^2 are known *a priori*, but these would need to be estimated in practice. We propose to partition the subbands into blocks and assume that the field $\mathcal{G}$ is constant over such blocks, as are the noise statistics σ_V^2. The value of the field $\mathcal{G}$ over block l, which we denote as g_l, and

[1]Differential entropy is invariant to translation, and so the IFC is infinite for perfect fidelity within an additive constant in the absence of noise as well. However, since we are applying the IFC in the wavelet domain on 'AC' subbands only to which the GSM model applies, the zero-mean assumptions on $\mathcal{U}$ and $\mathcal{V}$ imply that this case will not happen.

the variance of the RF $\mathcal{V}$ over block l, which we denote as $\sigma^2_{V,l}$, 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 (3) are available:

$$\widehat{g}_l = \widehat{\mathrm{Cov}}(C,D)\widehat{\mathrm{Cov}}(C,C)^{-1} \quad (12)$$

$$\widehat{\sigma}^2_{V,l} = \widehat{\mathrm{Cov}}(D,D) - g_l\widehat{\mathrm{Cov}}(C,D) \quad (13)$$

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

C. Wavelet Bases and Inter-Coefficient Correlations

The derivation leading to (10) assumes that $\mathcal{U}$ is uncorrelated, and hence $\mathcal{C}$ is independent given $\mathcal{S}$. In practice, if the wavelet decomposition is orthogonal, the underlying $\mathcal{U}$ could be approximately uncorrelated. In such cases, one could use (10) for computing the IFC. However real cartesian-separable orthogonal wavelets are not good for image analysis since they have poor orientation selectivity, and are not shift invariant. In our implementation, we chose the steerable pyramid decomposition with six orientations [37]. This gives better orientation selectivity than possible with real cartesian separable wavelets. However the steerable pyramid decomposition is over-complete, and the neighboring coefficients $\mathcal{C}$ from the same subband are linearly correlated. In order to deal with such correlated coefficients, we propose two simple approximations that work well for quality assessment purposes.

1) Vector GSM: Our first approximation is to partition the subband into non-overlapping block-neighborhoods and assume that the neighborhoods are uncorrelated with each other. One could then use a vector form of the IFC by modeling each neighborhood as a vector random variable. This 'blocking' of coefficients results in an upper bound:

$$I(C^N;D^N|s^N) \leq \sum_{j=1}^{N/M} I(\overrightarrow{C}_j;\overrightarrow{D}_j|s_j)$$

where $\overrightarrow{C}_j = (C_{j,i}, i = 1 \ldots M)$ is a vector of M wavelet coefficients that form the j-th neighborhood. All such vectors, associated with non-overlapping neighborhoods, are assumed to be uncorrelated with each other. We now model the wavelet coefficient neighborhood as a vector GSM. Thus, the vector RF $\mathcal{C} = \{\overrightarrow{C}_i : i \in \mathrm{I}'\}$ on a lattice I' is a product of a *scalar* RF $\mathcal{S}$ and a zero-mean Gaussian *vector* RF $\mathcal{U} = \{\overrightarrow{U}_i : i \in \mathrm{I}'\}$ of covariance $\mathbf{C}_{\overrightarrow{U}}$. The noise $\mathcal{V}$ is also a zero-mean vector Gaussian RF of same dimensionality as $\mathcal{C}$, and has covariance $\mathbf{C}_{\overrightarrow{V}}$. If we assume that $\overrightarrow{U}_i$ is independent of $\overrightarrow{U}_j$, $\forall i \neq j$, it is quite easy to show (by using differential entropy for Gaussian vectors) that:

$$I(C^N;D^N|s^N) \leq \sum_{j=1}^{N/M} I(\overrightarrow{C}_j;\overrightarrow{D}_j|s_j) \quad (14)$$

$$= \frac{1}{2}\sum_{j=1}^{N/M} \log_2\left(\frac{|g_j^2 s_j^2 \mathbf{C}_{\overrightarrow{U}} + \mathbf{C}_{\overrightarrow{V}}|}{|\mathbf{C}_{\overrightarrow{V}}|}\right) \quad (15)$$

where the differential entropy of a continuous vector random vector $\overrightarrow{X}$ distributed as a multivariate Gaussian $\mathcal{N}(\overrightarrow{\mu},\mathbf{\Sigma})$, $h(\overrightarrow{X}) = 1/2\log_2(2\pi e)^d|\mathbf{\Sigma}|$ where $|.|$ denotes the determinant, and d is the dimension of $\overrightarrow{X}$ [36].

Recalling that $\mathbf{C}_{\overrightarrow{U}}$ is symmetric and can be factorized as $\mathbf{Q\Lambda Q^T}$ with orthonormal $\mathbf{Q}$ and eigenvalues λ_k, and that for a distortion model where $\mathbf{C}_{\overrightarrow{V}} = \sigma_V^2\mathbf{I}$, the IFC simplifies as follows[2]:

$$I(C^N; D^N|s^N) \leq \sum_{j=1}^{N/M} I(\overrightarrow{C}_j; \overrightarrow{D}_j|s_j) \tag{16}$$

$$= \frac{1}{2}\sum_{j=1}^{N/M} \log_2\left(\frac{|g_j^2 s_j^2 \mathbf{C}_{\overrightarrow{U}} + \sigma_V^2\mathbf{I}|}{|\sigma_V^2\mathbf{I}|}\right) \tag{17}$$

$$= \frac{1}{2}\sum_{j=1}^{N/M} \log_2\left(\frac{|g_j^2 s_j^2 \mathbf{Q\Lambda Q^T} + \sigma_V^2\mathbf{I}|}{\sigma_V^{2M}}\right) \tag{18}$$

$$= \frac{1}{2}\sum_{j=1}^{N/M} \log_2\left(\frac{|g_j^2 s_j^2 \mathbf{\Lambda} + \sigma_V^2\mathbf{I}|}{\sigma_V^{2M}}\right) \tag{19}$$

$$= \frac{1}{2}\sum_{j=1}^{N/M}\sum_{k=1}^{M} \log_2\left(1 + \frac{g_j^2 s_j^2 \lambda_k}{\sigma_V^2}\right) \tag{20}$$

where the numerator term inside the logarithm of (19) is the determinant of a diagonal matrix and hence equals the product of the diagonal terms. The bound in (16) shrinks as M increases. In our simulations we use vectors from 3×3 spatial neighborhoods and achieve good performance. Equation (20) is the form that is used for implementation.

For the vector GSM model, the maximum-likelihood estimate of s_j^2 can be found as follows [38]:

$$s_j^2 = \frac{\overrightarrow{C}_j^T \mathbf{C_u}^{-1} \overrightarrow{C}_j}{M} \tag{21}$$

where M is the dimensionality of $\overrightarrow{C}_j$. Estimation of the covariance matrix $\mathbf{C}_{\overrightarrow{U}}$ is also straightforward from the reference image wavelet coefficients [38]:

$$\widehat{\mathbf{C}}_{\overrightarrow{U}} = \frac{M}{N}\sum_{j=1}^{N/M} \overrightarrow{C}_j \overrightarrow{C}_j^T \tag{22}$$

In (21) and (22), $\frac{1}{N}\sum_{i=1}^{N} s_i^2$ is assumed to be unity without loss of generality [38].

2) Downsampling: Our second approximation is to use a subset of the coefficients by *downsampling* $\mathcal{C}$. Downsampling reduces the correlation between coefficients. We will assume that the downsampled subband is approximately uncorrelated, and then use (10) for scalar GSM on the downsampled subband. The underlying assumption in the downsampling approach is that the quality prediction from the downsampled subbands should be approximately the same as the prediction from the complete subband. This downsampling approach has an additional advantage that it makes it possible to substantially reduce the complexity of computing the wavelet decomposition since only a fraction of the subband coefficients need to be computed. In our simulations we discovered that the wavelet decomposition is the most computationally expensive step. Significant speedups are possible with the typical downsampling factors of twelve or fifteen in our simulations. We downsample a subband along and across the principal orientations of the respective filters. In our simulations, the downsampling was done using nearest-neighbor interpolation.

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

[2]Utilizing the structure of $\mathbf{C}_{\overrightarrow{U}}$ and $\mathbf{C}_{\overrightarrow{V}}$ helps in faster implementations through matrix factorizations.

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 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", which was mapped linearly on to a $1-100$ range. 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) [18] and then converted to Z-scores [39], 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 [1].

VI. Results

In this section we present results on validation of the IFC on the database presented in Section V, and comparisons with other quality assessment algorithms. Specifically, we will compare the performance of our algorithm against PSNR, SSIM [16], and the well known Sarnoff model (Sarnoff JND-Metrix 8.0 [40]). We present results for five versions of the IFC: scalar GSM, scalar GSM with downsampling by three along the principal orientation and five across, vector GSM, vector GSM using the horizontal and vertical orientations only, and vector GSM using horizontal and vertical orientations and only one eigenvalue in the summation of (20). Table I summarizes the validation results.

A. Simulation Details

Some additional simulation details are as follows. Although full color images were distorted in the subjective evaluation, the QA algorithms (except JND-Metrix) operated upon the luminance component only. For the scalar GSM with no downsampling, a 5×5 moving window was used for local variance estimation (s_i^2), and 16×16 non-overlapping blocks were used for estimating parameters g_l and $\sigma_{V,l}^2$. The blocking was done in order for the stationarity assumptions on the distortion model to approximately hold. For the scalar GSM with downsampling, all parameters were estimated on the downsampled signals. A 3×3 window was used for variance estimation, while 8×8 blocks were used for the distortion model estimation. For vector GSM, vectors were constructed from non-overlapping 3×3 neighborhoods, and the distortion model was estimated with 18×18 non-overlapping blocks.

In all versions of the IFC, only the subbands at the finest level were used in the summation of (11). Since the sizes of the images in the database were different, the IFC was normalized by the number of pixels in each image. MSSIM (Mean SSIM) was calculated on the luminance component after decimating (filtering and downsampling) it by a factor of 4 [16].

B. Calibration of the Objective Score

It is generally acceptable for a QA method to stably predict subjective quality within a monotonic 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 of the QA algorithm. Thus in both the VQEG Phase-I and Phase-II testing and validation, a monotonic non-linear mapping between the objective and the subjective scores was allowed, and all the performance validation metrics were computed *after* compensating for it [18]. This is true for the results in Table I, where a five-parameter non-linearity (a logistic function with additive linear term) is used for all methods except for the IFC, for which we used the mapping on the logarithm of the IFC. The quality predictions, after compensating for the mapping, are shown in Figure 2. The mapping function used is given in (23), while the fitting was done using MATLAB's *fminsearch*.

$$\mathrm{Quality}(x) = \beta_1 \mathrm{logistic}\left(\beta_2, (x - \beta_3)\right) + \beta_4 x + \beta_5 \tag{23}$$

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

C. Discussion

Table I shows that the IFC, even in its simplest form, is competitive with all state-of-the-art FR QA methods presented in this paper. The comparative results between MSSIM and Sarnoff's JND-Metrix are qualitatively similar to those reported in [16], only that both of these methods perform poorer in the presence of a wider range of distortion types than reported in [16]. However, MSSIM still outperforms JND-Metrix by a sizeable margin using any of the validation criteria in Table I.

The IFC also performs demonstrably better than Sarnoff's JND-Metrix under all of the alternative implementations of the IFC. The vector-GSM form of the IFC outperforms even MSSIM. Note that the downsampling approximation performs better than scalar IFC without downsampling, even though the downsampled version operates on signals that are fifteen times smaller, and hence it is a computationally more feasible alternative to other IFC implementations at a reasonably good performance. Also note that the IFC as well as MSSIM use only the luminance components of the images to make quality predictions whereas the JND-Metrix uses all color information. Extending the IFC to incorporate color could further improve performance.

An interesting observation is that when only the smaller eigenvalues are used in the summation of (20), the performance increases dramatically. The last row in Table I, and Figure 2 show results when only the smallest eigenvalue is used in the summation in (20). The performance is relatively unaffected up to an inclusion of five

Validation against DMOS					
Model	CC	MAE	RMS	OR	SROCC
PSNR	0.826	7.272	9.087	0.114	0.820
JND-Metrix	0.901	5.252	6.992	0.046	0.902
MSSIM	0.912	4.979	6.616	0.035	0.910
IFC (no ds)	0.911	5.078	6.652	0.041	0.908
IFC (ds 3/5)	0.913	5.009	6.587	0.041	0.909
IFC (vec)	0.917	4.919	6.437	0.039	0.915
IFC (h/v, vec)	0.919	4.855	6.366	0.032	0.918
IFC (h/v, 1 ev)	0.929	4.523	5.941	0.059	0.928

TABLE I

VALIDATION SCORES FOR DIFFERENT QUALITY ASSESSMENT METHODS. THE METHODS TESTED WERE PSNR, SARNOFF JND-METRIX 8.0 [40], MSSIM [16], IFC FOR SCALAR GSM WITHOUT DOWNSAMPLING, IFC FOR SCALAR GSM WITH DOWNSAMPLING BY 3 ALONG ORIENTATION AND 5 ACROSS, IFC FOR VECTOR GSM, IFC FOR VECTOR GSM USING HORIZONTAL AND VERTICAL ORIENTATIONS ONLY, AND IFC FOR VECTOR GSM AND HORIZONTAL/VERTICAL ORIENTATIONS WITH ONLY THE SMALLEST EIGENVALUE IN (20). 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).

smallest eigenvalues (out of nine). One hypothesis that could explain this observation is that a measurement noise could be present in IFC whose strength depends upon the strength of the signal used in the computation of IFC. Thus, ignoring components with high signal strength (corresponding to summing over low eigenvalues only in (20)) could lower the noise if the relationship between the noise variance and the signal variance is super-linear, for which an increase in signal strength would cause a *decrease* in the signal-to-noise ratio.

Another interesting observation is that when only the horizontal and vertical subbands are used in the computation of the IFC in (11) for the vector GSM IFC, the performance increases [3]. We first thought that this was due to the presence of JPEG distorted images in the database since the blocking artifact is represented more in the horizontal and vertical subbands than at other orientations. However, we discovered that the performance increase was consistent for *all* distortion types present in the database, and most notably for the JPEG2000 distortion. Also we do not get this increase in performance when we sum over other subbands; the performance in fact worsens. Table II gives the performance change of IFC on individual distortion types for horizontal and vertical subbands and the corresponding performance change when orientations of ± 60 degrees were summed in (11). We feel that this performance increase is due to the importance that the HVS gives to horizontal and vertical edge information in images in comparison with other orientations [34].

In our MATLAB implementation, the scalar GSM version of the IFC (without downsampling) takes about 10

[3]It does so for other IFC forms but we will not report those results here since they are mirrored by the ones presented.

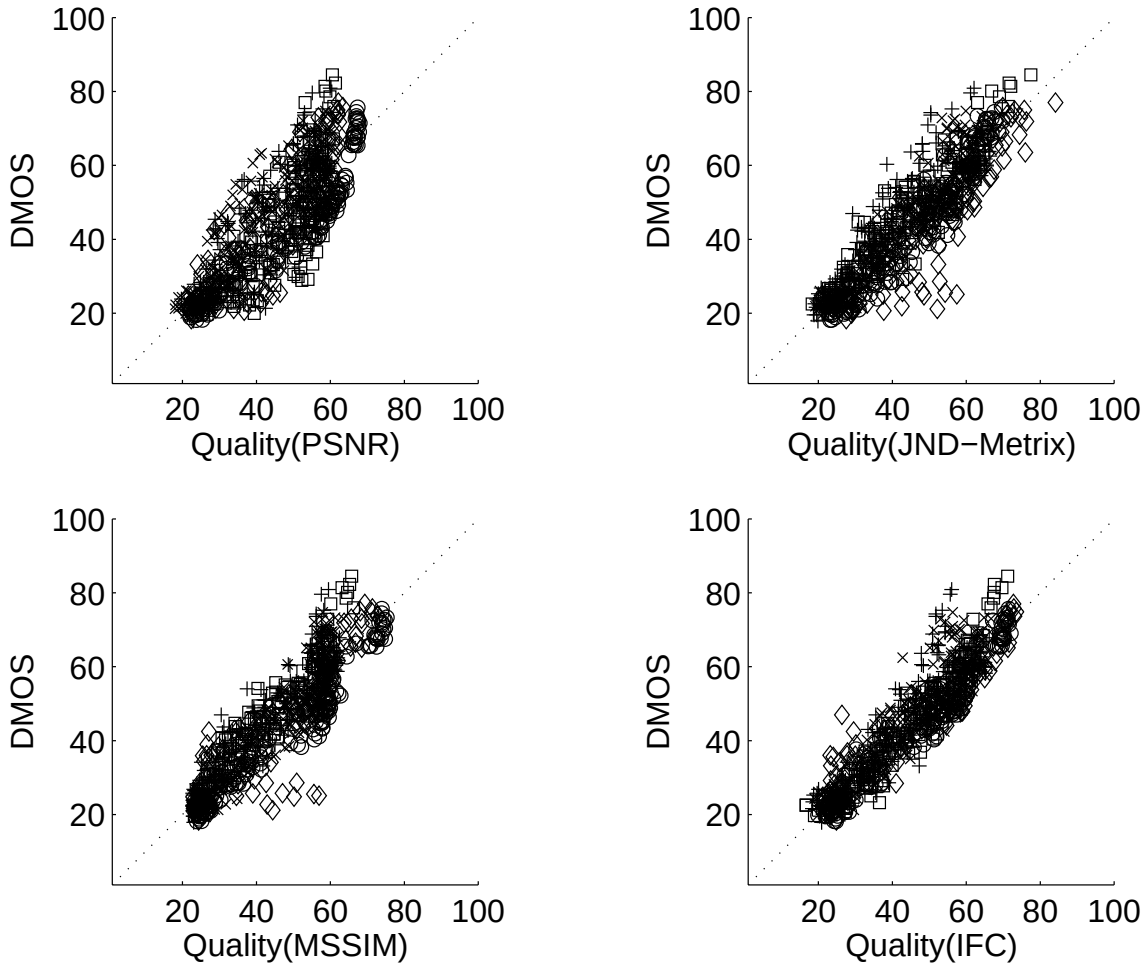

Fig. 2. Scatter plots for the quality predictions by the four methods after compensating for quality calibration: PSNR, Sarnoff's JND-metrix, MSSIM, and IFC for vector GSM. The IFC shown here uses only the horizontal and vertical subbands at the finest scale, and only the smallest eigenvalue in (20). The distortion types are: JPEG2000 (x), JPEG (+), white noise in RGB space (o), Gaussian blur (box), and transmission errors in JPEG2000 stream over fast-fading Rayleigh channel (diamond).

seconds for a 512×768 color image on a Pentium III 1 GHz machine. The vector GSM version (with horizontal and vertical subbands only) takes about 15 seconds. This includes the time required to perform color conversions, which is roughly 10% of the total time. We noted that about 40% to 50% of the time is needed for the computation of the wavelet decomposition.

We would like to point out the most salient feature of the IFC: it does not require any parameters from the HVS or viewing configuration, training data or stabilizing constants. In contrast, the JND-metrix requires a number of parameters for calibration such as viewing distance, display resolution, screen phosphor type, ambient lighting conditions etc. [40], and even SSIM requires two hand-optimized stabilizing constants. Despite being parameterless, the IFC outperforms both of these methods. It is reasonable to say that the performance of the IFC could improve further if these parameters, which are known to affect perceptual quality, were incorporated as well.

RMS in prediction against DMOS			
Distortion	All orientations	Hor./Vert.	± 60 deg.
JPEG2000	6.899	6.017	7.559
JPEG	6.542	6.237	6.927
White Noise	3.589	3.444	3.698
Gauss. Blur	4.166	3.873	4.521
Fast-fading	4.448	4.416	4.779

TABLE II

VALIDATION SCORES FOR THE VECTOR GSM IFC USING ALL ORIENTATIONS VERSUS USING: ONLY THE HORIZONTAL AND VERTICAL ORIENTATIONS, AND THE SUBBANDS ORIENTED AT ± 60 DEG. ONLY THE SMALLEST EIGENVALUE HAS BEEN USED IN (20) FOR GENERATING THIS TABLE.

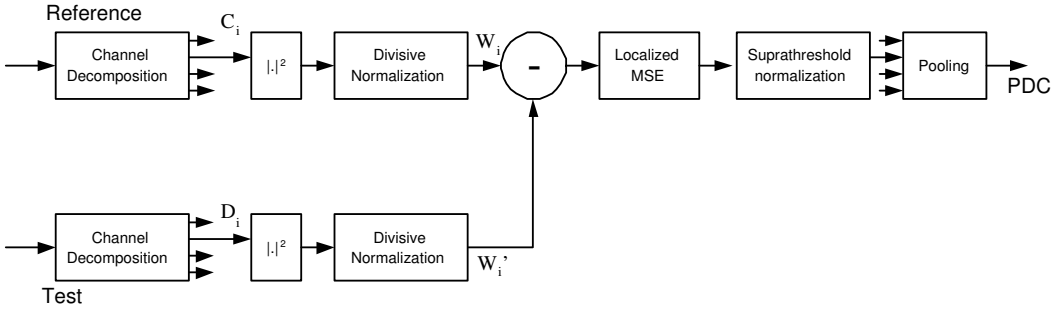

Fig. 3. An HVS based quality measurement system. We show that this HVS model is the dual of the scalar GSM based IFC of 11.

VII. SIMILARITIES WITH HVS BASED QA METHODS

We will now compare and contrast IFC with HVS based QA methods. Figure 3 shows an HVS based quality measurement system that computes the error signal between the processed reference and test signals, and then processes the error signal before computing the final perceptual distortion measure. A number of key similarities with most HVS based 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.

In the appendix we show the following relationship between the scalar version of the IFC in (10) and the HVS model of Figure 3 for one subband:

$$I(C^N; D^N|s^N) \quad \approx \quad \alpha \sum_{i=1}^{N} \log_2(\mathrm{MSE}(W_i, W_i'|s_i)) + \beta \tag{25}$$

where W_i and W_i' are as shown in Figure 3. The MSE computation in Figure 3 and (25) is a *localized* error strength measure. The logarithm term can be considered to be modeling of the suprathreshold effect. Suprathreshold effect is the name given to the fact that the same amount of distortion becomes perceptually less significant as the overall distortion level increases. Thus a change in MSE of, say, 1.0 to 2.0 would be more annoying than the same change

from 10.0 to 11.0. Researchers have previously modeled suprathreshold effects using visual impairment scales that map error strength measures through concave non-linearities, qualitatively similar to the logarithm mapping, so that they emphasize the error at higher quality [41]. Also, the pooling in (25) can be seen to be Minkowski pooling with exponent 1.0. Hence with the stated components, the IFC can be considered to be a particular HVS based quality assessment algorithm, the perceptual distortion criterion (PDC), within multiplicative and additive constants that could be absorbed into the calibration curve:

$$\text{PDC} = \sum_{k \in \text{subbands}} \sum_{i=1}^{N_k} \log_2(\text{MSE}(W_{k,i}, W'_{k,i}|s_{k,i})) \tag{26}$$

$$\text{IFC}_{\text{scalar}} \approx \alpha(\text{PDC}) + N_{sub}\beta \tag{27}$$

where k denotes the index of the k-th subband, and N_{sub} is the number of subbands used in the computation.

We can make the following observations regarding PDC of (26), which is the HVS dual of the IFC (using the scalar GSM model), in comparison with other HVS based FR QA methods:

- Some components of the HVS are not modeled in Figure 3 and (27), such as the optical point spread function and the contrast sensitivity function.
- The masking effect is modeled differently from some HVS based methods. While the divisive normalization mechanism for masking effect modeling has been employed by some QA methods [11]–[13], most methods divisively normalize the *error* signal with visibility thresholds that are dependent on neighborhood signal strength.
- Minkowski error pooling occurs in two stages: first a localized pooling in the computation of the localized MSE (with exponent 2) and then a global pooling after the suprathreshold modeling with an exponent of unity. Thus the perceptual error calculation is different from most methods, in that it happens in two stages with suprathreshold effects in between.
- In (26), the non-linearity that maps the MSE to a suprathreshold-MSE is a logarithmic non-linearity and it maps the MSE to a suprathreshold distortion that is later pooled into a quality score. Watson *et al*. have used threshold power functions to map objective distortion into *subjective* JND by use of two-alternative forced choice experiments [41]. However, their method applies the supratreshold non-linearity *after* pooling, as if the suprathreshold effect only comes into play at the global quality judgement level. The formulation in (26) suggests that the suprathreshold modeling should come *before* a global pooling stage but after localized pooling, and that it affects visual quality at a *local* level.
- One significant difference is that the IFC using the scalar GSM model, or the PDC of (26), which are duals of each other, is notably inferior to the vector GSM based IFC. We believe that this is primarily due to the underlying assumption about the uncorrelated nature of the wavelet coefficients being inaccurate. This dependence of perceptual quality on the correlation among coefficients is hard to investigate or model using HVS error sensitivities, but the task is greatly simplified by approaching the same problem with NSS modeling. Thus we feel that HVS based QA methods need to account for the fact that natural scenes are correlated within

subbands, and that this inter-coefficient correlation in the reference signal affects human perception of quality[4].

- Another significant difference between IFC/PDC and other HVS based methods is distinct modeling of signal attenuation. Other HVS based methods ignore signal gains and attenuations, constraining $\mathcal{G}$ to be unity, and treat such variations as additive signal errors as well. In contrast, a generalized gain g in the IFC/PDC ensures that signal gains are handled differently from additive noise components.
- One could conjecture that the conditioning on $\mathcal{S}$ in the IFC is paralleled in the HVS by the computation of the local variance and divisive normalization. Note that the high degree of self-correlation present in $\mathcal{S}$ enables its adequate estimation from $\mathcal{C}$ by local variance estimation. Since this divisive normalization occurs quite early in the HVS model [5] and since the visual signal is passed to the rest of the HVS after it has been *conditioned* by divisive normalization by the estimated s_i^2, we could hypothesize that the rest of the HVS analyzes the visual signal *conditioned on the prior knowledge of* $\mathcal{S}$, just as the IFC analyzes the mutual information between the test and the reference conditioned on the prior knowledge of $\mathcal{S}$.
- One question that should arise when one compares the IFC against the HVS error model is regarding HVS model parameters. Specifically, one should notice that while functionally the IFC captures HVS sensitivities, it does so without using actual HVS model parameters. We believe that some of the HVS model parameters were either incorporated into the calibration curve, or they did not affect performance significantly enough under the testing and validation experiments reported in this paper. Parameters such as the characteristics of the display devices or viewing configuration information could easily be understood to have approximately similar affect on all images for all subjects since the experimental conditions were approximately the same. Other parameters and model components, such as the optical point spread function or the contrast sensitivity function, which depend on viewing configuration parameters as well, are perhaps less significant for the scope and range of quality of our validation experiments. It is also reasonable to say that incorporating these parameters could further enhance the performance of IFC. We are continuing efforts into developing an IFC for a unified model that consists of source, distortion, and HVS models, and we feel that deeper insights into perception of quality would be gained.
- We would like to remind the readers at this point that although the IFC is similar to an HVS based distortion measure, it has *not* been derived using any HVS knowledge, and its derivation is completely independent. The similarities exist due to the similarities between NSS and HVS models. The difference is subtle, but profound!

VIII. Conclusions and Future Work

In this paper we presented an information fidelity criterion for image quality assessment using natural scene statistics. We showed that using signal source and distortion models, one could quantify the mutual information between the reference and the test images, and that this quantification, the information fidelity criterion, quantifies

[4]Equation (20) suggests that the same noise variance would cause a greater loss of information fidelity if the wavelet coefficients of the reference image were correlated than if they were uncorrelated.

[5]Divisive normalization has been discovered to be operational in the HVS [21].

perceptual quality. The IFC 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 showed that despite its competitive performance, the IFC is parameterless. We also showed that the IFC, under certain conditions, is quantitatively similar to an HVS based QA method, and we compared and contrasted the two approaches and hypothesized directions in which HVS based methods could be refined and improved.

We are continuing efforts into improving the IFC by combining HVS models with distortion and signal source models, incorporating color statistics, and inter-subband correlations. We are hopeful that this new approach will give new insights into visual perception of quality.

IX. Acknowledgements

The authors would like to thank Dr. Eero Simoncelli and Dr. Zhou Wang at the Center for Neural Science, New York University, for insightful comments.

Appendix

In this appendix we shall quantify the similarities between the scalar GSM version of the IFC of (10) and the HVS based QA assessment method shown in Figure 3. The model in Figure 3 is based on calculating MSE in the perceptual space and then processing it further to yield the final perceptual distortion measure. Here we will only deal with coefficients in one subband and a scalar GSM model.

We start by giving the formulation for the divisive normalization stage, which divides the input by its localized average. Considering the input to the squaring block, this turns out to be normalization by the estimated local variance of the input of the squaring block:

$$W_i = C_i^2 \left(\frac{1}{K} \sum_{j \in \mathcal{N}(i)} C_j^2 \right)^{-1} \approx \frac{C_i^2}{s_i^2} = U_i^2 \tag{28}$$

$$W_i' = D_i^2 \left(\frac{1}{K} \sum_{j \in \mathcal{N}(i)} D_j^2 \right)^{-1} \approx \frac{D_i^2}{g_i^2 s_i^2 + \sigma_V^2} \tag{29}$$

Here we have assumed that $s_j \approx s_i$ for $j \in \mathcal{N}(i)$, that is, the variance is approximately constant over the K pixels neighborhood of i, which we denote by $\mathcal{N}(i)$. Also note that the term inside the parentheses in an estimate of the conditional local variance of C (or D) at i given $S_i = s_i$, which could be approximated by the actual value. We have also assumed, without loss of generality, that $\mathrm{E}[U_i^2] = \sigma_U^2 = 1$, since any non-unity variance of $\mathcal{U}$ could be absorbed into $\mathcal{S}$. The MSE between W_i and W_i' given $S_i = s_i$ could now be analyzed:

$$\mathrm{MSE}(W_i, W_i'|s_i) = \mathrm{E}[(W_i' - W_i)^2 | s_i] \tag{30}$$

$$\approx \mathrm{E}\left[\left(\frac{D_i^2}{g_i^2 s_i^2 + \sigma_V^2} - U_i^2 \right)^2 | s_i \right] \tag{31}$$

$$= \mathrm{E}\left[\frac{\left(V_i^2 + 2 g_i C_i V_i - \sigma_V^2 U_i^2 \right)^2}{\left(g_i^2 s_i^2 + \sigma_V^2 \right)^2} | s_i \right] \tag{32}$$

where we have used $D_i = g_i C_i + V_i$ and that given $S_i = s_i$, $C_i = s_i U_i$. Expanding the above expression and taking expectation, and using independence between $\mathcal{U}$ and $\mathcal{V}$, the fact that $\mathcal{C}$, $\mathcal{U}$, and $\mathcal{V}$ are all zero-mean, and the fact that for zero-mean Gaussian variables $\mathrm{E}[X^4] = 3\sigma^4$, where σ^2 is the variance of X, we get:

$$\mathrm{MSE}(W_i, W_i'|s_i) \approx \frac{4\sigma_V^2}{g_i^2 s_i^2 + \sigma_V^2} \tag{33}$$

The goal of this derivation is to compare the information fidelity criterion of (10) and HVS based MSE criterion:

$$I(C^N; D^N|s^N) = \frac{1}{2}\sum_{i=1}^{N} \log_2\left(1 + \frac{g_i^2 s_i^2}{\sigma_V^2}\right) \tag{34}$$

$$= -\frac{1}{2}\sum_{i=1}^{N} \log_2\left(\frac{\sigma_V^2}{g_i^2 s_i^2 + \sigma_V^2}\right) \tag{35}$$

$$\approx -\frac{1}{2}\sum_{i=1}^{N} \left(\log_2(\mathrm{MSE}(W_i, W_i'|s_i)) - \log_2 4\right) \tag{36}$$

Hence we have an approximate relation between the information fidelity criterion and the HVS based MSE:

$$I(C^N; D^N|s^N) \approx \alpha \sum_{i=1}^{N} \log_2(\mathrm{MSE}(W_i, W_i'|s_i)) + \beta \tag{37}$$

where α and β are constants.

REFERENCES

[1] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE image quality assessment database," 2003, available at http://live.ece.utexas.edu/research/quality.

[2] 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.

[3] 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.

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

[5] 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.

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

[7] J. Lubin, "A visual discrimination mode 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.

[8] 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.

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

[10] 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.

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

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

[13] 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.

[14] 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.

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

[16] 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.

[17] VQEG: The Video Quality Experts Group,, *http://www.vqeg.org/*.

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

[19] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II," *ftp://ftp.its.bldrdoc.gov/dist/ituvidq/frtv2˙final˙report/VQEGII˙Final˙Report.pdf*, Aug. 2003.

[20] 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.

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

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

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] B. A. Wandell, *Foundations of Vision*. Sinauer Associates, Inc., 1995.

[35] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Processing*, vol. 4, no. 4, pp. 636–650, Apr. 2000.

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

[37] 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.

[38] 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.

[39] 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.

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

[41] A. B. Watson and L. Kreslake, “Measurement of visual impairment scales for digital video,” in *Human Vision, Visual Processing, and Digital Display, Proc. SPIE*, vol. 4299, 2001.