AN INFORMATION THEORETIC VIDEO QUALITY METRIC BASED ON MOTION MODELS

Kalpana Seshadrinathan and Alan C. Bovik

Department of Electrical and Computer Engineering,
1, University Station C0803, The University of Texas at Austin, Austin, TX - 78712

ABSTRACT

Accurate objective quality metrics are of great potential benefit to the video industry, as they promise the means to evaluate the performance of acquisition, display, coding and communication systems. Although the area of image quality assessment has attained maturity in recent years, video quality assessment still has a long way to go to before it reaches the levels of success achieved by still image quality metrics. In this paper, we propose a novel quality metric for video sequences, which we call the Video Information Fidelity Criterion (V-IFC), that utilizes motion information in video sequences, which is the main difference in moving from images to video. We previously proposed a model that describes the statistics of natural video sequences and this model is used in the development of V-IFC. This metric is capable of capturing temporal artifacts in video sequences, in addition to spatial distortions. Results are presented that demonstrate the efficacy of our quality metric by comparing model performance against subjective scores on the database developed by the Video Quality Experts Group (VQEG).

Index Terms— Quality Assessment, Video Signal Processing, motion compensation, Video Quality Experts Group (VQEG), Information Fidelity

1. INTRODUCTION

With the rapid increase in popularity of multimedia applications such as Video On Demand, wireless video, digital cinema etc., it is critical to be able to monitor the quality of video as it passes through distortion channels. Distortion channels are created due to the processing of the video sequences by common operations such as compression, channel coding, transmission errors, error concealment and decoding etc. In an overwhelming majority of applications, the end-user of the video sequence is a human observer. It is hence of interest to evaluate the quality of a video sequence, *as seen by a human observer*, as opposed to generic distortion measures that are commonly used for any data signal such as Mean Square Error (MSE). Video Quality Assessment (VQA) algorithms attempt to assess *perceptual degradations* introduced by any signal processing operations performed on video sequences. Unfortunately, despite rapid advances in video processing and communication technology, the performance of video quality assessment algorithms remains poor and there is considerable room for improvement [1].

Although progress in the development of accurate and reliable VQA algorithms has been slow, great strides have recently been made in assessing the quality of still images [2, 3]. In this paper, we develop a full reference quality metric for video signals by making natural extensions of the powerful information fidelity framework for still images to the spatio-temporal (video) domain. Full reference quality metrics assume the availability of a "perfect" reference video and attempt to assess the fidelity of the test video with respect to this pristine original.

Most of the research on VQA in the literature has focused on methods that attempt to model the Human Visual System (HVS). The approach adopted by HVS-based metrics is to process the video data using models that simulate the initial stages of the visual pathway. Various quality metrics developed for quality assessment differ in the aspects of the Human Visual System (HVS) that are chosen to be incorporated in the quality assessment system, as well as the computational model that is used to describe these effects. Examples of video quality metrics based on the HVS-based philosophy include the Digital Video Quality (DVQ) metric [4], the Sarnoff JND model [5] and the Perceptual Distortion Model (PDM) [6]. However, studies conducted by the Video Quality Experts Group indicate that the performance of HVS-based VQA algorithms leaves considerable room for improvement [1]. HVS-based VQA metrics suffer from inaccurate modeling of the HVS. In particular, inadequate modeling of temporal mechanisms in the HVS play a key role in the performance loss of video quality metrics, as opposed to still image quality metrics. For example, all of the VQA metrics mentioned above use either one or two temporal channels and only model the temporal tuning of the neurons in area V1 of the visual cortex. These models are too simple to describe motion processing in the HVS. In particular, activity of neurons in area MT of the extra-striate cortex, which play a very important role in motion perception, is not accounted for in any of these models.

Preliminary extensions of the information fidelity approach has been proposed for VQA [7] using a simple implementation of the still image quality metric on spatio-temporal video blocks. However, this metric does not utilize *motion information* or model temporal artifacts in video that can affect the quality of the video sequence. The human eye is quite sensitive to motion and can accurately judge the velocity and direction of moving objects, skills that are essential to the survival of an organism. Considerable resources in the HVS are devoted to motion perception and it is hence essential for video quality metrics to incorporate some form of motion modeling. Further, video sequences suffer from *spatio-temporal* artifacts and quality metrics must take into consideration temporal distortions in videos. Example of such temporal artifacts include ghosting, jitter, motion compensation mismatch, smearing, mosquito noise etc. Furthermore, the model presented in [7] uses a natural scene model that was proposed for natural *images* to model the scene statistics of natural *video* and the accuracy of such a model remains open to question.

We believe that the performance of video quality assessment techniques can be improved by the introduction of meaningful models that describe motion in video sequences, as well as model spatio-temporal distortions in the video stream. To our knowledge, none of the quality metrics proposed in the literature explicitly model motion

or temporal artifacts in video sequences. The novelty of this work lies in the use of motion models in predicting visual quality. In this paper, we present a Video Information Fidelity Criterion, known as V-IFC, that incorporates motion modeling using optical flow, which results in a *motion compensated* implementation of the IFC for still images [2]. We then demonstrate the performance of our metric on the VQEG database that contains distorted sequences as well as subjective scores assigned by human observers to these sequences [1].

2. MOTION IN THE FREQUENCY DOMAIN

In this paper, we consider the apparent motion of image intensities, namely the *optical flow*. The term velocity denotes the optical flow vector and not the true three dimensional velocity of motion. Let $i(x, y)$ denote an image and let $\tilde{I}(u, v)$ denote its Fourier transform, where (x, y) denotes the spatial axes and (u, v) denotes the spatial frequency axes. Assuming that this image undergoes translation with a velocity $\vec{\lambda} = (\lambda_x, \lambda_y)$, the resulting video sequence is given by $l(x, y, t) = i(x - v_x t, y - v_y t)$. Then, $\tilde{L}(u, v, w)$, the Fourier transform of $l(x, y, t)$, lies entirely along a plane in the frequency domain [8]. This plane is defined by:

$$\lambda_x u + \lambda_y v + w = 0$$

Additionally, the frequencies along this plane in the spatio-temporal frequency domain are *identical* to the spatial frequencies in the image $i(x, y)$.

In line with the assumptions used in several video compression standards, including MPEG-2 and H.264, we assume that short segments of video consist of local image patches undergoing translation, which is a reasonable approximation as long as there are no scene changes. This model can be used *locally* to describe video sequences, since translation is a linear approximation to more complex types of motion. Frequency domain approaches are also well suited to our study of human perception of video signals due to the presence of bandpass visual channels in the HVS [6]. Hence, in the proposed V-IFC video quality assessment system, the video sequence is spatio-temporally filtered using a family of band-pass filters. We use a model for the statistics of these sub-band filtered coefficients that we have developed previously to assess the quality of video sequences [9]. This model is summarized in the next section.

2.1. Statistical Model for Video

We previously proposed a model that describes the statistics of natural video sequences, which we briefly summarize here [9]. Translational motion of local image patches was combined with a statistical model for natural images in the frequency domain in the development of our model.

The wavelet coefficients of natural (still) images exhibit strong dependencies along neighboring spatial locations, scales, orientations. The Gaussian Scale Mixture (GSM) model nicely describes the wavelet coefficients distributions of natural still images [10, 11], which motivated the development of our information theoretic metrics for still image quality assessment [2, 3]. A continuous-frequency version of the GSM model can be posed as follows. Given a naturalistic image i with Fourier Transform $\tilde{I}$, we make the local model

$$\tilde{I}(u, v) \sim zU(u, v) \tag{1}$$

for all (u, v) in any sub-band of an oriented scale-space decomposition, where z is the mixing density and U is a complex, zero-mean, white Gaussian random field. Assume that an image patch that is described by these statistics undergoes translation with velocity $\vec{\lambda} = (\lambda_x, \lambda_y)$. Next, consider methods of capturing local video statistics using this simple model. Filter the video signal $l(x, y, t)$ with a family of spatio-spectrally localized 3-D sub-band filters $f_i(x, y, t) \leftrightarrow \tilde{F}_i(u, v, w)$, resulting in sub-band filtered coefficients $c_i(x, y, t)$. Assuming the mixing density is estimated, the coefficients $c_i(x, y, t)$ conditioned on the mixing density $\hat{z}_i(x, y, t)$ are zero-mean Gaussian random variables with variances [9]:

$$\sigma_i^C = (2\pi)^{-6} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{z}_i^2 |\tilde{F}_i(u, v, -\lambda_x u - \lambda_y v)|^2 \mathrm{d}u\mathrm{d}v \tag{2}$$

where $\tilde{F}(u, v, -\lambda_x u - \lambda_y v)$ is a 2-D slice of the filter along the plane containing the spectrum of the translating video signal. The dependence of σ_i, $\hat{z}_i$ and $\vec{\lambda}$ on (x, y, t) is dropped for notational convenience. From this, it is apparent that large-magnitude coefficients will appear where the energy of the variance field is large and where the oriented plane significantly intersects the filter pass-bands. Additionally, $c_i(x_1, y_1)$ is independent of $c_i(x_2, y_2)$ when conditioned on $\hat{z}$, since $U(x, y)$ was assumed to be white.

The z field is often modeled as gamma-distributed [10]; however, we do not assume any prior distribution of z and estimate it using a wavelet analysis of local signal energy, denoting the estimate using $\hat{z}$. Further details can be found in [9].

3. IFC INDEX FOR IMAGES

The IFC index was first developed for still images [2] and since the design of our metric closely follows this development, we briefly overview it here. The reference and test images are first passed through a scale-space oriented decomposition to generate filtered coefficients for the reference and test images.

Let $\vec{C}_i(x)$ denote a set of coefficients at adjacent spatial locations (an $R \times R$ window of coefficients, for example) of the reference image. Here, the index i denotes a single filter from the entire family used in the decomposition and we assume N filters, i.e. $i = 1, 2, \ldots N$. x denotes a spatial index for the vector of coefficients and $x = 1, 2, \ldots M$. In the IFC framework, this vector $\vec{C}_i(x)$ is assumed to be modeled well using the GSM model developed for natural images [10]. Thus, $\vec{C}_i(x)$ can be modeled as a Gaussian random vector of zero mean and covariance $\mathbf{C}_U$, conditioned on the estimated value of the mixing density $\hat{z}_i(x)$. In addition, the distorted image is assumed to be generated from the reference image using a simple blur and additive noise distortion model (also known as channel model). Let $\vec{D}_i(x)$ denote a set of coefficients from the distorted image at corresponding spatial locations to those chosen from the reference image. Then, the channel model is given by:

$$\vec{D}_i(x) = G_i(x)\vec{C}_i(x) + \vec{N}_i(x) \tag{3}$$

where $G_i(x)$ denotes a deterministic scalar gain field and $\vec{N}_i(x)$ is assumed to be an Additive White Gaussian Noise (AWGN) field with covariance matrix $\sigma_i^N(x)\mathbf{I}$.

Then, the IFC index between these coefficients is given by the *mutual information* between these vector fields and can be shown to be:

$$\mathrm{IFC}(\vec{C}(x), \vec{D}(x)) = \frac{1}{2}\log_2\left(1 + \frac{|G_i(x)^2 \hat{z}_i^2 \mathbf{C_U} + \sigma_i^N \mathbf{I}|}{|\sigma_i^N \mathbf{I}|}\right) \tag{4}$$

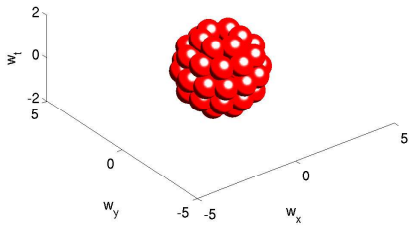

Fig. 1. Geometry of the Gabor filterbank in the frequency domain.

The overall quality index of the entire image is then calculated as the sum of the IFC indices over all values of x. This quality measure was shown to perform very well in predicting the quality of still images [2].

4. V-IFC INDEX FOR VIDEO SEQUENCES

4.1. Selection of sub-band filter family

In Section 2, we discussed the simple form that motion in video sequences takes in the frequency domain. This motivated us to develop the statistical model for sub-band filtered coefficients of video sequences. Decomposition of the video into bandpass channels in the frequency domain helps us achieve two goals, namely optical flow estimation and quality computation, both of which are accomplished using the outputs of these bandpass filters.

Although any filter family can be used to decompose the video sequence into bandpass channels, we opt to use Gabor filters in our implementation. Gabor filters attain the theoretical lower bound on the uncertainty in the frequency and spatial variables and thus, visual neurons can be said to optimize the uncertainty in information resolution [12]. Additionally, development of the video quality metric in Section 4.3 requires estimation of the optical flow vectors. Gabor filters have been successfully used for this purpose in the literature [13].

A Gabor filter $f(x, y, t)$ is simply the product of a Gaussian window and a complex exponential:

$$f(x,y,t) = \frac{1}{(\sqrt{2\pi})^3 \sigma_x \sigma_y \sigma_t} e^{-\left(\frac{x^2}{2\sigma_x^2}+\frac{y^2}{2\sigma_y^2}+\frac{t^2}{2\sigma_t^2}\right)} e^{i(Ux+Vy+Wt)} \tag{5}$$

where (U, V, W) is the center frequency of the Gabor filter and $(\sigma_x, \sigma_y, \sigma_t)$ is the spread of the Gaussian window in space-time. Then, the Fourier transform of the Gabor filter is a Gaussian whose standard deviation in the frequency domain is $(1/\sigma_x, 1/\sigma_y, 1/\sigma_t)$.

$$\tilde{F}(w_x, w_y, w_t) = e^{-\frac{1}{2}\left[\sigma_x^2(w_x-U)^2+\sigma_y^2(w_y-V)^2+\sigma_t^2(w_t-W)^2\right]} \tag{6}$$

The filters used in our implementation have the same geometry as the Gabor filters described in [13] and are illustrated in Figure 1. We used a family of filters consisting of $N = 22$ filters all at the same scale, i.e., all filters are tuned to the same spatio-temporal frequency band. Figure 1 shows the isosurface contours of the resulting filter bank in the frequency domain. The spatial spread of the Gaussian filters is the same along all axes and hence, the iso-surface contours are spherical.

4.2. Optical flow estimation

The proposed V-IFC algorithm uses motion information from the reference video sequence in the form of the optical flow vectors. We briefly describe the optical flow estimation algorithm. We used the Fleet and Jepson phase based algorithm for optical flow estimation with slight modifications [13]. The Fleet and Jepson algorithm is designed under the assumption that the evolution of phase contours of bandpass filtered outputs closely approximates the projected motion field. Constant phase contours are computed using the derivatives of the Gabor filter outputs, which is computed using a 5-point central difference kernel in [13]. However, we chose to perform the derivative computation by convolving the video sequence with filters that are derivatives of the Gabor kernels denoted by $f'_x(x, y, t), f'_y(x, y, t), f'_y(x, y, t)$.

$$f'_x(x,y,t) = f(x,y,t)\left(\frac{-x}{\sigma_x^2} + iU\right) \tag{7}$$

Similar definitions apply for the derivatives along y and t directions. This filter is more accurate in computing the derivative of the Gabor outputs, and produced better optical flow estimates in our experiments. We wish to point out that the Fleet and Jepson algorithm does not produce flow estimates with 100% density, i.e. flow estimates are not computed at each and every pixel of the video sequence. Finally, note that our current implementation uses only one scale of filters and cannot compute optical flow in fast moving regions of the video sequence due to temporal aliasing [13].

4.3. Proposed quality index for video sequences

Motion plays a very important role in the human perception of video sequences. Distorted videos suffer from artifacts that are *spatio-temporal* as described in Section 1. The main drawback of most video quality metrics in the literature, including the information theoretic quality metric for video developed earlier [7], was the failure to model motion or temporal artifacts in video sequences. In Section 4.2, we described a method to estimate the motion in a video sequence that has been proposed in the literature. In this section, we will use this statistical model just developed to develop a novel information theoretic video quality metric, that closely follows the development of the IFC metric for still images

Let $\{C_i(x), i = 1, 2, \ldots N\}$ denote the output of Gabor filter $f_i(x, y, t)$ operating on the reference video sequence. Note that x denotes a *spatio-temporal* index that represents the location of the sub-band filtered coefficients. Similarly, let $\{D_i(x), i = 1, 2, \ldots N\}$ denote coefficients at the corresponding spatio-temporal location obtained by filtering the distorted sequence, whose quality we wish to estimate, with the Gabor filter $f_i(x, y, t)$.

From the statistical model presented in Section 2.1, we know that $C_i(x)$ is distributed as a zero-mean Gaussian random variable with variance given by Eq. 2. We denote this resulting variance field by $\sigma_i^C(x)$. The main difference between the our quality metric and that presented in [7] is the introduction of this statistical model which was derived specifically for video sequences under the commonly used assumption of local translation. Although the heuristic model used in [7] was shown to work reasonably well by illustration, it was not tested systematically. Additionally, the model in [7] did not incorporate any motion information from the video sequence.

Similar to the IFC paradigm, we assume that the coefficients of the distorted video sequence are obtained by applying a distortion operator on the reference video coefficients. This distortion channel is modeled using a blur and additive noise model and is given by:

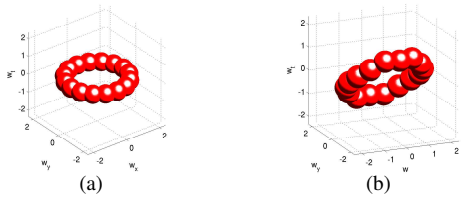

Fig. 2. Illustration of a set of motion compensated filters: (a) A static sequence (b) Sequence with motion.

$$D_i(x) = G_i(x)C_i(x) + N_i(x) \tag{8}$$

where $G_i(x)$ is a deterministic gain field and $N_i(x)$ is an AWGN field.Thus, $E[N_i(x)N_i(y)] = 0 \quad \forall \quad x \neq y$ and $E(N_i(x)^2) = \sigma_i^N(x)$. This distortion model is capable of handling both spatial as well as *temporal* distortions in the video sequence. Gabor filters form *spatio-temporal* bandpass channels in the frequency domain, whose iso-surface contours are ellipsoidal in shape. Assuming translational motion in the video sequence, the spectrum of the video sequence will lie along a plane. The orientation of this plane is defined by the optical flow vector $\vec{\lambda}$ and the frequency components along this plane are determined by the spatial frequency components of the image patch undergoing translation. Thus, distortions in the video that are *purely spatial*, i.e. intra-frame distortions, will result in changes in the frequency components along the plane, which will be captured by the Gabor filter outputs. Examples of such spatial distortions include blurring, blocking and ringing caused by compression, errors during acquisition, transmission through communication channels etc. Distortions in the video that are *purely temporal*, i.e. inter-frame distortions, will result in a change in the axis along which the plane intersects the Gabor filter. Examples of temporal distortions include motion compensation mismatch and mosquito noise due to compression, ghosting and temporal aliasing during acquisition, transmission through communication channels etc. Our distortion model captures both complementary forms of distortion and is hence, capable of handling a wide variety of distortion operators.

It is also instructive to note the differences between our proposed model and the distortion operator presented in [7]. Although a similar blur and additive noise model was proposed in [7], that distortion model was applied on the derivatives of the video sequence in the *pixel domain*. Thus, the model does not allow for an intuitive explanation of the kinds of distortions that it can handle.

We now define the quality index of the distorted sequence as the mutual information between the coefficients of the reference and distorted video sequences that are modeled as Gaussian random fields, conditioned on the estimated values of the mixing density $\hat{z}_i(x)$ and the gain field $G_i(x)$. It is fairly straightforward to compute this mutual information and the V-IFC index is hence defined by:

$$\text{V-IFC}(C_i(x), D_i(x)) = \frac{1}{2}\log_2\left(1 + \frac{G_i(x)^2\sigma_i^C(x)}{\sigma_i^N(x)}\right) \tag{9}$$

We now modify the IFC framework further to develop a *motion-compensated* implementation of the proposed quality metric. The optical flow computation on the reference sequence provides us with an estimate of the local orientation of the plane containing the frequency spectrum of the video sequence. We then identify the Gabor filters that overlap significantly with this plane. Although in the original IFC framework, quality indices are computed for *all* the Gabor filters, we define a selection criterion and only compute V-IFC indices for the outputs of the filters that satisfy this criterion.

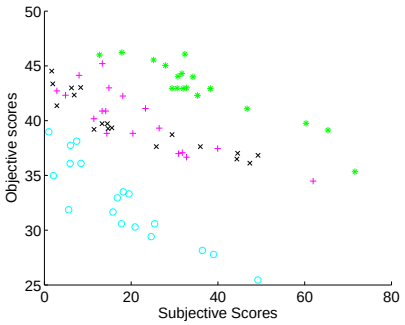

Fig. 3. Illustration of the dependence of the range of V-IFC values on the reference sequence. Each marker represents data points from a different reference video.

In our implementation, we require that the plane lie within one standard deviation of the Gabor filter in the frequency domain. Thus, if the optical flow vector at a pixel is (λ_x, λ_y) and the center frequency of the Gabor filter is (U_0, V_0, W_0), then the plane that contains the spectrum of the video sequence is described by $v_x w_x + v_y w_t + w_t = 0$. Thus, our rule for selection of the filter is:

$$S = \left\{k : \left|\frac{\lambda_x U + \lambda_y V + W}{\sqrt{\lambda_x^2 + \lambda_y^2 + 1}}\right| \leq \frac{1}{\sigma}\right\} \tag{10}$$

where S denotes a set that contains the selected filter indices and σ denotes the standard deviation of the Gabor filter along any axis in the space domain, since we use spherical Gabor filters. This filter selection procedure is illustrated in Fig. 2. Fig. 2(a) shows the filters that are selected for a static video sequence that undergoes no motion and consists of the same image repeated over frames. Fig. 2(b) illustrates the filters selected for a hypothetical sequence undergoing translation.

We hypothesize that our proposed metric is capable of handling a wide variety of both spatial as well as temporal artifacts. Note that in the absence of temporal artifacts, the proposed metric is simply a motion compensated implementation of the IFC metric for still images. This is a desirable property. Furthermore, motion compensation provides us with a way to model temporal distortions in the video. This is because the proposed system results in the filtering of the distorted sequence along the motion trajectories of the reference video sequence and estimating IFC indices using these filter outputs.

4.4. Implementation Details

The optical flow estimates described in Section 4.2 provide estimates of (λ_x, λ_y). Estimation details of the parameter $\hat{z}$ were also described in Section 2.1. The parameters $G_i(x)$ and $\sigma_i^N(x)$ were estimated by computing the least squares regression fit between coefficients extracted using a $7 \times 7 \times 7$ window from $C_i(x)$ and $D_i(x)$ centered on the pixel location x.

The integral in Eq. 2 can be evaluated in closed form. This integral was computed for complex Gabor filters in [14]. We used sine phase Gabor filters, which have zero DC response, in our implementation and evaluated this integral.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\tilde{F}_s(u,v,-\lambda_x u-\lambda_y v)|^2 \mathrm{d}u\mathrm{d}v = K(T_1 - T_2)$$

$$T_1 = e^{-\left(-\sigma_u^2\sigma_v^2\sigma_w^2 \frac{(\lambda_x U+\lambda_y V+W)^2}{r}\right)}$$

$$T_2 = e^{-(U^2\sigma_u^2+V^2\sigma_v^2+W^2\sigma_w^2)}$$

$$r = \lambda_x^2\sigma_v^2\sigma_w^2 + \lambda_y^2\sigma_u^2\sigma_w^2 + \sigma_u^2\sigma_v^2$$

$$K = \frac{0.5\pi}{\sqrt{r}}$$

where $\tilde{F}_s(u,v,w)$ denotes the Fourier transform of the sine phase Gabor filter, (U,V,W) denotes the center frequency of the filter and $(\sigma_u,\sigma_v,\sigma_w)$ denotes the standard deviation of the Gabor filter in the frequency domain.

Our flow estimation algorithm does not produce flow estimates at each pixel of the video sequence. At pixels without motion information, we simply set $\lambda_x = \lambda_y = 0$. This results in the computation of V-IFC indices that capture spatial distortions alone at these pixels. Additionally, to avoid computing IFC indices in low signal-to-noise regions, which may occur when the signal energy is insignificant inside the span of the Gabor filter, we computed IFC indices only at pixels where the magnitude of the response was at least 5% of the maximum response of the filter to the frame that contained the pixel.

5. RESULTS

We tested our proposed V-IFC index on the VQEG database [1]. This database contains 20 reference video sequences, test sequences obtained by distorting each of these reference videos with 16 different distortion operations and subjective scores for all test sequences [1]. The distorted sequences are further sub-divided into a low quality and high quality data set and for each reference sequence, two of the distorted versions are included in both the high and low quality data sets. We chose to treat the resulting subjective scores in the low and high quality tests as independent data points in computing the correlation coefficients. The current implementation of our optical flow estimation uses filters at just one scale. The results we present are for 16 of the 20 reference sequences, since the flow estimation algorithms failed to produce outputs for 4 sequences that were fast moving. These excluded sequences were sequence 6 (Formula 1 racing car), sequence 8 (scrolling text), sequence 9 (rugby game) and sequence 19 (football game). The VQEG test sequences are interlaced and similar to the approach adopted in [7], our algorithm operates only on the odd fields of the interlaced sequences. To reduce the computational burden, flow and V-IFC indices were not computed for each frame, but only for one in 16 frames.

We tested our metric on the remaining 16 reference sequences with 288 data points by computing the Spearman Rank Order Correlation Coefficient (SROCC) between subjective and objective scores for different video quality metrics. SROCC is one of the metrics specified by the VQEG that tests the prediction monotonicity of a video quality assessment system.

On analyzing our results, we noted that the range of values that the V-IFC index takes depended on the reference sequence. This is illustrated in Fig. 3. Each marker symbol denotes the data points obtained from all distorted versions of the same reference sequence. We show data for four different sequences from the VQEG database in this figure. We believe that the reason for this is the fact that the IFC is not a normalized metric [2]. The V-IFC, as specified in Eq.(9), depends not only on the distortion parameters, but also on the variance or energy of the coefficients of the reference image, σ_i^C. We believe that this behaviour of the IFC metric can be rectified by developing a normalized information theoretic measure, similar to [3] in the future. To overcome this problem, we fitted the objective and subjective scores of each reference sequence independently using a logistic function specified in [2]. The SROCC was then computed between the subjective and objective scores after passing the V-IFC scores through this optimal logistic function.

Prediction Model	SROCC
Peak Signal to Noise Ratio	0.786
Proponent P8 (Swisscom)	0.803
Structural Distortion Measurement [15]	0.812
Visual Information Fidelity [7]	0.849
Proposed V-IFC Metric	0.876

Table 1. Comparison of SROCC values for different video quality assessment algorithms.

The results of our experiments are summarized in Table 1, which shows the SROCC values for different metrics. PSNR does not correlate well with subjective scores as seen in Table 1. Proponent P8 is the best performing metric amongst the 10 different proponent models tested by the VQEG in terms of the SROCC metric [1]. We compared our results against the better performing version of the two metrics proposed in [15]. We also compared our results against the metric presented in [7]. The results clearly indicate that our V-IFC index performs quite well and is competitive with other video quality assessment systems. Although these preliminary results are promising, it is important to remember that the SROCC values reported here are somewhat optimistic since the scores were fitted individually for each reference sequence. In some sense, this reduces the burden on the quality assessment algorithm to predict visual quality well across *different content* in the video database.

6. CONCLUSIONS AND FUTURE WORK

In conclusion, we presented a novel information theoretic quality metric for video sequences, that uses a statistical model for video that we developed previously as well as a novel distortion model to predict subjective quality of video data. We validated our model by testing it on the VQEG - Phase I FR-TV database and showed that our metric is competitive with other state-of-the-art video quality metrics. In the future, we would like to seek improvements in our optical flow estimation techniques. Additionally, we would like to evaluate other distortion models that directly model the distortion in the flow field of the video.

7. REFERENCES

[1] (2000) Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI

[2] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *Image Processing, IEEE Transactions on*, vol. 14, no. 12, pp. 2117–2128, 2005.

[3] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *Image Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 430–444, 2006.

[4] A. B. Watson, J. Hu, and J. F. McGowan III, "Digital video quality metric based on human vision," *J. Electron. Imaging*, vol. 10, no. 1, pp. 20–29, Jan. 2001.

[5] (2003) Sarnoff corporation, JNDMetrix Technology. [Online]. Available: http://www.sarnoff.com/products services/video vision/ jndmetrix/downloads.asp

[6] S. Winkler, "Perceptual distortion metric for digital color video," in *Proc. SPIE Int. Soc. Opt. Eng.*, vol. 3644, no. 1. San Jose, CA, USA: SPIE, May 1999, pp. 175–184.

[7] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, AZ, Jan. 23-25 2005.

[8] A. B. Watson and J. Ahumada, A. J., "Model of human visual-motion sensing," *Journal of the Optical Society of America A (Optics and Image Science)*, vol. 2, no. 2, pp. 322–342, 1985.

[9] K. Seshadrinathan and A. C. Bovik, "Statistical video models and their application to quality assessment," in *Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, AZ, Jan. 23-26 2006.

[10] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," in *Advances in Neural Information Processing Systems*, S. A. Solla, T. Leen, and S.-R. Muller, Eds., vol. 12, 1999, pp. 855–861.

[11] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *Image Processing, IEEE Transactions on*, vol. 8, no. 12, pp. 1688–1701, 1999.

[12] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America A (Optics and Image Science)*, vol. 2, no. 7, pp. 1160–1169, 1985.

[13] D. Fleet and A. Jepson, "Computation of component image velocity from local phase information," *International Journal of Computer Vision*, vol. 5, no. 1, pp. 77–104, 1990.

[14] D. J. Heeger, "Optical flow using spatiotemporal filters," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 279–302, 1987.

[15] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–132, Feb. 2004.