A VISUAL INFORMATION FIDELITY APPROACH TO VIDEO QUALITY ASSESSMENT

Hamid R. Sheikh

Texas Instruments Inc., USA.
Email: hamid.sheikh@ieee.org

Alan C. Bovik

The University of Texas at Austin, USA.
Email: bovik@ece.utexas.edu

ABSTRACT

Measurement of visual quality is crucial for many image and video processing applications. Traditionally, quality assessment (QA) algorithms predict visual quality by comparing a distorted signal against a reference, typically by modeling the Human Visual System (HVS). In this paper, we adopt a new paradigm for video quality assessment that is an extension of our previous work on still image QA. We propose an information fidelity criterion that quantifies the Shannon information that is shared between the reference and the distorted videos relative to the information contained in the reference video 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 the VQEG Phase I dataset, and show that the information-fidelity framework is competitive with state of the art quality assessment methods.

1 Introduction

Measurement of visual quality is becoming increasingly important in many image and video processing applications, such as acquisition, compression, communication, restoration, enhancement and reproduction. The goal of quality assessment (QA) methods is to assess the quality of images and videos in a perceptually consistent manner and in close agreement with subjective human judgments.

Traditionally, researchers have focused mainly on measuring visual quality of videos by modeling the salient features of the human visual system (HVS) using the so-called full reference (FR) QA paradigm. In the FRQA framework, the algorithm measures the quality of a distorted (or test) video against a reference video that is assumed to have perfect quality. Since the mean squared error (MSE) between the test and the reference videos is not a good measure of visual quality, FRQA algorithms typically measure the distance between the test and reference signals in some perceptual space using HVS models. A review of recent video QA methods can be found in [1].

We previously proposed a novel information fidelity approach to image quality assessment, which is an information-theoretic framework using natural scene statistics (NSS) models [2]. Images and videos of the 3-D visual environment belong to a common class: the class of natural scenes. Due to the nature of image formation, the class of natural scenes is an extremely tiny subset of the set of all possible signals, and researchers have developed sophisticated models to characterize the statistical properties of this class. Most real-world distortions disturb these statistics and make the signals unnatural. In the information-fidelity framework, we use a common statistical fidelity measure, the mutual information, to quantify this unnaturalness. We have previously shown that this new framework outperforms current state-of-the-art still image quality assessment algorithms by a sizeable margin [2]. This paper presents extensions of our framework for still image QA to video QA. As a product of this extension we derive a simple video QA algorithm based on natural scene statistics whose performance is competitive with state-of-the-art video QA methods, and which outperforms the proponents in VQEG Phase-I study.

H. R. Sheikh was previously affiliated with the Dept. of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA. This work was supported by a grant from the National Science Foundation.

2 Visual Information Fidelity

We previously proposed a novel method for quantifying still image quality, the information fidelity paradigm, which is an information theoretic framework based on NSS models [2]. The setup is shown in Figure 1. Natural images and videos of perfect quality (the reference signals) are modeled as the output of a stochastic source. In the absence of any distortions, this signal passes through the HVS channel and is received by cognitive processes in the brain. The distorted signals, however, pass through another channel, such as a compression process or blurring, before it passes through the HVS and finally into the receiver. The success of the visual communication process in the presence of distortion intuitively relates to the amount of information that can be extracted by the brain from the distorted signal (that is the information flowing through the lower path in Figure 1) *relative* to the information that can be extracted from the reference signal (the information flowing through the upper path in Figure 1).

We previously proposed using mutual information as a measure of statistical information for quality assessment of still images. The visual information fidelity (VIF) measure for video QA that we propose in this paper is also derived from a quantification of two mutual information quantities: the mutual information between the input and the output of the HVS channel when no distortion is present (we call this the reference image information) and the mutual information between the input of the distortion channel and the output of the HVS channel for the test signal. In order to quantify the mutual information quantities, we need stochastic models for the source, distortion, and the HVS. Below we will outline the models that we use in this paper.

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 [3]. The model that we use in this paper is the Gaussian scale mixture (GSM) model in the spa-

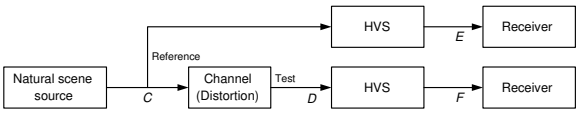

Fig. 1. Mutual information between $\mathcal{C}$ and $\mathcal{E}$ quantifies the information that the brain could ideally extract from the reference signal, whereas the mutual information between $\mathcal{C}$ and $\mathcal{F}$ quantifies the corresponding information that could be extracted from the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

tiotemporal derivative (horizontal, vertical, and temporal discrete-derivatives [4]) domain. We have previously successfully used the GSM model in the wavelet domain for image quality assessment using a similar information fidelity setup [2]. In order to use this model for video, we first need to motivate the suitability of this model for videos.

2.1.1 Image Formation

Researchers have argued that the peculiar statistics of natural images come from the physics of image formation [5, 6]. The laws that govern the formation of light stimulus that emanates from the environment dictate the statistics of the signal. Researchers have identified *occlusion* as the image formation rule that leads to many of the observed statistics of natural images. That is, images are formed when objects occlude each other, generally blocking the light signal coming from objects 'behind them'. This leads to images that contain smooth or textured regions separated by strong edges at object boundaries. Thus, when such an image is filtered through a zero-mean kernel (a high-pass or a band-pass filter), the resulting coefficients tend to have a histogram that has a sharp peak at zero and tails that fall off slowly. The smooth textures give rise to generally low magnitude coefficients (and hence the sharp peak at zero in the histogram), while the strong edges yield high magnitude coefficients (heavy tails) typically observed in natural images.

2.1.2 Occlusion in Natural Videos

It is obvious that occlusion continues to operate even for natural videos. Thus, applying *spatial* zero-mean kernels would still give the characteristic histograms. Moreover, it is easy to explain why such kernels would give similar histograms even when applied along the temporal domain. Consider two one-dimensional objects in Figure 2 that are being observed through a camera in a two dimensional world. When object A moves and occludes object B, one could see occlusion of B by A in a spatiotemporal plane much like the occlusion of objects in still images. The spatiotemporal plane consists of generally smooth 'objects' separated by sharp edges that result from 'occlusion' of spatiotemporal objects, and hence the temporal channel should exhibit similar statistical properties as the spatial channels. The argument could easily be extended to a three dimensional case.

Indeed, this hypothesis holds when one observes the histograms of temporally filtered natural videos. Figure 3 shows the histogram of the temporal derivative of 'Mobile and Calendar'. As expected, the histogram has a sharp peak at zero and heavy tails. Moreover, the temporal coefficients also exhibit linear and nonlinear dependencies similar to those observed in spatial coefficients of images [7]. This is illustrated in Figure 4, which shows the joint histogram of the logarithm of the magnitude of two temporally adjacent coefficients C_1 and C_2 in the temporal-derivative channel[1]. The linear dependency between the logarithm of the coefficient magnitudes demonstrates a multiplicative dependency between the coefficient magnitudes, which can be modeled well though a GSM model [7]. Moreover, divisively normalizing the coefficients by estimates of local standard deviation removes this dependency, as shown in Figure 4. We therefore conclude that the spatial and temporal channels of natural videos (such as those resulting from discrete derivatives for optical flow estimation or other 3D transformations such as wavelet decompositions) could be modeled well by GSMs.

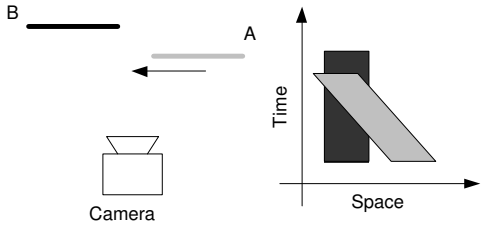

Fig. 2. Spatiotemporal representation of motion. Object A moves and occludes B, which is stationary. Occlusion manifests in the spatiotemporal domain as spatiotemporal objects 'occluding' each other, similar to spatial-only images.

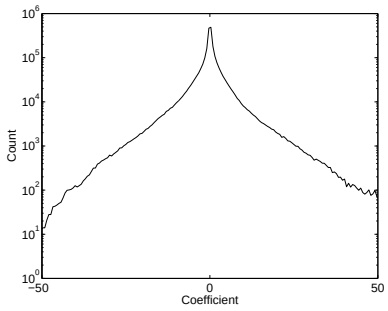

Fig. 3. Histogram of coefficients of a bandpass channel of 'Mobile and Calendar' video shows a sharp peak and heavy tails.

2.1.3 Representation of Motion

The use of intensity derivatives for optical flow estimation is widely known in the image processing community. Researchers have also explored possible connections between motion representation in HVS and spatiotemporal derivatives and their approximation using bandpass kernels [8]. In the information-fidelity framework, although we do not deal directly with motion representation, derivative information does form a bound on motion information. Any loss in derivative information (which we shall quantify later in this paper) would necessarily cause a loss in motion information in the signal.

2.1.4 Gaussian Scale Mixture Model

A GSM is a random field (RF) that can be expressed as a product of two independent RFs [7] . 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

[1]Linear dependency between C_1 and C_2 was removed before plotting the histograms to highlight the nonlinear dependency.

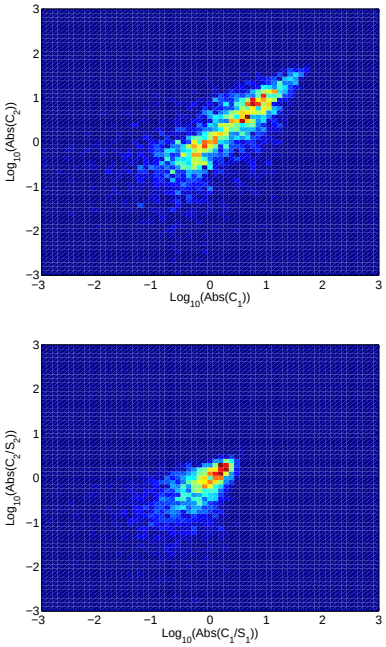

Fig. 4. Joint histograms of temporally adjacent coefficients before and after normalization. See text.

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. In this paper we model each channel (output of a spatiotemporal kernel) as a GSM. We partition a channel into non-overlapping spatiotemporal blocks, and model each block as the vector $\overrightarrow{C}_i$.

We model each channel with a separate GSM. However, we will only deal with one channel here and later generalize the results for multiple channels.

2.2 The Distortion Model

The distortion model that we use is a signal gain and additive noise model:

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

where $\mathcal{C}$ denotes the RF from a channel in the reference signal, $\mathcal{D} = \{\overrightarrow{D}_i : \ i \in \mathrm{I}\}$ denotes the RF from the corresponding channel 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 simple, yet effective, distortion model has been shown to work well for still image QA purposes [2]. It captures two important, and complementary, distortion types: noise (by the noise RF $\mathcal{V}$) and blur (by measuring the loss in higher frequencies using the scalar attenuation field $\mathcal{G}$). Moreover, the GSM source model in concert with the gain-and-additive-noise distortion model can be shown to be approximately equivalent to the contrast gain control model of visual masking [9].

2.3 The Human Visual System Model

The HVS model that we use is also described separately for each channel. Since HVS models are the dual of NSS models [10], many aspects of the HVS are already modeled in the NSS description. The components missing include the optical point spread function, the contrast sensitivity function, internal neural noise etc. In the information-fidelity setup, an HVS model is needed to serve as a *distortion baseline* against which the distortion added by the distortion channel could be compared. In such we observed that lumping all sources of the uncertainty that the HVS adds into a *visual noise* component serves as an adequate model of HVS uncertainty for QA purposes. The visual noise model that we use is an additive white Gaussian noise model, where we model the visual 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 $\overrightarrow{C}_i$:

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

$$\mathcal{F} = \mathcal{D} + \mathcal{N}' \ \text{ test signal} \tag{4}$$

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

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

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

2.4 The Visual Information Fidelity Criterion

With the source and the distortion models as described above, the visual information fidelity (VIF) criterion that we propose can be derived. In fact, using the generalization provided by the vector models, the formulation of VIF for video remains the same as that for still images [2]. Thus, assuming that the model parameters $\mathcal{G}$, σ_v^2 and σ_n^2 are known, as well as the underlying variance field $\mathcal{S}$, the reference and distorted image information is given by:

$$I(\overrightarrow{C}^N; \overrightarrow{E}^N | s^N) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{M}\log_2\left(1 + \frac{s_i^2\lambda_k}{\sigma_n^2}\right) \tag{6}$$

$$I(\overrightarrow{C}^N; \overrightarrow{F}^N | s^N) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{M}\log_2\left(1 + \frac{g_i^2 s_i^2\lambda_k}{\sigma_v^2 + \sigma_n^2}\right) \tag{7}$$

where i is an index to a spatiotemporal block and λ_k are the eigenvalues of $\mathbf{C}_U$. Note that as in [2], we consider only the conditional mutual information between $\mathcal{C}$ and $\mathcal{E}$ (or $\mathcal{F}$) given $\mathcal{S}$. The reason for this conditioning is to *tune* the GSM model for a particular reference signal because we are interested in measuring the quality of a particular reference-test pair and not the 'quality' of a distortion channel for the whole ensemble of natural signals. Thus the given field $S_i = s_i$ becomes the model parameters of a set of independent but not identically distributed vector Gaussian random variables whose covariance at index i is given by $s_i^2\mathbf{C}_U$.

$I(\overrightarrow{C}^N; \overrightarrow{E}^N | s^N)$ and $I(\overrightarrow{C}^N; \overrightarrow{F}^N | s^N)$ represent the information that can ideally be extracted by the brain from a particular channel in the reference and the test videos respectively. The visual information fidelity measure is simply the fraction of the reference image information that could be extracted from the test signal. Also we have only dealt with one channel so far. One could easily incorporate multiple channels by assuming that each channel 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:

$$\text{VIF}=\frac{\sum_{j\in\text{channels}} I(\vec{C}^{N,j};\vec{F}^{N,j}|s^{N,j})}{\sum_{j\in\text{channels}} I(\vec{C}^{N,j};\vec{E}^{N,j}|s^{N,j})} \quad (8)$$

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

3 Implementation

A number of implementation assumptions need to be made before the VIF of (8) could be implemented.

Assumptions about the source model. Mutual information (and hence the VIF) can only be calculated between RF's and not their *realizations*, that is, a particular reference/test video pair 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. A number of known estimation methods are available for estimating s_i^2 and $\mathbf{C}_U$ (for example [11]).

Assumptions about the distortion model. We propose to partition the channel into spatiotemporal 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 [2].

Assumptions about the HVS model. The parameter σ_n^2 in our simulations was hand optimized over a few values by running the algorithm and observing its performance.

4 Results

In this section we present the results of our simulations using the VIF in (8) using the data from VQEG Phase-I study. In our simulations, we use separable derivative kernels of length 5 in the spatial directions (horizontal and vertical) and 3 in the temporal direction. The channels that we use in (8) are the derivative channels on all three components in the YUV color space, a total of nine channels. The vectors C_i were constructed from $3 \times 3 \times 2$ blocks. The implementation runs only on one field of the interlaced frames. The results are given in Table 1.

It can be seen that even with a very simple implementation using separable kernels and only one level of decomposition, the VIF outperforms the highest performing proponent in the VQEG study [12]. Moreover, four of the 20 source sequences in the dataset are unnatural videos (computer generated or scrolling text), and Table 1 also gives the results when these sequences are excluded from testing.

5 Conclusions

In this paper we have presented a novel visual information fidelity criterion that quantifies the Shannon information present in the distorted video relative to the information present in the reference video. 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 the VQEG Phase-I dataset, and showed that the proposed method is competitive with the state-of-the-art methods and outperforms them in our simulations.

	CC	SROCC
PSNR ([12])	0.779	0.786
VQEG P8 ([12])	0.827	0.803
VIF	0.874	0.849
VIF (natural only)	0.891	0.865

Table 1. VIF performance compared against PSNR and the best performing proponent in [12] (P8) on all data using linear correlation coefficient (CC) after nonlinear regression with logistic function and the Spearman rank order correlation coefficient (SROCC) [12].

6 References

[1] 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.

[2] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Processing*, Sept. 2004, Accepted.

[3] 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.

[4] H. Farid and E. Simoncelli, "Differentiation of multidimensional signals," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 496–508, 2004.

[5] Ann B. Lee, David Mumford, and Jinggang Huang, "Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model," *International Journal of Computer Vision*, vol. 41, no. 1/2, pp. 35–59, 2001.

[6] D. L. Donoho and A. G. Flesia, "Can recent innovations in harmonic analysis 'explain' key findings in natural image statistics," *Vision Research*, vol. 12, no. 3, pp. 371–393, 2001.

[7] Martin J. Wainwright and Eero P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," *Advances in Neural Information Processing Systems*, vol. 12, pp. 855–861, 2000.

[8] Eero P. Simoncelli, "Vision and the statistics of the visual environment," *Current Opinion in Neurobiology*, vol. 13, Apr. 2003.

[9] 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*, Mar. 2004, Accepted.

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

[11] 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.

[12] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," `http://www.vqeg.org/`, Mar. 2000.