IMAGE INFORMATION AND VISUAL QUALITY

Hamid R. Sheikh and Alan C. Bovik

Laboratory for Image and Video Engineering, Department of Electrical and Computer Engineering,
The University of Texas at Austin, Austin, TX 78712-1084, USA.
Email: hamid.sheikh@ieee.org, bovik@ece.utexas.edu

ABSTRACT

Measurement of image quality is crucial for many image-processing algorithms. Traditionally, image quality assessment algorithms predict visual quality by comparing a distorted image against a reference image, typically by modeling the Human Visual System (HVS), or by using arbitrary signal fidelity criteria. In this paper we adopt a new paradigm for image quality assessment. We propose an information fidelity criterion that quantifies the Shannon information that is shared between the reference and the distorted images relative to the information contained in the reference image itself. We use Natural Scene Statistics (NSS) modeling in concert with an image degradation model and an HVS model. We demonstrate the performance of our algorithm by testing it on a data set of 779 images, and show that our method is competitive with state of the art quality assessment methods, and outperforms them in our simulations.

1. INTRODUCTION

Digital image and video processing systems are generally involved with signals that are meant to convey reproductions of visual information for 'human consumption'. Tradeoffs between system resources and the visual quality are typically involved in designing such systems, and accurate quality measurement algorithms are needed in order to make these tradeoffs efficiently. 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 to discover automatic ways of accurately measuring visual quality.

Traditionally, researchers have focussed on measuring quality by quantifying the similarity between a distorted (or test) image and a reference image that is assumed to have perfect quality. The Mean Squared Error (MSE), which is the L_2 norm of the arithmetic difference between the test and the reference images, is widely used to quantify (the loss of) visual quality. Unfortunately MSE, which is typically transformed into the Peak Signal to Noise Ratio (PSNR), does not correlate strongly enough with perceptual quality for most applications. In order to quantify the similarity between the test and the reference images in a perceptually meaningful manner, researchers have explored measuring error strength after processing the test and the reference images with HVS models. The underlying premise is that the sensitivities of the HVS 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. This *error sensitivity paradigm* is a *bottom-up* approach in which researchers model the low-level features of the HVS to achieve consistent quality predictions. Although such methods have met with good success, there are many questions that arise in their design [1]. Some researchers have therefore also explored arbitrary signal fidelity criteria that are not affected by assumptions about HVS models, but are motivated instead by the need to capture the loss of visual *structure* in the signal that the HVS hypothetically extracts for cognitive understanding. Such *top-down* methods have also met with good success [1]. A review of recent QA methods can be found in [2].

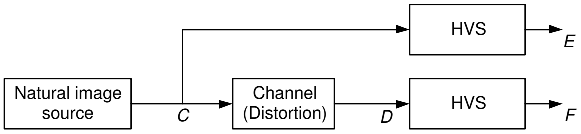

Fig. 1. Mutual information between $\mathcal{C}$ and $\mathcal{E}$ quantifies the information that the brain could ideally extract from the reference image, whereas the mutual information between $\mathcal{C}$ and $\mathcal{F}$ quantifies the corresponding information that could be extracted from the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

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*. Previously, we proposed using natural scene models in conjunction with distortion models to quantify the Shannon information shared between the test and the reference images, and showed that this shared image information is an aspect of fidelity that correlates well with visual quality [3]. In this paper we also quantify the information content of the reference image, since perceptual quality is likely to vary with the *relative* information loss, and propose a unified information fidelity criterion based on NSS, distortion and HVS modeling. This is shown pictorially in Figure 1.

H. R. Sheikh and A. C. Bovik are affiliated with the Laboratory for Image and Video Engineering, The University of Texas at Austin.

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

2. INFORMATION FIDELITY

Natural images of perfect quality can be modelled as the output of a stochastic source, which are then distorted by a 'channel' (the distortion operator) to give the test images. The mutual information between the test and the reference images is a measure of the information that is shared between the output of the channel and its input, and in the context of natural image sources it could quantify perceptual image fidelity [3]. We propose that this mutual information can also be compared against the information content of the reference image in order to quantify information fidelity *relative* to the information content of the reference image. We discuss the components of the proposed method in this section.

2.1. The Source Model

Natural scenes, that is, images and videos of the three dimensional visual environment captured using the visible spectrum, comprise only a tiny subset of the space of all possible signals. Many researchers have attempted to understand the structure of this subspace of natural images by studying their statistics. A good review on NSS models can be found in [4]. 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 [5]. 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.

The natural scene model that we use in this paper is the Gaussian scale mixture (GSM) model in the wavelet domain [6]. A GSM is a random field (RF) that can be expressed as a product of two independent RFs. 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 the RF $\mathcal{U}$ to be white. In this paper we model each subband of a scale-space-orientation wavelet decomposition (such as the steerable pyramid [7]) as a GSM. We partition the subband into non-overlapping blocks of M coefficients each, assume each block to be independent of others, and model each block as the vector $\overrightarrow{C}_i$. With such a construction, it is easy to make the following observations: $\mathcal{C}$ is normally distributed given $\mathcal{S}$, and that $\overrightarrow{C}_i$ are conditionally independent give $\mathcal{S}$ [6]. The GSM model has been shown to capture key statistical features of natural images, such as the heavy-tailed marginal distributions of, and the non-linear dependencies between, the wavelet coefficients of natural images [6].

We model each subband in the wavelet decomposition with a separate GSM. However, we will only deal with one subband here and later generalize the results for multiple subbands.

2.2. The Distortion Model

The distortion model that we use is a signal gain 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 (attenuation) 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 $\mathcal{U}$. This is a simple, yet effective, distortion model that works well for quality assessment purposes. It captures two important, and complementary, distortion types: white noise (by the noise RF $\mathcal{V}$) and blur (by measuring the loss in higher frequencies using the scalar attenuation field $\mathcal{G}$). We also believe that this model adequately reflects 'natural distortion operators' in response to which the HVS has evolved over eons. A more detailed motivation for this model has been presented in [3].

2.3. 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 [5], many aspects of the HVS are already modelled in the NSS description. The components missing include the optical point spread function, the contrast sensitivity function, internal neural noise etc. A more detailed comparison of quality assessment using NSS models versus HVS models has been discussed in [3]. We found from experiments that just modeling the internal neural noise gives marked improvement in performance in terms of the ability of the overall algorithm to predict quality.

The internal neural noise model that we use is an additive white Gaussian noise model. Thus we model the neural noise as the RF $\mathcal{N} = \{\overrightarrow{N}_i : i \in \mathrm{I}\}$, where $\overrightarrow{N}_i$ are zero-mean uncorrelated multivariate Gaussian with the same dimensionality as $\overrightarrow{C}_i$:

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

$$\mathcal{F} = \mathcal{D} + \mathcal{N} \quad \text{test image} \tag{4}$$

where $\mathcal{E}$ and $\mathcal{F}$ denote the visual signal at the output of the HVS model from the reference and the test images respectively, from which the brain extracts cognitive information. We model the covariance of the additive noise as:

$$\mathbf{C}_N = \sigma_n^2\mathbf{I} \tag{5}$$

where σ_n^2 is an HVS model parameter (variance of the internal neuron noise).

2.4. The Visual Information Fidelity Criterion

With the source and the distortion 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. In order to keep the problem analytically tractable, we will analyze the conditional mutual information between $\mathcal{C}$ and $\mathcal{E}$ (or $\mathcal{F}$) given $\mathcal{S}$. This is also in line with some divisive normalization based HVS models which process the signals after 'conditioning' them with local variance estimates [6].

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:

$$I(\overrightarrow{C}^N; \overrightarrow{E}^N | s^N) = \sum_{j=1}^{N} \sum_{i=1}^{N} I(\overrightarrow{C}_i; \overrightarrow{E}_j | \overrightarrow{C}^{i-1}, \overrightarrow{E}^{j-1}, s^N) \tag{6}$$

$$=\sum_{i=1}^{N} I(\overrightarrow{C}_i;\overrightarrow{E}_i|s_i) \tag{7}$$

$$=\sum_{i=1}^{N}(h(\overrightarrow{C}_i+\overrightarrow{N}_i|s_i)-h(\overrightarrow{N}_i|s_i)) \tag{8}$$

$$=\frac{1}{2}\sum_{i=1}^{N}\log_2\left(\frac{|s_i^2\mathbf{C_U}+\sigma_n^2\mathbf{I}|}{|\sigma_n^2\mathbf{I}|}\right) \tag{9}$$

where we get (6) from chain rule [8], and (7) from the conditional independence of $\mathcal{C}$ and $\mathcal{N}$ given $\mathcal{S}$. Similarly we can show that for the test image

$$I(\overrightarrow{C}^N;\overrightarrow{F}^N|s^N)$$

$$= \sum_{i=1}^{N}(h(g_i\overrightarrow{C}_i+\overrightarrow{V}_i+\overrightarrow{N}_i|s_i)-h(\overrightarrow{V}_i+\overrightarrow{N}_i|s_i)) \tag{10}$$

$$= \frac{1}{2}\sum_{i=1}^{N}\log_2\left(\frac{|g_i^2s_i^2\mathbf{C_U}+(\sigma_v^2+\sigma_n^2)\mathbf{I}|}{|(\sigma_v^2+\sigma_n^2)\mathbf{I}|}\right) \tag{11}$$

Since $\mathbf{C}_U$ is symmetric, it can be factored as $\mathbf{C}_U = \mathbf{Q\Lambda Q^T}$, 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:

$$I(\overrightarrow{C}^N;\overrightarrow{E}^N|s^N)=\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{M}\log_2\left(1+\frac{s_i^2\lambda_k}{\sigma_n^2}\right) \tag{12}$$

$$I(\overrightarrow{C}^N;\overrightarrow{F}^N|s^N)=\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{M}\log_2\left(1+\frac{g^2s_i^2\lambda_k}{\sigma_v^2+\sigma_n^2}\right) \tag{13}$$

$I(\overrightarrow{C}^N;\overrightarrow{E}^N|s^N)$ and $I(\overrightarrow{C}^N;\overrightarrow{F}^N|s^N)$ represent the information that can ideally be extracted by the brain from a particular subband in the reference and the test images respectively. The visual information fidelity measure that we propose in this paper is simply the fraction of the reference image information that could be extracted from the test image. 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 our visual information fidelity (VIF) measure is given by:

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

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.

3. IMPLEMENTATION

A number of implementation assumptions need to be made before the VIF of (14) could be implemented.

Assumptions about the source model. Note that mutual information (and hence the 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. For the vector GSM model, the maximum-likelihood estimate of s_i^2 can be found as follows [9]:

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

Estimation of the covariance matrix $\mathbf{C_U}$ is also straightforward from the reference image wavelet coefficients [9]:

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

In (15) and (16), $\mathrm{E}[S_i^2]$ is assumed to be unity without loss of generality [9].

Assumptions about the distortion model. 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 the variance of the RF $\mathcal{V}$ over block l, which we denote as $\sigma_{v,l}^2$, 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:

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

$$\widehat{\sigma_{v,l}^2}=\widehat{\mathrm{Cov}}(D,D)-\widehat{g_l}\widehat{\mathrm{Cov}}(C,D) \tag{18}$$

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

Assumptions about the HVS model. The parameter σ_n^2 in our simulations was hand optimized by running the algorithm using different values and observing the performance of the algorithm.

4. RESULTS

In this section we present the results of our simulations using the VIF proposed in (14). We compare the performance of our algorithm against the well known Sarnoff model [10] and SSIM [1].

4.1. Simulation Details

For the wavelet decomposition, we used the steerable pyramid with six orientations [7]. These wavelets have better orientation selectivity, as well as a loose shift-invariance, than the commonly used cartesian-separable wavelets. Vectors $\overrightarrow{C}_i$ and $\overrightarrow{D}_i$ were constructed from non-overlapping 3×3 neighborhoods, and the distortion model was estimated with 18×18 non-overlapping blocks. Only the horizontal and vertical subbands at the finest level were used in the summation of (14). The value of σ_n^2 used was 0.1. MSSIM (Mean SSIM) was calculated on the luminance component after decimating (filtering and downsampling) it by a factor of 4 [1]. The JND-Metrix was evaluated on full color images, whereas the VIF and SSIM operated upon the luminance component only.

4.2. Subjective Validation Study

We tested our algorithm's performance on an extensive subjective study. 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. Twenty-nine high-resolution 24-bits/pixel RGB color images (typically 768×512) were distorted using five distortion types: JPEG2000, JPEG, white noise

Validation against DMOS		
Model	CC	RMSE
PSNR	0.826	9.087
Sarnoff	0.901	6.992
MSSIM	0.911	6.629
VIF (proposed)	0.950	5.046

Table 1. Validation scores for different quality assessment methods: PSNR, JND-Metrix 8.0 [10], MSSIM [1], and the VIF.

in the RGB components, Gaussian blur, and transmission errors in the JPEG2000 bit stream using a fast-fading Rayleigh channel model. A total of 779 distorted images were derived. 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. The raw scores were converted to difference scores (between the test and the reference) and then converted to Z-scores and finally a Difference Mean Opinion Score (DMOS) for each distorted image.

4.3. Discussion

The performance metrics that we report here are the linear correlation coefficient (CC) and the root mean squared error (RMSE) between DMOS and the predicted DMOS. 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 of the QA algorithm. In this paper we use a five-parameter non-linearity (a logistic function with additive linear term) for all methods except for the VIF, for which we used the mapping on the logarithm of the VIF. The mapping used was:

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

$$f(\tau, x) = \frac{1}{2} - \frac{1}{1 + \exp(\tau x)} \qquad (20)$$

Table 1 gives these results, which are also shown in Figure 2 for the VIF. The VIF gives notably superior performance over MSSIM and Sarnoff's JND-Metrix. The improvement over Sarnoff's JND-Metrix in our testing is roughly the same as that of JND-Metrix over PSNR.

5. CONCLUSIONS

In this paper we have presented a novel visual information fidelity criterion that quantifies the Shannon information present in the distorted image relative to the information present in the reference image. We showed that VIF is a competitive way of measuring fidelity that relates well with visual quality. We validated the performance of our algorithm using an extensive study involving 779 images, and we showed that the proposed method is competitive with the state-of-the-art methods and outperforms them in our simulations. We are currently working on extending our work for video quality assessment as well.

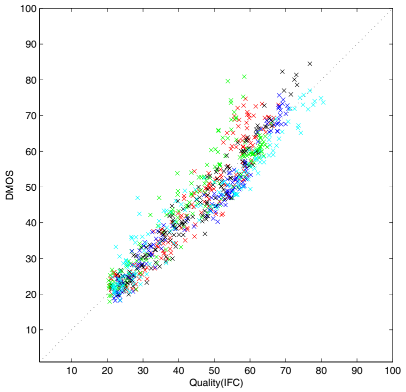

Fig. 2. Quality predictions by the VIF after compensating for quality calibration. The distortion types are: JPEG2000 (red), JPEG (green), white noise in RGB space (blue), Gaussian blur (black), and transmission errors in JPEG2000 stream (cyan).

6. REFERENCES

[1] 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*, Jan. 2004, To appear.

[2] 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.

[3] 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*, Sept. 2003, Submitted.

[4] 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.

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

[6] Martin J. Wainwright, Eero P. Simoncelli, and Alan 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.

[7] 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.

[8] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 1991.

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

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