Motion-based Perceptual Quality Assessment of Video

Kalpana Seshadrinathan and Alan C. Bovik

Laboratory for Image and Video Engineering
Department of Electrical and Computer Engineering
The University of Texas at Austin, Austin, TX - USA.

ABSTRACT

There is a great deal of interest in methods to assess the perceptual quality of a video sequence in a full reference framework. Motion plays an important role in human perception of video and videos suffer from several artifacts that have to deal with inaccuracies in the representation of motion in the test video compared to the reference. However, existing algorithms to measure video quality focus primarily on capturing spatial artifacts in the video signal, and are inadequate at modeling motion perception and capturing temporal artifacts in videos. We present an objective, full reference video quality index known as the MOtion-based Video Integrity Evaluation (MOVIE) index that integrates both spatial and temporal aspects of distortion assessment. MOVIE explicitly uses motion information from the reference video and evaluates the quality of the test video along the motion trajectories of the reference video. The performance of MOVIE is evaluated using the VQEG FR-TV Phase I dataset and MOVIE is shown to be competitive with, and even out-perform, existing video quality assessment systems.

Keywords: video quality, quality assessment, full reference, MOVIE

1. INTRODUCTION

The digital video revolution has caused widespread demand and use of consumer video appliances (cameras, camcorders, DVD players, set-top boxes, video enabled handheld devices, displays, video game consoles) and consumer video applications (digital television broadcasting, interactive Video on Demand (VoD), streaming video over IP networks, mobile video, video tele-conferencing, gaming) where digital videos are being delivered to humans. Due to the large plethora of applications that target human end users, automatic methods to determine the perceptual or visual quality of a video signal become important in quality monitoring, achieving Quality of Service (QoS) requirements and designing the video acquisition, communication or display system. This paper focuses on the full reference paradigm of video quality assessment (VQA), wherein the availability of a perfect quality, pristine reference video is assumed in addition to the test video whose visual quality is to be determined. The test video is assumed to be the result of an unknown distortion process operating on the pristine reference.

Section 2 presents a survey of existing full reference VQA algorithms. Videos typically suffer from spatial artifacts (that alter spatial aspects of a scene and are visible in individual frames of the video) and temporal artifacts (that alter the motion of pixels and are visible *across* frames of the video in time) that are described in Section 3. Existing algorithms to measure video quality focus primarily on capturing spatial artifacts in the video signal, and are inadequate at capturing temporal artifacts in videos. However, motion plays an important role in human perception of video due to effects such as visual perception of speed and direction of motion, visual tracking of moving objects and motion saliency. Further, although videos do suffer from spatial distortions such as blur, blockiness etc., several commonly occurring distortions in video such as motion compensation mismatch, jitter and ghosting have a temporal component. It is critical that objective VQA algorithms are able to account for the perceptual effects of both spatial and temporal distortions.

We seek to advance the VQA state-of-the-art by developing a full reference framework and an algorithm for VQA known as the MOtion-based Video Integrity Evaluation (MOVIE) index. MOVIE integrates both spatial and temporal aspects of distortion assessment and uses a spatio-temporally localized, multi-scale decomposition of the reference and test videos using a set of spatio-temporal Gabor filters. One component of the MOVIE framework known as the Spatial MOVIE index uses the output of the multi-scale decomposition of the reference and test videos to measure spatial distortions in the video. Spatial MOVIE is conceptually similar to several existing VQA systems. The key contribution of this work is the second component of the MOVIE index known as the Temporal MOVIE index, which captures temporal degradations in the video that form a critical component of visual perception of video quality. The Temporal MOVIE index computes and uses motion information from the reference video explicitly in quality measurement, and evaluates the quality of the test video along the motion trajectories of the reference video. Finally, the Spatial MOVIE index and the Temporal MOVIE index are combined to obtain a single measure of video quality known as the MOVIE index. The proposed MOVIE algorithm is described in detail in Section 4.

The MOVIE index is shown to deliver scores that correlate quite closely with human subjective judgment, using the Video Quality Expert Group (VQEG) FR-TV Phase 1 database as a test bed in Section 5. MOVIE is found to be quite competitive with, and even out-perform existing VQA systems. Finally, we conclude this paper in Section 6 with a brief discussion of future work.

2. BACKGROUND

The simplest, and perhaps most commonly used VQA index, is the Mean Squared Error (MSE) or equivalently the Peak Signal to Noise Ratio (PSNR) computed between the reference and test videos. The widespread use of PSNR even today despite its well documented failings as a visual quality index can be attributed to its simplicity and mathematical convenience.[1–3]

A large body of work has focused on using models of the various stages of processing that occur in the human eye-brain pathway in constructing an objective VQA index. This approach to VQA is intuitive since the goal of objective VQA systems is to match visual performance in predicting quality. Several popular VQA systems have adopted this approach.[4–8] HVS-based VQA algorithms have largely been derived from corresponding models for quality assessment of still images. For this reason, HVS-based systems typically include rather elaborate mechanisms to capture spatial aspects of human vision and distortion perception. However, HVS-based VQA systems do not do an adequate job in modeling temporal aspects of human vision and video distortions. Typically, HVS-based VQA systems only model the temporal mechanisms that occur in the early stages of processing in the visual cortex using either one or two kinds of linear filters - one lowpass and one bandpass - that are applied separably along the temporal dimension of the video after the spatial decomposition.[4–8] However, motion perception in the HVS is a complex visual task that is believed to occur in multiple stages of processing in the visual pathway and models of the temporal tuning of neurons in the visual cortex that are modeled in current HVS-based systems represent just the first stage of processing in motion computation.[9] It is well known that Visual Area MT/V5 of the extra-striate cortex plays an important role in motion perception and is believed to play a role in integrating local motion information into a global percept of motion, guidance of some eye movements, segmentation and structure computation in 3-dimensional space.[10] However, to the best of our knowledge, none of the existing HVS-based VQA systems attempt to model neuronal processing in Area MT to assess video quality. The importance of many of the functions of Area MT descibed above in the visual perception of video quality is obvious and motivates the need for better modeling of temporal aspects of human vision in VQA systems.

Recent years have seen a paradigm shift in VQA from algorithms that are based on models of the HVS to algorithms that utilize certain statistics or features computed from the reference and test videos in predicting visual quality. We classify such systems as feature based or signal statistic based VQA algorithms. For instance, features such as average brightness, contrast, edges, textures, color, blockiness and blur are computed from the reference and test videos and difference measures between feature values computed from the reference and test videos serve as indicators of visual quality. Several popular algorithms utilize a signal statistic based approach to VQA.[11–15] One of the prominent VQA algorithms that belongs to this class that deserves mention is the Video Quality Metric (VQM) developed at the National Telecommunications and Information Administration (NTIA).[14] VQM has been standardized by the American National Standards Institute (ANSI) and have been included as a normative method in two International Telecommunications Union (ITU) recommendations.[16] However, signal statistic based VQA algorithms suffer from similar drawbacks as HVS-based VQA systems and do not do an adequate job in capturing temporal aspects of human vision and distortion perception. While significant efforts are directed toward capturing spatial distortions in the video, features that are used to capture temporal degradations in the video are fairly limited. Some VQA systems operate frame-by-frame on the video[11] or utilize rudimentary temporal features such as adjacent frame differences[14] or normalized cross correlation between adjacent frames.[15] Such an approach does not do justice to the complexity of motion processing in the HVS and the significant role it plays in visual perception of video quality. We present an approach in this paper that explicitly utilizes motion models in a VQA framework and that can account for both spatial and temporal aspects of human vision and distortion assessment. Before we present this approach, we first discuss some commonly occurring spatial and temporal distortions in video in Section 3 to shed some insight on the challenging context of VQA.

3. VIDEO DISTORTIONS

Digital videos often undergo various processing stages such as compression and transmission through error prone communication channels, before reaching the ultimate receiver (the human eye in applications we consider).[17] A number of commonly occurring distortions in video may be classified as "spatial distortions", by which we refer to distortions that are visible in individual frames of the video and that alter the spatial aspects of various objects in a scene. This includes artifacts such as blocking, blur, ringing, mosaic patterns, false contouring and additive noise. Note that still images also suffer from these spatial distortions. An important distinction in moving from quality assessment of still images to videos is the fact that videos suffer from "temporal distortions", in addition to these spatial distortions. By temporal distortions, we refer to distortions that arise mainly from the occurrence of motion. These include artifacts such as motion compensation mismatch, jitter, ghosting and smearing that alter the movement trajectories of pixels in the test video relative to the reference. Other temporal distortions include mosquito noise and stationary area fluctuations that create a false perception of motion in certain regions of the test video that are stationary in the reference. Current VQA systems discussed in Section 2 focus on capturing spatial distortions in the video, but fail to do an adequate job in capturing temporal distortions in videos.

The discussion in Section 2 highlighted the fact that existing VQA algorithms do not do an adequate job in modeling the temporal aspects of human vision. The perceived visual annoyance of temporal distortions in video discussed in this section is closely tied to motion processing in the HVS. We will now present our proposed MOVIE index that simultaneously accounts for motion processing in the HVS and captures temporal distortions in videos (in addition to spatial distortions).

4. MOVIE INDEX

In this section, we detail the MOVIE algorithm. The MOVIE index operates by first decomposing the reference and test videos into multiple spatio-temporal bandpass channels using a family of Gabor filters. The resulting representation is then used to define the Spatial MOVIE index and the Temporal MOVIE index, each of which primarily captures spatial and temporal distortions in the video respectively. Finally, the two indices are combined to produce an overall video integrity score that we call the MOVIE index.

4.1 Linear Decomposition

Frequency domain approaches are well suited to the study of human perception of video signals and form the backbone of most VQA systems. This stage is often intended to mimic similar processing that occurs in the HVS since neurons in the visual cortex and the extra-striate cortex are spatial frequency and orientation selective.[18–20] In addition, a large number of neurons in the striate cortex, as well as Area MT which is devoted to movement perception, are known to be directionally selective; i.e., neurons respond best to a stimulus moving in a particular direction. Thus, both spatial characteristics and movement information in a video sequence are captured by a linear spatio-temporal decomposition.

In the MOVIE framework, we opt to use the Gabor basis functions for the linear decomposition. The responses of several spatio-temporally separable responses can be combined to encode the local speed and direction of motion of the video sequence.[21, 22] Spatio-temporal Gabor filters have been used in several models of the response of motion selective neurons in the visual cortex.[21, 23, 24] In our implementation of the ideas described here, we utilize the algorithm described by Fleet *et. al.*[25] that uses the outputs of a Gabor filter family to estimate motion. Thus, the same set of Gabor filtered outputs is used for motion estimation and for quality computation.

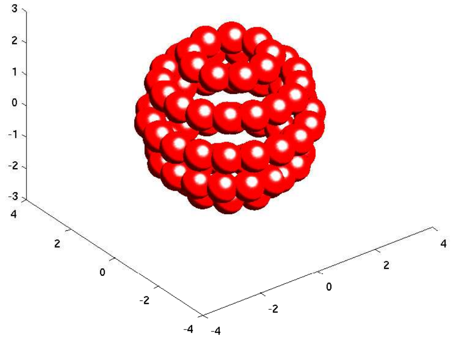

Figure 1. Geometry of the Gabor filterbank in the frequency domain. The figure shows iso-surface contours of all Gabor filters at the finest scale. The two horizontal axes denote the spatial frequency coordinates and the vertical axis denotes temporal frequency.

Our implementation uses separable Gabor filters that have equal standard deviations along both spatial frequency coordinates and the temporal coordinate. The design of the Gabor filter family used in MOVIE is very similar to that used by Fleet *et. al.* with the following distinctions.[25] A single scale of Gabor filters was used by Fleet *et. al.*,[25] which we found inadequate. Firstly, motion was not detected in fast moving regions of commonly occurring videos using a single scale of filters due to temporal aliasing.[25] Secondly, a multi-scale linear filter family is better suited for VQA since human perception of videos and video distortions is inherently a multi-scale process. Our implementation uses $P = 3$ scales of filters and as an additional contribution, we extended the single scale motion estimation algorithm of Fleet *et. al.* to multiple scales. Further, our filters have a narrower bandwidth of 0.5 octaves measured at one standard deviation of the Gabor frequency response, compared to the broader tuning of the Gabor filters used by Fleet *et. al.*[25] spanning 0.8 octaves. The Gabor filters are designed to intersect at one standard deviation of the frequency response resulting in a total of $N = 35$ filters at each scale.

Figure 1 shows iso-surface contours of the sine phase component of the filters tuned to the finest scale in the resulting filter bank in the frequency domain. The filters at coarser scales would appear as concentric spheres inside the sphere depicted in Fig. 1. Finally, a Gaussian filter is included at the center of the sphere to capture low frequencies in the reference and test videos and the standard deviation of the Gaussian is designed such that it intersects the coarsest scale of filters at one standard deviation of the frequency response.

4.2 Spatial MOVIE

Our approach to capturing spatial distortions in video of the kind described in Section 3 is inspired both by the Structural SIMilarity (SSIM) index and the information theoretic indices that have been developed for quality assessment of still images.[2, 13, 26] However, we will be using the outputs of the *spatio-temporal* Gabor filters to accomplish this. Hence, the model described here primarily captures spatial distortions in the video and at the same time, responds to temporal distortions in a limited fashion.

Let $r(\mathbf{i})$ and $d(\mathbf{i})$ denote the reference and distorted videos respectively, where $\mathbf{i}$ represents a pixel location in the video. The reference and distorted videos are passed through the Gabor filterbank to obtain bandpass filtered videos. Denote the Gabor filtered reference video by $\tilde{f}(\mathbf{i}, k)$ and the Gabor filtered distorted video by $\tilde{g}(\mathbf{i}, k)$, where $k = 1, 2, \ldots, K$ indexes the filters in the Gabor filterbank. Specifically, let $k = 1, 2, \ldots \frac{K}{P}$ correspond to the finest scale, $k = \frac{K}{P} + 1, \ldots, \frac{2K}{P}$ the second finest scale and so on.

All quality computations begin locally, using local windows B of coefficients extracted from each of the Gabor sub-bands, where the window B spans N pixels. Consider a pixel location $\mathbf{i}_0$. Let $\mathbf{f}(k)$ be a vector of dimension N, where $\mathbf{f}(k)$ is composed of the *complex magnitude* of N elements of $\tilde{f}(\mathbf{i}, k)$ spanned by the window B centered on $\mathbf{i}_0$. The Gabor coefficients $\tilde{f}(\mathbf{i}, k)$ are complex, but the vectors $\mathbf{f}(k)$ are real and denote the Gabor channel amplitude response.[20] Notice that we have just dropped the dependence on the spatio-temporal location $\mathbf{i}$ for notational convenience by considering a specific location $\mathbf{i}_0$. If the window B is specified by a set of relative indices, then $\mathbf{f}(k) = \{\tilde{f}(\mathbf{i}_0 + \mathbf{m}, k), \mathbf{m} \in B\}$. Similar definition applies for $\mathbf{g}(k)$. To index each element of $\mathbf{f}(k)$, we use the notation $\mathbf{f}(k) = [f_1(k), f_2(k), \ldots, f_N(k)]^T$.

The outputs of the Gabor filter-bank represent a decomposition of the reference and test video into bandpass channels. Individual Gabor filters respond to a specific range of spatio-temporal frequencies and orientations in the video, and any differences in the spectral content of the reference and distorted videos are captured by the Gabor outputs. Spatial distortions in the video such as blur, ringing, false contouring, blocking, noise and so on can be captured using errors computed between corresponding sub-bands of the reference and test videos.

Define the following spatial error from each subband response:

$$E_S(\mathbf{i}_0, k) = \frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}\left[\frac{f_n(k) - g_n(k)}{M(k) + C_1}\right]^2 \tag{1}$$

$$M(k) = \max\left(\sqrt{\frac{1}{N}\sum_{n=1}^{N}|f_n(k)|^2}, \sqrt{\frac{1}{N}\sum_{n=1}^{N}|g_n(k)|^2}\right) \tag{2}$$

Notice that the spatial error in (1) is computed as the MSE between $\mathbf{f}(k)$ and $\mathbf{g}(k)$ normalized by a masking function $M(k)$. Contrast masking refers to the reduction in visibility of a signal component (target) due to the presence of another signal component of similar frequency and orientation (masker) in a local spatial neighborhood.[3] In the context of VQA, the presence of large signal energy in the image content (masker) masks the visibility of noise or distortions (target) in these regions. The masking function in our model attempts to capture this feature of human visual perception and the masking function is a local energy measure computed from *both* the reference and distorted sub-bands. Masking models in the literature have adopted different approaches.[4, 27, 28] Our approach defined in (1) is a mutual masking function, wherein the divisive masking function is defined using both the reference and test videos (mutual masking functions were first used by Daly *et. al.*[28]). However, the use of the maximum in our approach in (2) differs significantly from the model used by Daly *et. al.* where a minimum is used. The reason for our choice is that MOVIE is intended to predict the visual annoyance of supra-threshold, easily visible distortions in the video, while the Daly model was intended the predict the visibility of differences between the reference and distorted images. We discovered that the use of maximum in (1) is well-suited to capture distortions such as compression and blur in the test video, that result in attenuation of Gabor sub-bands tuned to the finer scales. Additionally, use of the maximum prevents overprediction of errors in severely distorted regions of the video, such as significant loss of textures due to compression.

The Gaussian filter responds to the mean intensity or the DC component of the two images. A spatial error index can be defined using the output of the Gaussian filter operating at DC. Let $\mathbf{f}(\mathrm{DC})$ and $\mathbf{g}(\mathrm{DC})$ denote a vector of dimension N extracted at $\mathbf{i}_0$ from the output of the Gaussian filter operating on the reference and test videos respectively using the same window B. $\mathbf{f}(\mathrm{DC})$ and $\mathbf{g}(\mathrm{DC})$ are low pass filtered versions of the two videos. We first remove the effect of the mean intensity from each video before error computation, since this acts as a bias to the low frequencies present in the reference and distorted images that are captured by the Gaussian filter. We estimate the mean as the average of the Gaussian filtered output:

$$\mu_{\mathbf{f}} = \frac{1}{N}\sum_{n=1}^{N} f_n(\mathrm{DC}),\ \mu_{\mathbf{g}} = \frac{1}{N}\sum_{n=1}^{N} g_n(\mathrm{DC}) \tag{3}$$

An error index for the DC sub-band is then computed in a similar fashion as the Gabor sub-bands:

$$E_{\mathrm{DC}}(\mathbf{i}_0) = \frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}\left[\frac{|f_n(\mathrm{DC}) - \mu_{\mathbf{f}}| - |g_n(\mathrm{DC}) - \mu_{\mathbf{g}}|}{M_{\mathrm{DC}} + C_2}\right]^2 \tag{4}$$

$$M_{\mathrm{DC}} = \max\left(\sqrt{\frac{1}{N}\sum_{n=1}^{N}|f_n(\mathrm{DC}) - \mu_{\mathbf{f}}|^2}, \sqrt{\frac{1}{N}\sum_{n=1}^{N}|g_n(\mathrm{DC}) - \mu_{\mathbf{g}}|^2}\right) \tag{5}$$

The constants C_1 and C_2 in (1) and (4) are added to prevent instability when the denominators are small and are chosen to be $C_1 = 0.1, C_2 = 1$.

The spatial error indices computed from all of the Gabor sub-bands and the Gaussian sub-band can then be pooled to obtain an error index for location $\mathbf{i}_0$ using

$$E_S(\mathbf{i}_0) = \frac{\sum_{k=1}^{K} E_S(\mathbf{i}_0, k) + E_{\mathrm{DC}}(\mathbf{i}_0)}{K+1} \tag{6}$$

Finally, we convert the error index to a quality index at location $\mathbf{i}_0$ using

$$Q_S(\mathbf{i}_0) = 1 - E_S(\mathbf{i}_0) \tag{7}$$

4.3 Motion Estimation

To compute temporal quality, motion information is computed from the reference video sequence in the form of optical flow fields. The same set of Gabor filters used to compute the spatial quality component described above is used to calculate optical flow from the *reference video*. Our implementation uses the successful Fleet and Jepson algorithm that uses the phase of the complex Gabor outputs for motion estimation.[25] As an additional contribution, we realized a multi-scale version of the Fleet and Jepson algorithm that we describe very briefly below.

The algorithm by Fleet *et. al.* uses a 5-point central difference to perform derivative computation. However, we chose to perform the derivative computation more accurately by convolving the video sequence with filters that are derivatives of the Gabor kernels. To extend the algorithm by Fleet *et. al.* to multiple scales, we compute a 2D velocity estimate at each scale using the outputs of the Gabor filters at that scale only. It is important not to combine estimates across scales due to temporal aliasing.[21, 25] We also obtain an estimate of the residual error in the least squares solution for each scale of the Gabor filterbank. The final flow vector at each pixel of the reference video is set to be the 2D velocity computed at the scale with the minimum residual error.

4.4 Temporal MOVIE

The spatio-temporal Gabor decompositions of the reference and test video sequences, and the optical flow field computed from the *reference video* using the outputs of the Gabor filters can be used to estimate the temporal video quality. By measuring video quality along the motion trajectories, we expect to be able to account for the effect of temporal distortions of the type described in Section 3.

First, we discuss how translational motion manifests itself in the frequency domain. Let $a(x, y)$ denote an image patch and let $A(u, v)$ denote its Fourier transform. Assuming that this patch undergoes translation with a velocity $[\lambda, \phi]$ where λ and ϕ denote velocities along the x and y directions respectively, the resulting video sequence is given by $b(x, y, t) = a(x - \lambda t, y - \phi t)$. Then, $B(u, v, w)$, the Fourier transform of $b(x, y, t)$, lies entirely within a plane in the frequency domain.[29] Moreover, the magnitudes of the spatial frequencies do not change but are simply sheared:

$$B(u, v, w) = \begin{cases} A(u, v) & \text{if } \lambda u + \phi v + w = 0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Assume that short segments of video without any scene changes consist of local image patches undergoing translation. This is quite reasonable and is commonly used in video encoders that use motion compensation. This model can be used *locally* to describe video sequences, since translation is a linear approximation to more complex types of motion. Under this assumption, the reference and test videos $r(\mathbf{i})$ and $d(\mathbf{i})$ consist of local image patches (such as $a(x, y)$ in the example above) translating to create spatio-temporal video patches (such as $b(x, y, t)$).

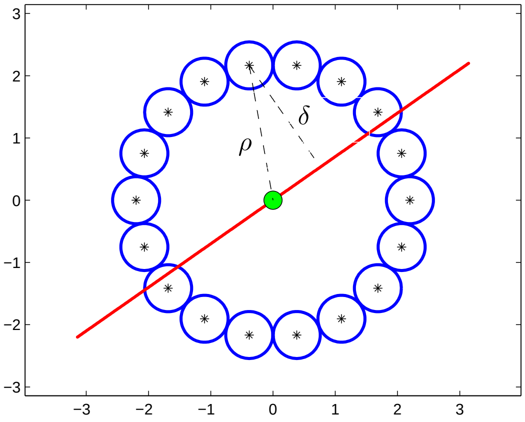

Figure 2. A slice of the Gabor filters and the spectral plane shown in 2 dimensions. The horizontal axis denotes horizontal spatial frequency and the vertical axis denotes temporal frequency. Each circle represents a Gabor filter and the centers of each filter are also marked. The radius ρ of the single scale of Gabor filters and the distance δ of the center frequency of one Gabor filter from spectral plane are marked.

Motion vectors from the reference can be used to construct responses from the reference and distorted Gabor outputs that are tuned to the speed and direction of movement of the reference.[22] This is accomplished by computing a weighted sum of the Gabor outputs, where the weight assigned to each individual filter is determined by its distance from the spectral plane of the reference video. Filters that lie very close to the spectral plane are assigned positive excitatory weights. Filters that lie away from the plane are assigned negative inhibitory weights. This achieves two objectives. First, the resulting response is tuned to the movement in the reference video. In other words, a strong response is obtained when the input video has a motion that is equal to the reference video signal. Additionally, any deviation from the reference motion is penalized due to the inhibitory weight assignment. An error computed between these motion tuned responses then serves to evaluate temporal video integrity. The weighting procedure is detailed in the following.

Let $\boldsymbol{\lambda}$ be a vector of dimension N, where $\boldsymbol{\lambda}$ is composed of N elements of the horizontal component of the flow field of the reference sequence spanned by the window B centered on $\mathbf{i}_0$. Similarly, $\boldsymbol{\phi}$ represents the vertical component of flow. Then, using (8), the spectrum of the reference video lies along:

$$\lambda_n u + \phi_n v + w = 0, n = 1, 2, \dots N \tag{9}$$

Define a sequence of distance vectors $\boldsymbol{\delta}(k), k = 1, 2, \dots, K$ of dimension N. Each element of this vector denotes the distance of the center frequency of the k^{th} filter from the plane containing the spectrum of the reference video in a window centered on $\mathbf{i}_0$ extracted using B. Let $\mathbf{U}_0(k) = [u_0(k), v_0(k), w_0(k)], k = 1, 2, \dots, K$ represent the center frequencies of all the Gabor filters. Then, $\boldsymbol{\delta}(k)$ represents the perpendicular distance of a point from a plane defined by (9) in a 3-dimensional space and is given by:

$$\delta_n(k) = \left| \frac{\lambda_n u_0(k) + \phi_n v_0(k) + w_0(k)}{\sqrt{\lambda_n^2 + \phi_n^2 + 1}} \right|, n = 1, 2, \dots, N \tag{10}$$

We now design a set of weights based on these distances. Our objective is to assign the filters that intersect the spectral plane to have the maximum weight of all filters. The distance of the center frequencies of these filters from the spectral plane is the minimum of all filters. First, define $\boldsymbol{\alpha}'(k), k = 1, 2, \dots, K$ using:

$$\alpha'_n k = \frac{\rho(k) - \delta_n(k)}{\rho(k)} \tag{11}$$

where $\rho(k)$ denotes the radius of the sphere along which the center frequency of the k^{th} filter lies in the frequency domain.

Figure 2 illustrates the geometrical computation specified in (11). Each of the circles represents the slice of a Gabor filter in 2 dimensions and the red line shows the projection of the spectral plane in 2 dimensions. The radius $\rho(k)$ and distance $\delta_n(k)$ are illustrated for one of the Gabor filters.

Since we want the weights to be excitatory and inhibitory, we shift all the weights at each scale to be zero-mean.[22] Finally, to make the weights insensitive to the filter geometry that was chosen, we normalize them so that the maximum weight is 1. This ensures that the maximum weight remains 1 irrespective of whether the spectral plane exactly intersects the center frequencies of the Gabor filters. We hence have a weight vector $\boldsymbol{\alpha}(k), k = 1, 2, \ldots, K$ with elements:

$$\alpha_n(k) = \frac{\alpha'_n(k) - \mu_\alpha}{\max_{k=1,2,\ldots,\frac{K}{P}} \left[\alpha'_n(k) - \mu_\alpha\right]},\ k = 1, 2, \ldots, \frac{K}{P} \tag{12}$$

where

$$\mu_\alpha = \frac{\sum_{k=1}^{\frac{K}{P}} \alpha'_n(k)}{\frac{K}{P}} \tag{13}$$

Similar definitions apply for other scales.

Motion tuned responses from the reference and distorted video sequences may be constructed using these weights. Define N-vectors $\boldsymbol{\nu}^r$ and $\boldsymbol{\nu}^d$ using:

$$\nu_n^r = \frac{(f_n(\mathrm{DC}) - \mu_{\mathbf{f}})^2 + \sum_{k=1}^{K} \alpha_n(k) f_n(k)^2}{(f_n(\mathrm{DC}) - \mu_{\mathbf{f}})^2 + \sum_{k=1}^{K} f_n(k)^2 + C_3} \tag{14}$$

$$\nu_n^d = \frac{(g_n(\mathrm{DC}) - \mu_{\mathbf{g}})^2 + \sum_{k=1}^{K} \alpha_n(k) g_n(k)^2}{(g_n(\mathrm{DC}) - \mu_{\mathbf{g}})^2 + \sum_{k=1}^{K} g_n(k)^2 + C_3} \tag{15}$$

The constant C_3 in (14) and (15) is added to prevent instability when the denominator terms are small and is chosen to be $C_3 = 100$. The vector $\boldsymbol{\nu}^r$ represents the response of the reference video to a mechanism that is tuned to *its own* motion. If the process of motion estimation was perfect and there was infinite translation resulting in a perfect plane, every element of $\boldsymbol{\nu}^r$ would be close to 1. The vector $\boldsymbol{\nu}^d$ represents the response of the distorted video to a mechanism that is tuned to the motion of the *reference video.* Thus, any deviation between the reference and distorted video motions is captured by (14) and (15).

The denominator terms in (14) and (15) ensure that temporal quality measurement is relatively insensitive to spatial distortions, thus avoiding redundancy in the spatial and temporal quality measurements. For example, in the case of blur, we would expect that the same Gabor filters are activated by the reference and distorted videos. However, the response of the finest scale filters are attenuated in the distorted video compared to the reference. Since each video is normalized by its own activity across all filters, the resulting response is not very sensitive to spatial distortions. Instead, the temporal mechanism responds strongly to distortions where the orientation of the spectral planes of the reference and distorted sequences differ.

Define a temporal error index using

$$E_T(\mathbf{i}_0) = \frac{1}{N} \sum_{n=1}^{N} (\nu_n^r - \nu_n^d)^2 \tag{16}$$

The error index in (16) is also exactly 0 when the reference and test images are identical. Finally, we convert the error index into a quality index using

$$Q_T(\mathbf{i}_0) = 1 - E_T(\mathbf{i}_0) \tag{17}$$

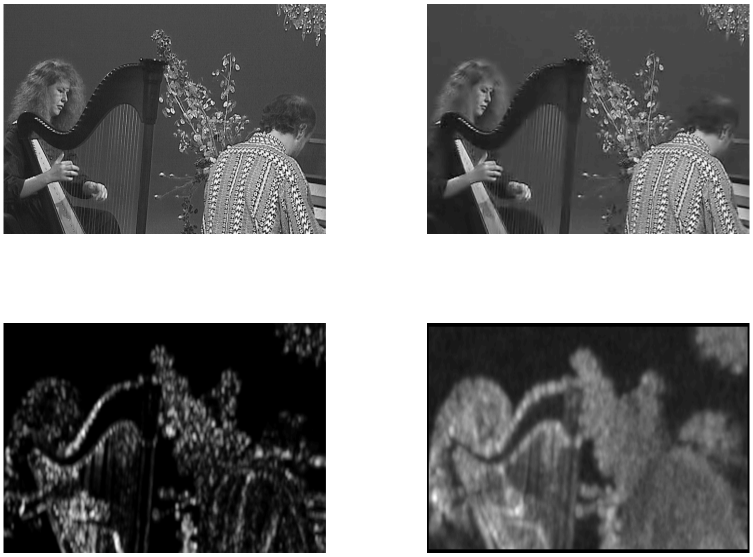

Figure 3. Illustration of the performance of the MOVIE index. Top left shows a frame from the reference video. Top right shows the corresponding frame from the distorted video. Bottom left shows a logarithmically compressed temporal quality map. Bottom right shows the spatial quality map. Notice that the spatial quality map responds to the blur in the test video. The temporal quality map responds to motion compensation mismatches surrounding the harp, heads of the two people and distortions in the strings.

4.5 MOVIE Index

The output of the spatial and temporal quality computation stages is two videos - a spatial quality video $Q_S(\mathbf{i})$ that represents the spatial quality at every pixel of the video sequence and a similar video for temporal quality denoted as $Q_T(\mathbf{i})$. Figure 3 illustrates one frame of the spatial and temporal quality videos generated by MOVIE on a representative video sequence in the VQEG database. The temporal quality map has been logarithmically compressed for visibility. First of all, it is evident that the kind of distortions captured by the spatial and temporal maps is different. The test video sequences in both examples suffer from significant blurring and the spatial quality map clearly reflects the loss of quality due to blur. The temporal quality map, however, shows poor quality along the edges of objects such as the harp where motion compensation mismatches are evident.

The final video quality index, which we call the MOVIE index, combines these into a single VQA index. Consider a set of specific time instants $t = \{t_0, t_1, \ldots, t_\tau\}$ which corresponds to frames in the spatial and temporal quality videos. We explored different pooling techniques to improve upon the commonly used strategy of using the mean of local quality indices (MOVIE quality maps in this instance) as the overall quality of the entire video. We found that the variance of the quality scores is also perceptually relevant. Indeed, a higher variance indicates a broader spread of both high and low quality regions in the video. Since lower quality regions affect the perception of video quality more so than do high quality regions, larger variances in the quality scores are indicative of lower perceptual quality. This is intuitively similar to pooling strategies based on percentiles, wherein the poorest percentile of the quality scores have been used to determine the overall quality.[14] A ratio of the standard deviation to the mean is often used in statistics and is known as the coefficient of variation. We have found that this moment ratio is a good predictor of the subjective quality of a video.

Define frame level quality indices for Spatial MOVIE using the ratio of the standard deviation to the mean of the Spatial MOVIE scores for that frame. Similar definition applies for Temporal MOVIE.

$$\mathrm{FQ}_S(t_j) = \frac{\sigma_{Q_S(x,y,t_j)}}{\mu_{Q_S(x,y,t_j)}},\ \mathrm{FQ}_T(t_j) = \frac{\sigma_{Q_T(x,y,t_j)}}{\mu_{Q_T(x,y,t_j)}} \tag{18}$$

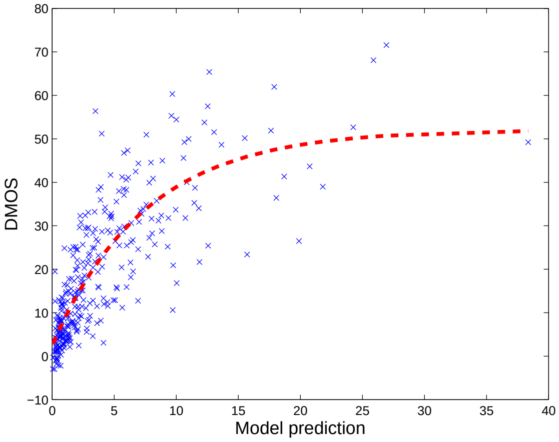

Figure 4. Scatter plot of the subjective DMOS scores against MOVIE scores on all sequences in the VQEG database. The best fitting logistic function used for non-linear regression is also shown.

The Spatial MOVIE index is then defined as the average of these frame level descriptors.

$$\text{Spatial MOVIE} = \frac{1}{\tau}\sum_{j=1}^{\tau} \mathrm{FQ}_S(t_j) \tag{19}$$

The range of values of the Temporal MOVIE scores is smaller than that of the spatial scores, due to the large divisive normalization in (14) and (15). To offset this effect, we use the square root of the temporal scores.

$$\text{Temporal MOVIE} = \sqrt{\frac{1}{\tau}\sum_{j=1}^{\tau} \mathrm{FQ}_T(t_j)} \tag{20}$$

The MOVIE index is defined as:

$$\text{MOVIE} = \text{Spatial MOVIE} \times \text{Temporal MOVIE} \tag{21}$$

5. RESULTS

We tested our algorithm on the VQEG FRTV Phase-1 database.[30] Since most of the VQEG videos are interlaced, our algorithm runs on just one field of the interlaced video. We ran our algorithm on the temporally earlier field for all sequences. We ignore the color component of the video sequences. The VQEG database contains 20 reference sequences and 16 distorted versions of each reference, for a total of 320 videos. Two distortions types in the VQEG database (HRC 8 and 9) contain two different subjective scores assigned by subjects corresponding to whether these sequences were viewed along with “high” or “low” quality videos.[30] We used the scores assigned in the “low” quality regime as the subjective scores for these videos.

The performance of our algorithm is reported for two metrics - the Spearman Rank Order Correlation Coefficient (SROCC) which is an indicator of the prediction monotonicity of the quality index and the Linear Correlation Coefficient (LCC) after non-linear regression. We used the same logistic function specified in[30] to fit the model predictions to the subjective data. The results are reported in Table 1.

PSNR provides a baseline for comparison of VQA models. Ten leading VQA models were tested by the VQEG in its Phase I study including a model from NTIA that was a precursor to VQM, as well as models from NASA, Sarnoff Corporation, KDD and EPFL.[30] Proponent P8 (Swisscom) is the best performing model of these ten models tested by the VQEG in terms of both SROCC and LCC after nonlinear regression.[30] SSIM (without weighting) refers to a frame-by-frame application of the SSIM index that was proposed for video.[11]

Table 1. Comparison of the performance of VQA algorithms.

Prediction Model	SROCC	LCC
Peak Signal to Noise Ratio	0.786	0.779
Proponent P8 (Swisscom)	0.803	0.827
SSIM (without weighting)	0.788	0.820
SSIM (weighting)	0.812	0.849
Spatial MOVIE	0.793	0.796
Temporal MOVIE	0.816	0.801
MOVIE	0.833	0.821

SSIM (weighting) refers to an extension of the SSIM index to video, that incorporated rudimentary motion information as weights for different regions of the video sequence.[11]

It is clear that the MOVIE index is competitive with other leading algorithms on the VQEG database. We wish to note that the proponents of the VQEG study did not have access to the database in the evaluation performed by the VQEG.[30] The performance of some of these algorithms have been improved since the publication of the study in 2000.[30] The performance of Spatial MOVIE is poorer than that of the Temporal MOVIE index, which powerfully illustrates the importance of capturing and assessing temporal video distortions. Using both in conjunction improves over using either separately. Scatter plots of the model prediction and DMOS values, along with the best fitting logistic function, for the MOVIE index are shown in Fig. 4.

6. CONCLUSIONS

In this paper, we presented a novel framework for spatio-temporal quality assessment of videos that integrates measurement of temporal distortions and motion related artifacts in video into the VQA framework. We developed an algorithm for VQA using this framework known as the MOVIE index. In addition to capturing spatial distortions in video, MOVIE computes and uses motion information from the reference video explicitly in quality measurement and evaluates the quality of the test video along the motion trajectories of the reference video. MOVIE was shown to deliver scores that correlate quite closely with human subjective judgment using the VQEG FR-TV Phase 1 database as a test bed. We would like to investigate extensions of the MOVIE index to utilize the color components of video in the future.

ACKNOWLEDGMENTS

This research was supported by a grant from the National Science Foundation (Award Number: 0728748).

REFERENCES

[1] Girod, B., "What's wrong with mean-squared error," in [*Digital Images and Human Vision*], Watson, A. B., ed., 207–220, The MIT Press (1993).

[2] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on* **13**(4), 600–612 (2004).

[3] Seshadrinathan, K., Pappas, T. N., Safranek, R. J., Chen, J., Wang, Z., Sheikh, H. R., and Bovik, A. C., "Image quality assessment," in [*The Essential Guide to Image Processing*], Bovik, A. C., ed., Elsevier (2008).

[4] Lubin, J., "The use of psychophysical data and models in the analysis of display system performance," in [*Digital Images and Human Vision*], Watson, A. B., ed., 163–178, The MIT Press (1993).

[5] van den Branden Lambrecht, C. J. and Verscheure, O., "Perceptual quality measure using a spatiotemporal model of the human visual system," in [*Proc. SPIE*], **2668**, 450–461, SPIE, San Jose, CA, USA (Mar. 1996).

[6] Winkler, S., "Perceptual distortion metric for digital color video," *Proc. SPIE* **3644**, 175–184 (May 1999).

[7] Watson, A. B., Hu, J., and McGowan III, J. F., "Digital video quality metric based on human vision," *J. Electron. Imaging* **10**, 20–29 (Jan. 2001).

[8] Masry, M., Hemami, S. S., and Sermadevi, Y., "A scalable wavelet-based video distortion metric and applications," *Circuits and Systems for Video Technology, IEEE Transactions on* **16**(2), 260–273 (2006).

[9] Wandell, B. A., [*Foundations of Vision*], Sinauer Associates Inc., Sunderland, MA (1995).

[10] Born, R. T. and Bradley, D. C., "Structure and function of visual area mt.," *Annu Rev Neurosci* **28**, 157–189 (2005).

[11] Wang, Z., Lu, L., and Bovik, A. C., "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication* **19**, 121–132 (Feb. 2004).

[12] Wang, Z. and Li, Q., "Video quality assessment using a statistical model of human visual speed perception.," *J Opt Soc Am A Opt Image Sci Vis* **24**, B61–B69 (Dec 2007).

[13] Sheikh, H. R. and Bovik, A. C., "A visual information fidelity approach to video quality assessment," in [*First International conference on video processing and quality metrics for consumer electronics*], (2005).

[14] Pinson, M. H. and Wolf, S., "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting* **50**, 312–322 (Sept. 2004).

[15] Hekstra, A. P., Beerends, J. G., Ledermann, D., de Caluwe, F. E., Kohler, S., Koenen, R. H., Rihs, S., Ehrsam, M., and Schlauss, D., "PVQM - A perceptual video quality measure," *Signal Proc.: Image Comm.* **17**, 781–798 (2002).

[16] ITU-T Rec. J. 144, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," (2004).

[17] Yuen, M. and Wu, H. R., "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Processing* **70**, 247–278 (Nov. 1998).

[18] Movshon, J. A., Thompson, I. D., and Tolhurst, D. J., "Spatial summation in the receptive fields of simple cells in the cat's striate cortex.," *J Physiol* **283**, 53–77 (Oct 1978).

[19] Daugman, J. G., "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)* **2**(7), 1160–1169 (1985).

[20] Bovik, A. C., Clark, M., and Geisler, W. S., "Multichannel texture analysis using localized spatial filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**, 55–73 (Jan. 1990).

[21] Heeger, D. J., "Optical flow using spatiotemporal filters," *International Journal of Computer Vision* **1**(4), 279–302 (1987).

[22] Simoncelli, E. P. and Heeger, D. J., "A model of neuronal responses in visual area MT," *Vision Res* **38**, 743–761 (Mar 1998).

[23] Adelson, E. H. and Bergen, J. R., "Spatiotemporal energy models for the perception of motion.," *J Opt Soc Am A* **2**, 284–299 (Feb 1985).

[24] Priebe, N. J., Lisberger, S. G., and Movshon, J. A., "Tuning for spatiotemporal frequency and speed in directionally selective neurons of macaque striate cortex.," *J Neurosci* **26**, 2941–2950 (Mar 2006).

[25] Fleet, D. and Jepson, A., "Computation of component image velocity from local phase information," *International Journal of Computer Vision* **5**(1), 77–104 (19900801).

[26] Sheikh, H. R. and Bovik, A. C., "Image information and visual quality," *IEEE Transactions on Image Processing* **15**(2), 430–444 (2006).

[27] Teo, P. C. and Heeger, D. J., "Perceptual image distortion," in [*Proceedings of the IEEE International Conference on Image Processing*], 982–986 vol.2, IEEE (1994).

[28] Daly, S., "The visible difference predictor: An algorithm for the assessment of image fidelity," in [*Digital Images and Human Vision*], Watson, A. B., ed., 176–206, The MIT Press (1993).

[29] Watson, A. B. and Ahumada, A. J., J., "Model of human visual-motion sensing," *Journal of the Optical Society of America A (Optics and Image Science)* **2**(2), 322–342 (1985).

[30] The Video Quality Experts Group, "Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment," (2000).