Motion Tuned Spatio-temporal Quality Assessment of Natural Videos

Kalpana Seshadrinathan*, *Member, IEEE* and Alan C. Bovik, *Fellow, IEEE*

***Abstract*—There has recently been a great deal of interest in the development of algorithms that objectively measure the integrity of video signals. Since video signals are being delivered to human end users in an increasingly wide array of applications and products, it is important that automatic methods of video quality assessment (VQA) be available that can assist in controlling the quality of video being delivered to this critical audience. Naturally, the quality of motion representation in videos plays an important role in the perception of video quality, yet existing VQA algorithms make little direct use of motion information, thus limiting their effectiveness. We seek to ameliorate this by developing a general, spatio-spectrally localized multiscale framework for evaluating dynamic video fidelity that integrates both spatial and temporal (and spatio-temporal) aspects of distortion assessment. Video quality is evaluated not only in space and time, but also in space-time, by evaluating motion quality along computed motion trajectories. Using this framework, we develop a full reference VQA algorithm for which we coin the term the MOtion-based Video Integrity Evaluation index, or MOVIE index. It is found that the MOVIE index delivers VQA scores that correlate quite closely with human subjective judgment, using the Video Quality Expert Group (VQEG) FRTV Phase 1 database as a test bed. Indeed, the MOVIE index is found to be quite competitive with, and even outperform, algorithms developed and submitted to the VQEG FRTV Phase 1 study, as well as more recent VQA algorithms tested on this database.**

I. Introduction

DIGITAL videos are increasingly finding their way into the day-to-day lives of people due to the rapid proliferation of networked video applications such as video on demand, digital television, video teleconferencing, streaming video over the Internet, video over wireless, consumer video appliances and so on. Quality control of videos from the capture device to the ultimate human user in these applications is essential in maintaining Quality of Service (QoS) requirements and methods to evaluate the perceptual quality of digital videos form a critical component of video processing and communication systems.

Humans can, almost instantaneously, judge the quality of an image or video that they are viewing, using prior knowledge and expectations derived from viewing millions of time-varying images on a daily basis. The right way to assess quality, then, is to ask humans for their opinion of the quality of an image or video, which is known as subjective assessment of quality. Indeed, subjective judgment of quality must be regarded as the ultimate standard of performance by which image quality assessment (IQA) or video quality assessment (VQA) algorithms are assessed. Subjective quality is measured by asking a human subject to indicate the quality of an image or video that they are viewing on a numerical or qualitative scale. To account for human variability and to assert statistical confidence, multiple subjects are required to view each image/video, and a Mean Opinion Score (MOS) is computed. While subjective methods are the only completely reliable method of VQA, subjective studies are cumbersome and expensive. For example, statistical significance of the MOS must be guaranteed by using sufficiently large sample sizes; subject naivety must be imposed; the dataset of images/videos must be carefully calibrated; and so on [1], [2]. Subjective VQA is impractical for nearly every application other than benchmarking automatic or objective VQA algorithms.

To develop generic VQA algorithms that work across a range of distortion types, full reference algorithms assume the availability of a "perfect" reference video, while each test video is assumed to be a distorted version of this reference.

We survey the existing literature on full reference VQA in Section II. The discussion there will highlight the fact that although current full reference VQA algorithms incorporate features for measuring spatial distortions in video signals, very little effort has been spent on directly measuring temporal distortions or motion artifacts. As described in Section II, several algorithms utilize rudimentary temporal information by differencing adjacent frames or by processing the video using simple temporal filters before feature computation. However, most existing VQA algorithms do not attempt to directly compute motion information in video signals to predict quality; notable exceptions include [3], [4], [5], [6], [7]. [3] is not a generic VQA algorithm and targets video coding applications, where models of visual motion sensors developed in [8] are utilized to perform computations that signal the direction of motion. In [4], [5], [6], motion information is only used to design weights to pool local *spatial* quality indices into a single quality score for the video. TetraVQM appeared subsequent to early submissions of this work [9] and computes motion compensated errors between the reference and distorted videos [7].

Yet, motion plays a very important role in human perception of moving image sequences [10]. Considerable resources in the human visual system (HVS) are devoted to motion perception. The HVS can accurately judge the velocity and direction of motion of objects in a scene, skills that are essential to survival. Humans are capable of making smooth pursuit eye movements to track moving objects. Visual attention is known to be drawn to movement in the periphery of vision, which makes humans and other organisms aware of approaching danger [10], [11]. Additionally, motion provides important clues about the shape

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

of three dimensional objects and aids in object identification. All these properties of human vision demonstrate the important role that motion plays in perception, and the success of VQA algorithms depends on their ability to model and account for motion perception in the HVS.

While video signals do suffer from spatial distortions, they are often degraded by severe *temporal* artifacts such as ghosting, motion compensation mismatch, jitter, smearing, mosquito noise (amongst numerous other types), as described in detail in Section III. It is imperative that video quality indices account for the deleterious perceptual influence of these artifacts, if objective evaluation of video quality is to accurately predict subjective judgment. Most existing VQA algorithms are able to capture spatial distortions that occur in video sequences (such as those described in Section III-A), but don't do an adequate job in capturing temporal distortions (such as those described in Section III-B).

We seek to address this by developing a general framework for achieving spatio-spectrally localized multiscale evaluation of dynamic video quality. In this framework, both spatial and temporal (and spatio-temporal) aspects of distortion assessment are accounted for. Video quality is evaluated not only in space and time, but also in space-time, by evaluating motion quality along computed motion trajectories.

Using this framework, we develop a full reference VQA algorithm which we call the MOtion-based Video Integrity Evaluation index, or MOVIE index. MOVIE integrates explicit motion information into the VQA process by tracking perceptually relevant distortions along motion trajectories, thus augmenting the measurement of spatial artifacts in videos. Our approach to VQA represents an evolution, as we have sought to develop principles for VQA that were inspired by the structural similarity and information theoretic approaches to IQA proposed in [12], [13], [14], [15]. The Structural SIMilarity (SSIM) index and the Visual Information Fidelity (VIF) criterion are successful still image quality indices that correlate exceedingly well with perceptual image quality as demonstrated in extensive psychometric studies [16]. Indeed, our early approaches were extensions of these algorithms, called Video SSIM and Video Information Fidelity Criterion (IFC) [9], [17], where, roughly speaking, quality indices were computed along the motion trajectories.

Our current approach, culminating in the MOVIE index, represents a significant step forward from our earlier work, as we develop a general framework for measuring both spatial and temporal video distortions over multiple scales, and along motion trajectories, while accounting for spatial and temporal perceptual masking effects. As we show in the sequel, the performance of this approach is highly competitive with algorithms developed and submitted to the VQEG FRTV Phase 1 study, as well as more recent VQA algorithms tested on this database.

We review the existing literature on VQA in Section II. To supply some understanding of the challenging context of VQA, we describe commonly occurring distortions in digital video sequences in Section III. The development of the MOVIE index is detailed in Section IV. We explain the relationship between the MOVIE model and motion perception in biological vision systems in Section V. We also describe the relationship between MOVIE and the SSIM and VIF still image quality models in that section. The performance of MOVIE is presented in Section VI, using the publicly available Video Quality Expert Group (VQEG) FRTV Phase 1 database. We conclude the paper in Section VII with a discussion of future work.

II. Background

Mathematically simple error indices such as the Mean Squared Error (MSE) are often used to evaluate video quality, mostly due to their simplicity. It is well known that the MSE does not correlate well with visual quality, which is the reason why research into full reference VQA techniques has been intensely studied [18]; see [19] for a review of VQA. Several types of weighted MSE and Peak Signal to Noise Ratio (PSNR) have also been proposed by researchers; see, for example, [20], [21], [22].

A substantial amount of the research into IQA and VQA has focused on using models of the HVS to develop quality indices, which we broadly classify as HVS-based indices. The basic idea behind these approaches is that the best way to predict the quality of an image or video, in the absence of any knowledge of the distortion process, is to attempt to "see" the image using a system similar to the HVS. Typical HVS-based indices use linear transforms separably in the spatial and temporal dimensions to decompose the reference and test videos into multiple channels, in an attempt to model the tuning properties of neurons in the front-end of the eye-brain system. Contrast masking, contrast sensitivity and luminance masking models are then used to obtain thresholds of visibility for each channel. The error between the test and reference video in each channel is then normalized by the corresponding threshold to obtain the errors in Just Noticeable Difference (JND) units. The errors from different channels at each pixel are then combined, using the Minkowski error norm or other pooling strategies, to obtain a space-varying map that predicts the probability that a human observer will be able to detect any difference between the two images.

Examples of HVS-based image quality indices include [23], [24], [25], [26], [27]; see [28] for a review. It is believed that two kinds of temporal mechanisms exist in the early stages of processing in the HVS, one lowpass and one bandpass, known as the sustained and transient mechanisms [29], [30]. Most HVS-based video quality indices have been derived from still image quality indices by the addition of a temporal filtering block to model these mechanisms. Popular HVS-based video quality indices such as the Moving Pictures Quality Metric (MPQM) [31], Perceptual Distortion Metric (PDM) [32] and the Sarnoff JND vision model [24] filter the videos using one bandpass and one lowpass filter along the temporal dimension. Other methods such as the Digital Video Quality (DVQ) metric [33] and the scalable wavelet based video distortion index in [34] utilize a single low pass filter along the temporal dimension. A VQA algorithm that estimates spatiotemporal distortions through a temporal analysis of spatial perceptual distortion maps was presented in [6].

One of the visual processing tasks performed by the HVS is the computation of speed and direction of motion of objects using the series of time-varying images captured by the retina. All the indices mentioned above use either one or two temporal channels and model the temporal tuning of only the neurons in early stages of the visual pathway such as the retina, lateral geniculate nucleus (LGN) and Area V1 of the cortex. This is the first stage of motion processing that occurs in primate vision systems, the outputs of which are used in latter stages of motion processing that occur in Area MT/V5 of the extra-striate cortex [35]. Visual area MT 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 [36]. Models of processing in MT is hence essential in VQA due to the critical role of these functions in the perception of videos by human observers. The response properties of neurons in Area MT/V5 are well studied in primates and detailed models of motion sensing have been proposed [37], [38], [39]. To our knowledge, no VQA index has attempted to incorporate these models to account for visual processing of motion in Area MT.

More recently, there has been a shift toward VQA techniques that attempt to characterize features that the human eye associates with loss of quality; for example, blur, blocking artifacts, fidelity of edge and texture information, color information, contrast and luminance of patches and so on. Part of the reason for this shift in paradigm has been the complexity and incompleteness of models of the HVS. A more important reason, perhaps, is the fact that HVS based models typically model threshold psychophysics, i.e., the sensitivity of the HVS to different features are measured at the threshold of perception [40], [41], [10]. However, while detection of distortions is important in some applications, VQA deals with supra-threshold perception, where artifacts in the video sequences are visible and algorithms attempt to quantify the *annoyance levels* of these distortions. Popular VQA algorithms that embody this approach include the Video Quality Metric (VQM) from NTIA [42], the SSIM index for video [4], [5], Perceptual Video Quality Measure (PVQM) [43] and other algorithms from industry [44], [45], [46], [47], [48]. However, these models also predominantly capture spatial distortions in the video sequence and fail to do an adequate job in capturing temporal distortions in video. VQM considers 3D spatio-temporal blocks of video in computing some features, and the only temporal component of the VQM method involves frame differences [42]. The extensions of the SSIM index for video compute local *spatial* SSIM indices at each frame of the video sequence and use motion information only as weights to combine these local quality measurements into one single quality score for the entire video [4], [5]. TetraVQM is a VQA algorithm that appeared subsequent to early submissions of this work, that utilizes motion estimation within a VQA framework, where motion compensated errors are computed between the reference and distorted images [7].

There is a need for improvement in the performance of objective quality indices for video. Most of the indices proposed in the literature have been simple extensions of quality indices for images. Biological vision systems devote considerable resources to motion processing. Presentations of video sequences to human subjects induce visual experiences of motion and the perceived distortion in video sequences is a combination of both spatial and motion artifacts. We argue that accurate representation of motion in video sequences, as well as of temporal distortions, have great potential to advance video quality prediction. We present such an approach to VQA in this paper.

III. Distortions in Digital Video

In this section, we discuss the kinds of distortions that are commonly observed in video sequences [49]. Distortions in digital video inevitably exhibit both spatial and temporal aspects. Even a process such as blur from a lens has a temporal aspect, since the blurred regions of the video tend to move around from frame to frame. Nevertheless, there are distortions that are primarily spatial, which we shall call "spatial distortions".

Likewise, there are certain distortions that are primarily temporal in that they arise purely from the occurrence of motion, although such distortions may affect individual frames of the video as well. We will refer to these as "temporal distortions".

A. Spatial Distortions

Examples of commonly occurring spatial distortions in video include blocking, ringing, mosaic patterns, false contouring, blur and noise [49]. *Blocking effects* result from block based compression techniques used in several Discrete Cosine Transform (DCT) based compressions systems including Motion Picture Experts Group (MPEG) systems such as MPEG-1, MPEG-2, MPEG-4 and H.263, H.264. Blocking appears as periodic discontinuities in each frame of the compressed video at block boundaries. *Ringing distortions* are visible around edges or contours in frames and appear as a rippling effect moving outward from the edge toward the background. Ringing artifacts are visible in non-block based compression systems such as Motion JPEG-2000 as well. *Mosaic Patterns* are visible in block based coding systems and manifest as a mismatch between the contents of adjacent blocks as a result of coarse quantization. *False contouring* occurs in smoothly textured regions of a frame containing gradual degradation of pixel values over a given area. Inadequate quantization levels result in step-like gradations having no physical correlate in the reconstructed frame. *Blur* is a loss of high frequency information and detail in video frames. This can occur due to compression, or as a by-product of image acquisition. *Additive Noise* manifests itself as a grainy texture in video frames. Additive noise arises due to video acquisition and by passage through certain video communication channels.

B. Temporal Distortions

Examples of commonly occurring temporal artifacts in video include motion compensation mismatch, mosquito noise, stationary area fluctuations, ghosting, jerkiness and smearing [49]. *Motion compensation mismatch* occurs due to the

assumption that all constituents of a macro-block undergo identical motion, which might not be true. This is most evident around the boundaries of moving objects and appears as the presence of objects and spatial characteristics that are uncorrelated with the depicted scene. *Mosquito effect* is a temporal artifact seen primarily as fluctuations in light levels in smooth regions of the video surrounding high contrast edges or moving objects. *Stationary area fluctuations* closely resemble the mosquito effect in appearance, but are usually visible in textured stationary areas of a scene. *Ghosting* appears as a blurred remnant trailing behind fast moving objects in video sequences. This is a result of deliberate lowpass filtering of the video along the temporal dimension to remove additive noise that may be present in the source. *Jerkiness* results from delays during the transmission of video over a network where the receiver does not possess enough buffering ability to cope with the delays. *Smearing* is an artifact associated with the non-instantaneous exposure time of the acquisition device, where light from multiple points of the moving object at different instants of time are integrated into the recording.

It is important to observe that temporal artifacts such as motion compensation mismatch, jitter and ghosting alter the movement trajectories of pixels in the video sequence. Artifacts such as mosquito noise and stationary area fluctuations introduce a false perception of movement arising from temporal frequencies created in the test video that were not present in the reference. The perceptual annoyance of these distortions is closely tied to the process of motion perception and motion segmentation that occurs in the human brain while viewing the distorted video.

IV. Motion Tuned Spatio-Temporal Framework for Video Quality Assessment

In our framework for VQA, separate components for spatial and temporal quality are defined. First, the reference and test videos are decomposed into spatio-temporal bandpass channels using a Gabor filter family. Spatial quality measurement is accomplished by a method loosely inspired by the SSIM index and the information theoretic methods for IQA [13], [15], [14]. Temporal quality is measured using motion information from the reference video sequence. Finally, the spatial and temporal quality scores are pooled to obtain an overall video integrity score known as the MOVIE index [50]. Figure 1 shows a block diagram of the MOVIE index and each stage of processing in MOVIE is detailed in the following.

A. Linear Decomposition

Frequency domain approaches are well suited to the study of human perception of video signals and form the backbone of most IQA and VQA systems. Neurons in the visual cortex and the extra-striate cortex are spatial frequency and orientation selective and simple cells in the visual cortex are known to act more or less as linear filters [51], [52], [53]. 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

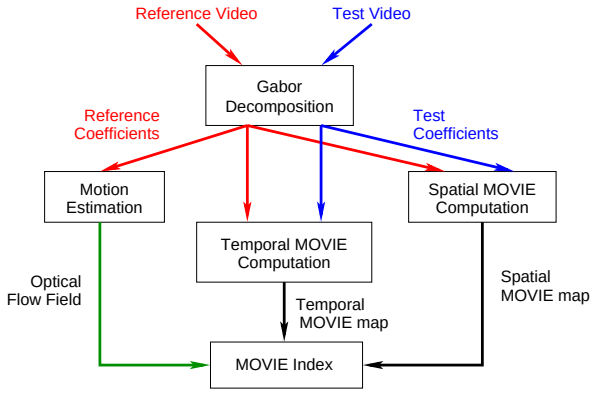

Fig. 1. Block diagram of MOVIE index. Flow of the reference video through the MOVIE VQA system is color coded in red, while flow of the test video is shown in blue. Both reference and test videos undergo linear decomposition using a Gabor filter family. Spatial and temporal quality is estimated using the Gabor coefficients from the reference and test videos. Temporal quality computation additionally uses reference motion information, computed using the reference Gabor coefficients. Spatial and temporal quality indices are then combined to produce the overall MOVIE index.

characteristics and movement information in a video sequence are captured by a linear spatio-temporal decomposition.

In our framework for VQA, a video sequence is filtered spatio-temporally using a family of bandpass Gabor filters and video integrity is evaluated on the resulting bandpass channels in the spatio-temporal frequency domain. Evidence indicates that the receptive field profiles of simple cells in the mammalian visual cortex are well modeled by Gabor filters [52]. The Gabor filters that we use in the algorithm we develop later are separable in the spatial and temporal coordinates and several studies have shown that neuronal responses in Area V1 are approximately separable [54], [55], [56]. Gabor filters attain the theoretical lower bound on uncertainty in the frequency and spatial variables and thus, visual neurons approximately optimize this uncertainty [52]. In our context, the use of Gabor basis functions guarantees that video features extracted for VQA purposes will be optimally localized.

Further, the responses of several spatio-temporally separable responses can be combined to encode the local speed and direction of motion of the video sequence [57], [58]. Spatio-temporal Gabor filters have been used in several models of the response of motion selective neurons in the visual cortex [57], [59], [39]. In our implementation of the ideas described here, we utilize the algorithm described in [60] 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.

A Gabor filter $h(\mathbf{i})$ in three dimensions is the product of a Gaussian window and a complex exponential:

$$h(\mathbf{i}) = \frac{1}{(2\pi)^{\frac{3}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{\mathbf{i}^T\Sigma^{-1}\mathbf{i}}{2}\right) \exp\left(j\mathbf{U_o}^T\mathbf{i}\right) \quad (1)$$

where $\mathbf{i} = (x, y, t)$ is a vector denoting a spatio-temporal location in the video sequence and $\mathbf{U_o} = (U_0, V_0, W_0)$ is the center frequency of the Gabor filter. Σ is the covariance matrix of the Gaussian component of the Gabor filter. The Fourier transform of the Gabor filter is a Gaussian with covariance

matrix Σ^{-1}:

$$H(\mathbf{u}) = \exp\left(-\frac{(\mathbf{u}-\mathbf{U_0})^T\Sigma(\mathbf{u}-\mathbf{U_0})}{2}\right) \quad (2)$$

Here, $\mathbf{u} = (u, v, w)$ denotes the spatio-temporal frequency coordinates.

Our implementation uses separable Gabor filters that have equal standard deviations along both spatial frequency coordinates and the temporal coordinate. Thus, Σ is a diagonal matrix with equal valued entries along the diagonal. Our filter design is very similar to the filters used in [60]. However, our filters have narrower bandwidth and are multi-scale as described below.

All the filters in our Gabor filter bank have constant octave bandwidths. We use $P = 3$ scales of filters, with 35 filters at each scale. Figure 2(a) 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. 2(a). We used filters with rotational symmetry and the spatial spread of the Gabor filters is the same along all axes. The filters have an octave bandwidth of 0.5 octaves, measured at one standard deviation of the Gabor frequency response. The center frequencies of the finest scale of filters lie on the surface of a sphere in the frequency domain, whose radius is 0.7π radians per sample. Each of these filters has a standard deviation of 2.65 pixels along both spatial coordinates and 2.65 frames along the temporal axis. In our implementation, the Gabor filters were sampled out to a width of three standard deviations; so the support of the kernels at the finest scale are 15 pixels and 15 frames along the spatial and temporal axes respectively. The center frequencies of the filters at the coarsest scale lie on the surface of a sphere of radius 0.35π, have a standard deviation of 5.30 pixels (frames) and a support of 33 pixels (frames).

Nine filters are tuned to a temporal frequency of 0 radians per sample corresponding to no motion. The orientations of these filters are chosen such that adjacent filters intersect at one standard deviation; hence the orientations of these filters are chosen to be multiples of 20° in the range $[0^\circ, 180^\circ)$. Seventeen filters are tuned to horizontal or vertical speeds of $s = 1/\sqrt{3}$ pixels per frame and the temporal center frequency of each of these filters is given by $\rho * \frac{s}{\sqrt{s^2+1}}$ radians per sample, where ρ is the radius of the sphere that the filters lie on [60]. Again, the orientations are chosen such that adjacent filters intersect at one standard deviation and the orientations of these filters are multiples of 22° in the range $[0^\circ, 360^\circ)$. The last nine filters are tuned to horizontal or vertical velocities of $\sqrt{3}$ pixels per frame. The orientations of these filters are multiples of 40° in the range $[0^\circ, 360^\circ)$.

Figure 2(b) shows a slice of the sine phase component of the Gabor filters along the plane of zero temporal frequency ($w = 0$) and shows the three scales of filters with constant octave bandwidths. Figure 2(c) shows a slice of the sine phase component of the Gabor filters along the plane of zero vertical spatial frequency. Filters along the three radial lines are tuned to the three different speeds of $(0, \frac{1}{\sqrt{3}}, \sqrt{3})$ pixels per frame.

Finally, a Gaussian filter is included at the center of the Gabor structure to capture the low frequencies in the signal. The standard deviation of the Gaussian filter is chosen so that it intersects the coarsest scale of bandpass filters at one standard deviation.

B. Spatial MOVIE Index

Our approach to capturing spatial distortions in the video of the kind described in Section III-A is inspired both by the SSIM index and the information theoretic indices that have been developed for IQA [13], [61], [14]. 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. We will hence term this part of our model the "Spatial MOVIE Index", taking this to mean that the model primarily captures spatial distortions. We explain how the Spatial MOVIE index relates to and improves upon prior approaches in Section V.

Let $r(\mathbf{i})$ and $d(\mathbf{i})$ denote the reference and distorted videos respectively, where $\mathbf{i} = (x, y, t)$ is a vector denoting a spatio-temporal location in the video sequence. 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 subbands, 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. 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 $\mathbf{f}(k) = [f_1(k), f_2(k), \ldots, f_N(k)]^T$.

Contrast masking is a property of human vision that 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 [62]. 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. Contrast masking has been modeled using a mechanism of contrast gain control that often takes the form of a divisive normalization [63], [25], [64]. Models of contrast gain control using divisive normalization arise in psychophysical literature from studies of the non-linear response properties of neurons in the primary visual cortex [65], [66], [67] and have also been shown to be well-suited for efficient encoding of natural signals by the visual system [68].

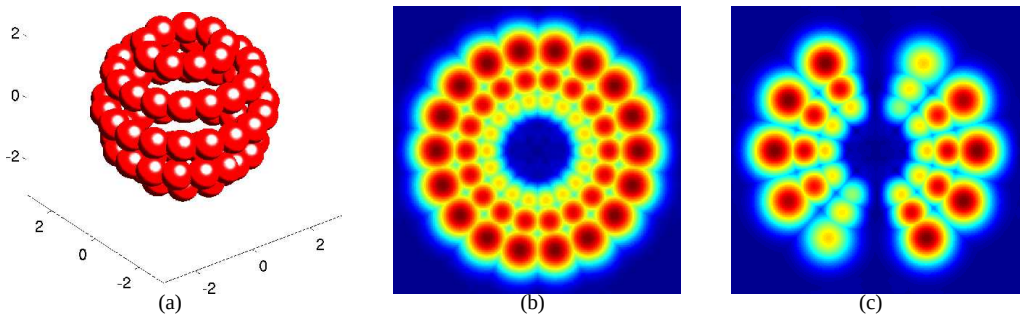

Fig. 2. (a) Geometry of the Gabor filterbank in the frequency domain. The figure shows iso-surface contours of all Gabor filters at the finest scale. The two horizontal axes denote the spatial frequency coordinates and the vertical axis denotes temporal frequency. (b) A slice of the Gabor filter bank along the plane of zero temporal frequency. The x-axis denotes horizontal spatial frequency and the y-axis denotes vertical spatial frequency. (c) A slice of the Gabor filter bank along the plane of zero vertical spatial frequency. The x-axis denotes horizontal spatial frequency and the y-axis denotes temporal frequency.

The Spatial MOVIE index attempts to capture this property of human vision and we define the spatial error from each subband response using:

$$E_S(\mathbf{i}_0,k) = \frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}\left[\frac{f_n(k)-g_n(k)}{M(k)+C_1}\right]^2 \tag{3}$$

where $M(k)$ is defined as

$$M(k) = \max\left(\sqrt{\frac{1}{N}\sum_{n=1}^{N}|f_n(k)|^2}, \sqrt{\frac{1}{N}\sum_{n=1}^{N}|g_n(k)|^2}\right) \tag{4}$$

C_1 is a small positive constant that is included to prevent numerical instability when the denominator of (3) goes to 0. This can happen in smooth regions of the video (for instance, smooth backgrounds, sky, smooth object surfaces and so on), where most of the bandpass Gabor outputs are close to 0. Additionally, since the divisive normalization in (3) is modeled within a sub-band, the denominator in (3) can go to zero for certain sub-bands in sinusoid-like image regions, high frequency sub-bands of edge regions and so on.

In summary, 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 MOVIE then uses a divisive normalization approach to capture contrast masking wherein the visibility of errors between the reference and distorted images ($\mathbf{f}(k)$ and $\mathbf{g}(k)$) are inhibited divisively by $M(k)$, which is a local energy measure computed from the reference and distorted sub-bands. (3) detects primarily spatial distortions in the video such as blur, ringing, false contouring, blocking, noise and so on.

The error index $E_S(\mathbf{i}_0,k)$ is bounded and lies between 0 and 1:

$$\begin{aligned} E_S(\mathbf{i}_0,k) &= \frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}\left[\frac{f_n(k)-g_n(k)}{M(k)+C_1}\right]^2 \\ &= \frac{1}{2}\left\{\frac{\frac{1}{N}\sum_{n=1}^{N}f_n(k)^2}{[M(k)+C_1]^2} + \frac{\frac{1}{N}\sum_{n=1}^{N}g_n(k)^2}{[M(k)+C_1]^2} \right. \\ &\qquad \left. - 2\frac{\frac{1}{N}\sum_{n=1}^{N}f_n(k)g_n(k)}{[M(k)+C_1]^2}\right\} \\ &\le \frac{1}{2}\left\{\frac{\frac{1}{N}\sum_{n=1}^{N}f_n(k)^2}{[M(k)+C_1]^2} + \frac{\frac{1}{N}\sum_{n=1}^{N}g_n(k)^2}{[M(k)+C_1]^2}\right\} \quad (5) \\ &\le \left[\frac{M(k)}{M(k)+C_1}\right]^2 \quad (6) \end{aligned}$$

(5) uses the fact that $f_n(k)$ and $g_n(k)$ are non-negative. (6) follows from the definition of $M(k)$. Therefore, $E_S(\mathbf{i}_0,k)$ lies between 0 and 1. Observe that the spatial error in (3) is exactly 0 when the reference and distorted videos are identical.

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 vectors 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:

$$\mu_{\mathbf{f}} = \frac{1}{N}\sum_{n=1}^{N}f_n(\mathrm{DC}), \quad \mu_{\mathbf{g}} = \frac{1}{N}\sum_{n=1}^{N}g_n(\mathrm{DC}) \tag{7}$$

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

$$E_{\mathrm{DC}}(\mathbf{i}_0) = \frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}\left[\frac{|f_n(\mathrm{DC})-\mu_{\mathbf{f}}| - |g_n(\mathrm{DC})-\mu_{\mathbf{g}}|}{M_{\mathrm{DC}}+C_2}\right]^2 \tag{8}$$

where M_{DC} is defined as

$$M_{\text{DC}} = \max\left(\sqrt{\frac{1}{N}\sum_{n=1}^{N}|f_n(\text{DC}) - \mu_{\mathbf{f}}|^2}, \sqrt{\frac{1}{N}\sum_{n=1}^{N}|g_n(\text{DC}) - \mu_{\mathbf{g}}|^2}\right) \quad (9)$$

C_2 is a constant added to prevent numerical instability when the denominator of (8) goes to 0. This can happen in smooth image regions since the DC sub-band is close to constant in these regions.

It is straightforward to verify that $E_{\text{DC}}(\mathbf{i}_0)$ also lies between 0 and 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

$$E_S(\mathbf{i}_0) = \frac{\sum_{k=1}^{K} E_S(\mathbf{i}_0, k) + E_{\text{DC}}(\mathbf{i}_0)}{K+1} \quad (10)$$

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

$$Q_S(\mathbf{i}_0) = 1 - E_S(\mathbf{i}_0) \quad (11)$$

C. 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 [60] algorithm that uses the *phase* of the complex Gabor outputs for motion estimation. Notice that we only used the complex magnitude in the spatial quality computation and, as it turns out, we only use the complex magnitudes to evaluate the temporal quality. As an additional contribution, we have realized a multi-scale version of the Fleet and Jepson algorithm, which we briefly describe in the Appendix.

D. Temporal MOVIE Index

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 distortions of the type described in Section III-B. Once again, the model described here primarily captures temporal distortions in the video, while responding to spatial distortions in a limited fashion. We hence call this stage of our model the "Temporal Movie Index".

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 [8]. This plane is defined by:

$$\lambda u + \phi v + w = 0 \quad (12)$$

Moreover, the magnitudes of the spatial frequencies do not change but are simply sheared:

$$B(u, v, w) = \begin{cases} A(u, v) \text{ if } \lambda u + \phi v + w = 0 \\ 0 \qquad \text{otherwise} \end{cases} \quad (13)$$

Spatial frequencies in the video signal provide information about the spatial characteristics of objects in the video sequence such as orientation, texture, sharpness and so on. Translational motion shears these spatial frequencies to create orientation along the temporal frequency dimension without affecting the magnitudes of the spatial frequencies. Translational motion has an easily accessible representation in the frequency domain and these ideas have been used to build motion estimation algorithms for video [8], [57], [58].

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 $b(x, y, t)$). Observe that (12) and (13) assume infinite translation of the image patches [8], which is not practical. In actual video sequences, local spectra will not be planes, but will in fact be the convolution of (13) with the Fourier transform of a truncation window (a sinc function). However, the rest of our development will assume infinite translation and it will be clear as we proceed that this will not significantly affect the development.

The optical flow computation on the reference sequence provides an estimate of the local orientation of this spectral plane at every pixel of the video. Assume that the motion of each pixel in the distorted video sequence *exactly* matches the motion of the corresponding pixel in the reference. We would then expect that the filters that lie along the motion plane orientation identified from the reference are activated by the distorted video and that the outputs of all Gabor filters that lie away from this spectral plane are negligible. However, when temporal artifacts are present, the motion in the reference and distorted video sequences do not match. This situation happens, for example, in motion compensation mismatches, where background pixels that are static in the reference move with the objects in the distorted video due to block motion estimation. Another example is ghosting, where static pixels surrounding moving objects move in the distorted video due to temporal low-pass filtering. Other examples are mosquito noise and stationary area fluctuations, where the visual appearance of motion is created from temporal frequencies in the distorted video that were not present in the reference. All of these artifacts shift the spectrum of the

distorted video to lie along a different orientation than the reference.

The 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. 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 (12), the spectrum of the reference video lies along:

$$\lambda_n u + \phi_n v + w = 0, n = 1, 2, \ldots N \tag{14}$$

Define a sequence of distance vectors $\boldsymbol{\delta}(k), k = 1, 2, \ldots, K$ 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 $\mathbf{U}_0(k) = [u_0(k), v_0(k), w_0(k)], k = 1, 2, \ldots, K$ 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 (14) in a 3-dimensional space and is given by:

$$\delta_n(k) = \left| \frac{\lambda_n u_0(k) + \phi_n v_0(k) + w_0(k)}{\sqrt{\lambda_n^2 + \phi_n^2 + 1}} \right|, n = 1, 2, \ldots, N \tag{15}$$

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 $\boldsymbol{\alpha}'(k), k = 1, 2, \ldots, K$ using:

$$\alpha_n' k = \frac{\rho(k) - \delta_n(k)}{\rho(k)} \tag{16}$$

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 3 illustrates the geometrical computation specified in (16).

From the geometry of the Gabor filterbank, it is clear that $0 \leq \alpha_n'(k) \leq 1 \forall n, k$ since the spectral plane specified by (14) always passes through the origin. If the spectral plane passes through the center frequency of a Gabor filter k, then it passes through the corresponding Gabor filter at all scales. $\alpha_n'(k) = 1$ for this filter and the corresponding filters at other scales. If the center frequency of a Gabor filter k lies along a plane that passes through the origin and is perpendicular to the spectral plane of the reference video, then $\alpha_n'(k) = 0$.

Since we want the weights to be excitatory and inhibitory, we shift all the weights at each scale to be zero-mean [58]. 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. Although the weights are invariant to the filter geometry, observe that due to the Gaussian falloff in the frequency response of the Gabor filters, the Gabor responses themselves are not insensitive to the filter geometry. We hence have a weight vector $\boldsymbol{\alpha}(k), k = 1, 2, \ldots, K$ with elements:

$$\alpha_n(k) = \frac{\alpha_n'(k) - \mu_\alpha}{\max_{k=1,2,\ldots,\frac{K}{P}} [\alpha_n'(k) - \mu_\alpha]}, \; k = 1, 2, \ldots, \frac{K}{P} \tag{17}$$

where

$$\mu_\alpha = \frac{\sum_{k=1}^{\frac{K}{P}} \alpha_n'(k)}{\frac{K}{P}} \tag{18}$$

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:

$$\nu_n^r = \frac{(f_n(\mathrm{DC}) - \mu_\mathbf{f})^2 + \sum_{k=1}^K \alpha_n(k) f_n(k)^2}{(f_n(\mathrm{DC}) - \mu_\mathbf{f})^2 + \sum_{k=1}^K f_n(k)^2 + C_3} \tag{19}$$

$$\nu_n^d = \frac{(g_n(\mathrm{DC}) - \mu_\mathbf{g})^2 + \sum_{k=1}^K \alpha_n(k) g_n(k)^2}{(g_n(\mathrm{DC}) - \mu_\mathbf{g})^2 + \sum_{k=1}^K g_n(k)^2 + C_3} \tag{20}$$

The constant C_3 is added to prevent numerical instability when the denominators of (19) or (20) go to 0. This can happen in smooth image regions.

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 are captured by (19) and (20).

The denominator terms in (19) and (20) 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

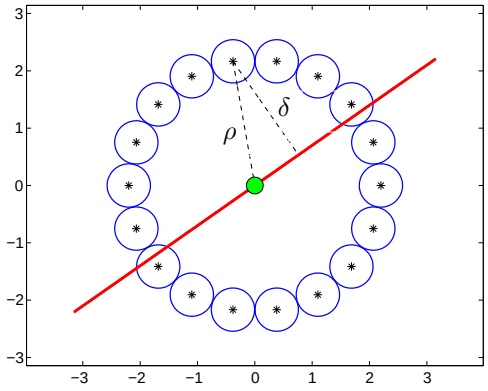

Fig. 3. A slice of the Gabor filters and the spectral plane shown in 2 dimensions. The horizontal axis denotes horizontal spatial frequency and the vertical axis denotes temporal frequency. Each circle represents a Gabor filter and the centers of each filter are also marked. The radius ρ of the single scale of Gabor filters and the distance δ of the center frequency of one Gabor filter from the spectral plane are marked.

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

$$E_T(\mathbf{i}_0) = \frac{1}{N}\sum_{n=1}^{N}(\nu_n^r - \nu_n^d)^2 \tag{21}$$

The error index in (21) is also exactly 0 when the reference and test images are identical. Finally, we convert the error index into a quality index using

$$Q_T(\mathbf{i}_0) = 1 - E_T(\mathbf{i}_0) \tag{22}$$

E. Pooling Strategy

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 $Q_T(\mathbf{i})$. The MOVIE index combines these local quality indices into a single score for the entire video. 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 refer to these frames of the quality videos, $Q_S(x, y, t_0)$ and $Q_T(x, y, t_0)$ for instance, as "quality maps".

To obtain a single score for the entire video using the local quality scores obtained at each pixel, several approaches such as probability summation using psychometric functions [26], [24], mean of the quality map [13], weighted summation [4], percentiles [42] and so on have been proposed. In general, the distribution of the quality scores depends on the nature of the scene content and the distortions. For example, distortions tend to occur more in "high activity" areas of the video sequences such as edges, textures and boundaries of moving objects. Similarly, certain distortions such as additive noise affect the entire video, while other distortions such as compression or packet loss in network transmission affect specific regions of the video. Selecting a pooling strategy is not an easy task since the strategy that humans use to evaluate quality based on their perception of an entire video sequence is not known.

We explored different pooling strategies and found that use of the the mean of the MOVIE quality maps as an indicator of the overall visual quality of the video suffered from certain drawbacks. Quality scores assigned to videos that contain a lot of textures, edges, moving objects and so on using the mean of the quality map as the visual quality predictor is consistently lower than quality scores computed for videos that contain smooth regions (backgrounds, objects). This is because many distortions such as compression alter the appearance of textures and other busy regions of the video much more significantly than the smooth regions of the video. However, people tend to assign poor quality scores even if only parts of the video appear to be distorted.

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 [42]. 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 perceptual error between the reference and test videos.

Define frame level *error* indices for both spatial and temporal components of MOVIE at a frame t_j using:

$$\mathrm{FE}_S(t_j) = \frac{\sigma_{Q_S(x,y,t_j)}}{\mu_{Q_S(x,y,t_j)}}, \quad \mathrm{FE}_T(t_j) = \frac{\sigma_{Q_T(x,y,t_j)}}{\mu_{Q_T(x,y,t_j)}} \tag{23}$$

Use of the coefficient of variation in pooling, with the standard deviation appearing in the numerators of (23), results in frame level error indices, as opposed to frame level quality indices. However, this ensures that the frame level MOVIE indices do not suffer from numerical instability issues due to very small values appearing in the denominator. The frame level error indices in (23) are exactly zero when the reference and distorted videos are identical, since $Q_S(x, y, t_j) = 1$ for all x, y. The error indices increase whenever the standard deviation of the MOVIE quality scores increases or the mean of the MOVIE quality scores decreases, which is desirable. Notice that the standard deviation term in the coefficient of variation captures the spread in quality that occurs when videos contain smooth regions, thus avoiding the drawback of using just the mean.

Video quality is fairly uniform over the duration of the video sequence (for instance, compression distortions behave this way) in the VQEG FRTV Phase 1 database that we use to evaluate MOVIE in Section VI. We adopted the simple pooling strategy of using the mean of the frame level descriptors for temporal pooling, although more advanced temporal pooling strategies may be investigated for future improvements of the MOVIE index. The Spatial MOVIE index is defined as the

average of these frame level descriptors.

$$\text{Spatial MOVIE} = \frac{1}{\tau}\sum_{j=1}^{\tau} \text{FE}_S(t_j) \tag{24}$$

The range of values of the Temporal MOVIE scores is smaller than that of the spatial scores, due to the large divisive normalization in (19) and (20). To offset this effect, we use the square root of the temporal scores.

$$\text{Temporal MOVIE} = \sqrt{\frac{1}{\tau}\sum_{j=1}^{\tau} \text{FE}_T(t_j)} \tag{25}$$

We adopt the simple strategy of defining the overall MOVIE index for a video using the product of the Spatial and Temporal MOVIE indices. This causes the MOVIE index to respond equally strongly to percentage changes in either the Spatial or Temporal MOVIE indices and makes MOVIE relatively insensitive to the range of values occupied by the Spatial and Temporal MOVIE indices. The MOVIE index is defined as:

$$\text{MOVIE} = \text{Spatial MOVIE} \times \text{Temporal MOVIE} \tag{26}$$

F. Implementation Details and Examples

We now discuss some implementation details of MOVIE. To reduce computation, instead of filtering the entire video sequence with the set of Gabor filters, we centered the Gabor filters on every 16^{th} frame of the video sequence and computed quality maps for only these frames. We selected multiples of 16 since our coarsest scale filters span 33 frames and using multiples of 16 ensures reasonable overlap in the computation along the temporal dimension. The window B was chosen to be a 7×7 window. To avoid blocking artifacts caused by a square window, we used a Gaussian window of standard deviation 1 sampled to a size of 7×7 [13]. If we denote the Gaussian window using $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \ldots, \gamma_N\}$ with $\sum_{n=1}^{N} \gamma_n = 1$, (3) and (4) are modified as:

$$E_S(\mathbf{i}_0, k) = \frac{1}{2}\sum_{n=1}^{N} \gamma_n \left[\frac{f_n(k) - g_n(k)}{M(k) + C_1}\right]^2 \tag{27}$$

$$M(k) = \max\left(\sqrt{\sum_{n=1}^{N} \gamma_n |f_n(k)|^2}, \sqrt{\sum_{n=1}^{N} \gamma_n |g_n(k)|^2}\right) \tag{28}$$

Similar modifications apply for (7), (8) and (9). (21) is modified as:

$$E_T(\mathbf{i}_0) = \sum_{n=1}^{N} \gamma_n (\nu_n^r - \nu_n^d)^2 \tag{29}$$

There are three parameters in MOVIE: C_1,C_2 and C_3. The role of these constants have been described in detail in [69]. The divisive nature of the masking model in (3) and (19) makes them extremely sensitive to regions of low signal energy in the video sequences. The constants serve to stabilize the computation in these regions and are included in most divisive normalization models [65], [67], [24], [64], [68]. We chose the parameters C_1, C_2 and C_3 to be of the same order of magnitude as the quantities in the denominators of (3), (8) and (19) that they are intended to stabilize. We selected the constants to be: $C_1 = 0.1$, $C_2 = 1$ and $C_3 = 100$. C_1, C_2 are chosen differently since the Gaussian filter is lowpass and produces larger responses than bandpass Gabor filters. This is intuitively reasonable from the power spectral properties of natural images [70]. C_3 is larger because it is intended to stabilize (19) and (20), where the denominator terms correspond to sums of the squares of all Gabor coefficients. We found that MOVIE is not very sensitive to the choice of constant as long as the constant used was not too small. Using small values for the constants leads to incorrect predictions of poor qualities in smooth regions of the videos due to the instability of the divisive models, which does not match visual perception.

Figure 4 illustrates quality maps generated by MOVIE on one of the videos in the VQEG FRTV Phase 1 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 suffers 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 edges of objects such as the harp where motion compensation mismatches are evident. Of course, the spatial and temporal quality values are not completely independent. This is because the spatial computation uses the outputs of *spatio-temporal* Gabor filters and the constant C_3 in (19) and (20) permits the temporal computation to respond to blur.

V. Relation to Existing Models

The MOVIE index has some interesting relationships to spatial IQA indices and to visual perception.

A. Spatial MOVIE

The spatial quality in (3) is closely related to contrast gain control models that use divisive normalization to model the response properties of neurons in the primary visual cortex [65], [67], [66]. Several HVS modeling based IQA algorithms account for contrast masking in human vision using divisive normalization models of contrast gain control [25], [64], [24]. Additionally, the spatial quality in (3) is closely related to the structure term of the SSIM index and the information theoretic basis of IQA [69]. Indeed, in previous work, we have established that the Gaussian Scale Mixture (GSM) image model assumption used by the information theoretic indices made them equivalent to applying the structure term of the SSIM index in a sub-band domain. Spatial MOVIE falls out naturally from our analysis in [69] and represents an improved version of these metrics.

We also discuss the relation of both SSIM and IFC to contrast masking models in human vision based IQA systems in [69]. The structure term of the SSIM index applied between sub-band coefficients (without the stabilizing constant and

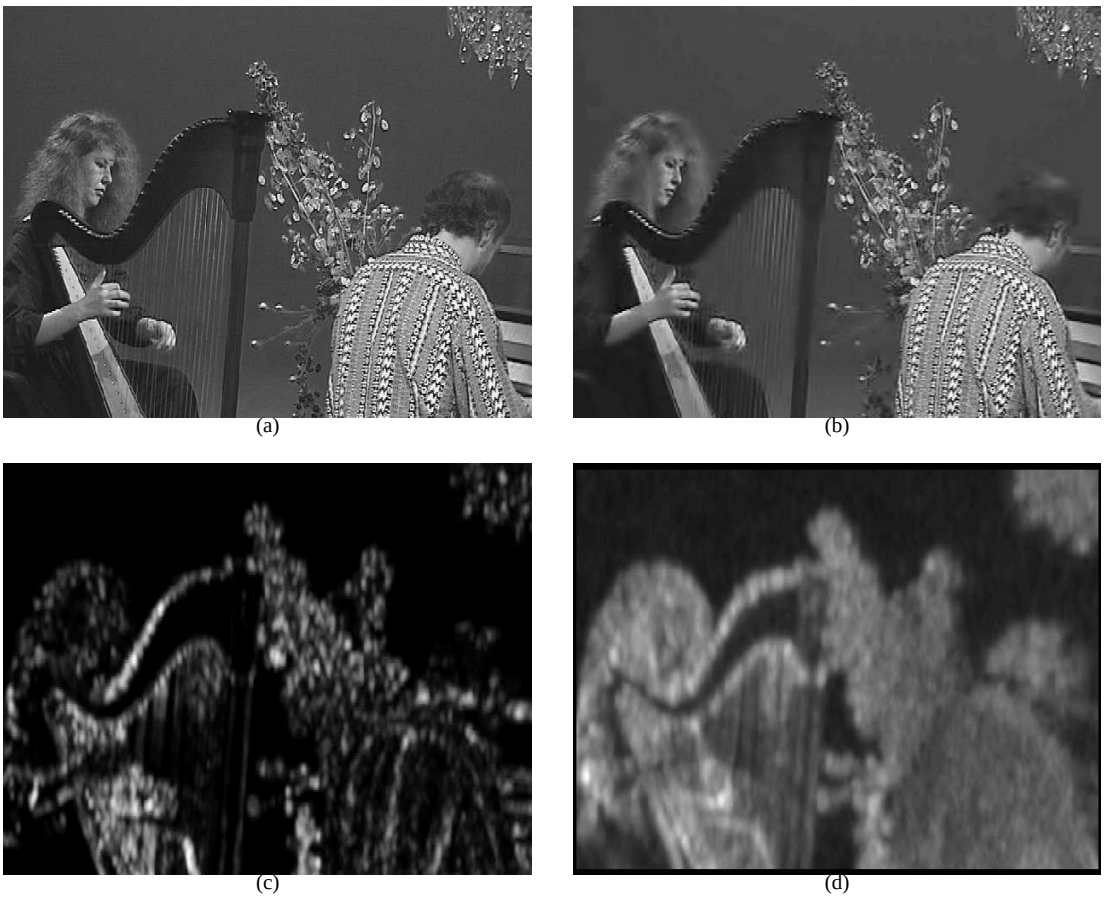

Fig. 4. Illustration of the performance of the MOVIE index. Top left - frame from reference video, Top right - corresponding frame from distorted video, Bottom left - logarithmically compressed temporal quality map, Bottom right - spatial quality map. Bright regions correspond to regions of poor quality.

assuming zero mean sub-band coefficients) is given by [69]:

$$\frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}\left[\frac{f_n(k)}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}|f_n(k)|^2}}-\frac{g_n(k)}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}|g_n(k)|^2}}\right]^2 \tag{30}$$

Divisive normalization is performed in (30), wherein divisive inhibition is modeled within the sub-band, while the divisive inhibition pool (in the denominator of (30)) is composed of coefficients from the same sub-band but at adjacent spatial locations. The divisive inhibition pool and divisive normalization model used here differ from other contrast gain control models. For example, Lubin models divisive inhibition within the same sub-band, while the Teo and Watson models seek to account for cross-channel inhibition [24], [25], [64].

A chief distinction between the divisive normalization in the SSIM index in (30) and the Spatial MOVIE index in (3) is the fact that we have chosen to utilize both the reference and distorted coefficients to compute the masking term. This is described as "mutual masking" in the literature [26]. Masking the reference and test image patches using a measure of their own signal energy in (30) ("self masking") is not an effective measure of blur in images and videos. Blur manifests itself as attenuation of certain sub-bands of the reference image and it is easily seen that the self masking model in (30) does not adequately capture blur.

However, our model is differs from mutual masking models such as [26], where the minimum of the masking thresholds computed from the reference and distorted images is used. Using a minimum of the masking thresholds is well suited for determining whether an observer can distinguish between the reference and test images, as in [26]. However, MOVIE is intended to predict the annoyance of supra-threshold, visible distortions. Using the maximum of the two masking thresholds in (3) causes the spatial quality index to saturate in the presence of severe distortions (loss of textures, severe blur, severe ringing and so on). This prevents over-prediction of errors in these regions. An additional advantage of using the maximum is that it guarantees bounded quality scores.

B. Temporal MOVIE

Motion perception is a complex procedure involving low-level and high-level processing. Although motion processing begins in the striate cortex (Area V1), Area MT/V5 in the extra-striate cortex is known to play a significant role in movement processing. Several papers in psychophysics and vision science study the properties of neurons in these areas in primates such as the macaque monkey. The properties of neurons in Area V1 that project to Area MT have been well studied [35]. This study reveals that cells in V1 that

project to MT may be regarded as local motion energy filters that are spatio-temporally separable and tuned to a specific frequency and orientation (such as the Gabor filters used here). Area MT receives directional information from V1 and performs more complex computations using the preliminary motion information computed by V1 neurons [35]. A subset of neurons in Area MT have been shown to be *speed tuned*, where the speed tuning of the neuron is independent of the spatial frequency of the stimulus [39], [71]. Models for such speed tuned neurons have been constructed by combining the outputs of a set of V1 cells whose orientation is consistent with the desired velocity [58]. Our temporal quality computation bears several similarities with the neuronal model of MT in [58], [72]. Similarities include the weighting procedure based on the distance between the linear filters and the motion plane and the normalization of weighted responses. The models in [58], [72] are rather elaborate, physiologically plausible mechanisms designed to match the properties of visual neurons. Our model is designed from an engineering standpoint of capturing distortions in videos. Differences between the two models include the choice of linear decomposition and our derivation of analytic expressions for the weights based on filter geometry. Interestingly, the models of Area MT construct neurons tuned to different speeds and use these responses to determine the speed of the stimulus. Our model computes the speed of motion using the Fleet and Jepson algorithm and then constructs speed tuned responses based on the computed motion.

To the best of our knowledge, none of the existing VQA algorithms attempt to model the properties of neurons in Area MT despite the availability of such models in the vision research community. Our discussion here shows that our proposed VQA framework can match visual perception of video better, since it integrates concepts from motion perception.

VI. Performance

We tested our algorithm on the VQEG FRTV Phase 1 database [73] since this is the largest publicly available VQA database to date. Although the VQEG has completed and is in the process of conducting several other studies on video quality, the videos from these subsequent studies have not been made public due to licensing and copyright issues [74]. Since most of the videos in the VQEG FRTV Phase 1 database 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, although color might represent a direction for future improvements of MOVIE. 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 [73]. We used the scores assigned in the "low" quality regime as the subjective scores for these videos.

Table I show the performance of MOVIE in terms of the Spearman Rank Order Correlation Coefficient (SROCC), the

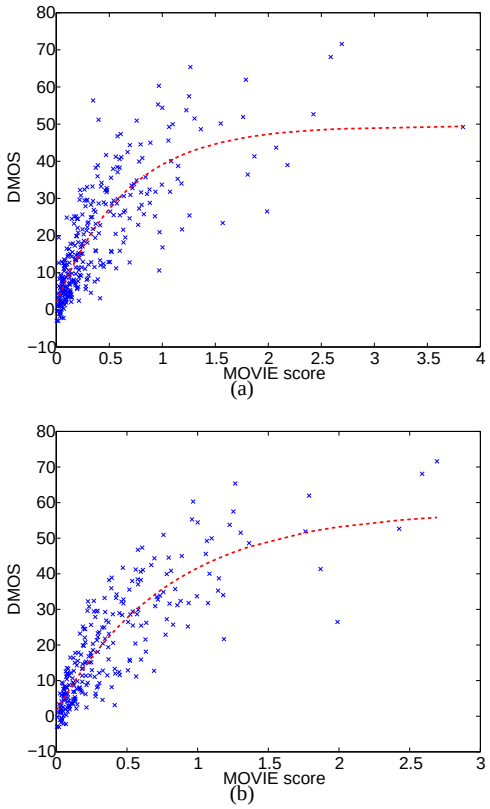

Fig. 5. Scatter plot of the subjective DMOS scores against MOVIE scores on the VQEG database. Each point on the plot represents one video in the database. The best fitting logistic function used for non-linear regression is also shown. (a) On all sequences in the VQEG database (b) After omitting the animated videos.

Linear Correlation Coefficient (LCC) after non-linear regression and the Outlier Ratio (OR). We used the same logistic function specified in [73] to fit the model predictions to the subjective data. PSNR provides a baseline for comparison of VQA models. Ten leading VQA models were tested by the VQEG in its Phase 1 study including a model from NTIA that was a precursor to VQM, as well as models from NASA, Sarnoff Corporation, KDD and EPFL [73]. Proponent P8 (Swisscom) was the best performing model of these ten models tested by the VQEG [73]. SSIM (without weighting) refers to a frame-by-frame application of the SSIM index that was proposed for video in [4]. SSIM (weighting) refers to the model in [4] that incorporated rudimentary motion information as weights for different regions of the video sequence. Speed SSIM refers to the VQA algorithm in [5] that incorporates a model of human visual speed perception to design spatiotemporal weighting factors that are used to weight local SSIM indices in the pooling stage.

The Root Mean Squared Error (RMSE) between subjective scores and MOVIE scores after non-linear regression on the entire VQEG database is 8.76. Outliers are defined by the VQEG as points for which the absolute error between the DMOS score and model prediction is larger than twice the standard deviation of the DMOS score and the outlier ratio is defined as the ratio of the number of outlier videos to the total number of videos [73]. A more standard way to

define outliers is using the three sigma rule, where outliers are defined as points for which the absolute error between the DMOS score and model prediction is larger than three times the standard deviation of the DMOS scores [75]. Use of the three sigma rule guarantees that the probability that a point lies outside the range of three standard deviations is $\leq 0.3\%$ assuming that the errors are normally distributed. The outlier ratio for MOVIE using the three sigma rule is 0.488 on the entire VQEG database.

The VQEG database contains 4 sequences that are animated (sources 4,6,16 and 17). Animated videos are quite distinct from natural videos and often contain perfectly smooth and constant regions, perfect step edges, text and so on that seldom occur in natural images. Natural images have several characteristic statistical properties such as self-similarity across scales, heavy tailed wavelet marginal distributions and so on [70], [76], that do not occur in synthetic videos of these types. Although our model does not explicitly assume any statistical model for the images or videos, our spatial quality model is closely related to the IFC, which assumes that the reference images are the output of a natural scene statistical model [69]. Several aspects of our VQA model such as the choice of Gabor filters, scale invariant processing of the Gabor sub-bands and divisive normalization in the spatial and temporal quality computation are implicitly geared toward natural videos. Indeed, it has been suggested that the divisive normalization that is used in both Spatial and Temporal MOVIE results in efficient encoding, since it reduces the statistical dependencies that are present when natural images are decomposed using linear filters [68]. Hence, the divisive normalization in MOVIE can be interpreted as a dual of natural scene statistical modeling. A further discussion of the relation between natural scene statistics and the SSIM and IFC IQA techniques can be found in [69]. The presence of text in three of these animations is further cause for concern, since the subjective perception of these videos might have been influenced by the readability of the text in the distorted video.

We also present performance indices of our VQA model for *only* the 16 natural videos and their distorted versions (a total of 256 videos) in the VQEG database in Table II. We present these results in a separate table since these numbers are not directly comparable against the reported performance of other quality models on all the videos in the database. Table II also shows the performance of PSNR and SSIM (without weighting) on the same set of natural videos in the VQEG database. For a fair comparison with MOVIE, we used only the luminance component of the video to compute PSNR and SSIM (without weighting) on these natural videos. Note that the performance of PSNR and SSIM is slightly worse in Table II than on the entire dataset as reported in Table I. The outlier ratio for MOVIE on only the natural videos is 0.461 at three standard deviations.

Scatter plots of the model prediction and DMOS values, along with the best fitting logistic function, for the MOVIE index are shown in Fig. 5 on the entire VQEG database and after omitting animations.

It is clear that the MOVIE index is competitive with other leading algorithms on the VQEG database. Note that the reported performance of the VQEG proponents is from [73], where the proponents did not have access to the VQEG database. The performance of some of these algorithms have been improved since the publication of the study in 2000 [73]. VQM from NTIA is the only publicly available algorithm of the ten proponents in the study. However, since VQM was trained using the VQEG data, we are unable to report the performance of VQM on the VQEG dataset [42]. None of the parameters of the MOVIE index were trained using the VQEG data. The results in Tables I and II show the competitive performance of MOVIE with other leading VQA techniques whose performance has been reported on the VQEG dataset in the eight years since the study.

Prediction Model	SROCC	LCC	OR
Peak Signal to Noise Ratio	0.786	0.779	0.678
Proponent P8 (Swisscom)	0.803	0.827	0.578
SSIM (without weighting)	0.788	0.820	0.597
SSIM (weighting)	0.812	0.849	0.578
Spatial MOVIE	0.793	0.796	0.666
Temporal MOVIE	0.816	0.801	0.647
MOVIE	0.833	0.821	0.644

TABLE I
COMPARISON OF THE PERFORMANCE OF VQA ALGORITHMS USING SROCC, LCC AND OR.

Prediction Model	SROCC	LCC	OR	RMSE
PSNR	0.739	0.718	0.699	10.968
SSIM (without weighting)	0.802	0.810	0.633	9.245
Spatial MOVIE	0.825	0.830	0.656	8.803
Temporal MOVIE	0.835	0.825	0.621	8.902
MOVIE	0.860	0.858	0.656	8.093

TABLE II
COMPARISON OF THE PERFORMANCE OF VQA ALGORITHMS AFTER OMITTING THE ANIMATION SEQUENCES USING SROCC, LCC, OR AND RMSE. PSNR AND SSIM (WITHOUT WEIGHTING) ARE COMPUTED USING ONLY THE LUMINANCE COMPONENT OF THE VIDEO IN THIS TABLE FOR A FAIR COMPARISON WITH MOVIE.

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. It is also seen from Table II that the performance of MOVIE is considerably better on just the natural videos in the VQEG database. The performance of MOVIE in Table I is particularly impressive because it does not use color information and uses only one field of the interlaced video sequence.

VII. CONCLUSIONS AND FUTURE WORK

We have introduced a new, motion-based paradigm for VQA that successfully captures temporal distortions as well as spatial distortions. The performance of the resulting algorithm, known as MOVIE, bears out the underlying philosophy that such distortions contribute significantly to the perception of video quality, and is in agreement with physiological findings. An obvious avenue for improving MOVIE that we wish to investigate is the inclusion of color information. Additionally, there is a need for more diverse publicly available databases of

reference videos, distorted videos, and statistically significant subjective scores taken under carefully controlled measurement conditions to enable improved verification and testing of VQA algorithms. Such a database will be of great value to the VQA research community, particularly in view of the fact that the videos from recent VQEG studies (including the VQEG FRTV-Phase 2 study and the Multimedia study) are not being made public [74]. Toward this end, we are creating such a database of videos that will complement the existing LIVE Image Quality Database [77] and which seeks to improve the accessibility and diversity of such data. The upcoming LIVE Video Quality Database will be described in future reports.

Lastly, there naturally remains much open field for improving current competitive VQA algorithms. We believe that these will be improved by the development of better models for naturalistic videos, for human image and motion processing, and by a better understanding of the nature of distortion perception. Important topics in these directions include scalability of VQA, utilizing models of visual attention and human eye movements in VQA [78], [79], [80], [6], exploration of advanced spatial and temporal pooling strategies for VQA [80], reduced reference VQA, and no reference VQA. However, in our view, the most important development in the future of both IQA and VQA is the deployment of the most competitive algorithms for such diverse and important tasks as establishing video Quality of Service (QoS) in real-time applications; benchmarking the performance of competing image and video processing algorithms, such as compression, restoration, and reconstruction; and optimizing algorithms using IQA and VQA indices to establish perceptual objective functions [81]. This latter goal is the most ambitious owing to the likely formidable analytical challenges to be overcome, but may also prove to be the most significant.

APPENDIX
OPTICAL FLOW COMPUTATION VIA A NEW MULTI-SCALE APPROACH

The Fleet and Jepson algorithm attempts to find constant phase contours of the outputs of a Gabor filterbank to estimate the optical flow vectors [60]. Constant phase contours are computed by estimating the derivative of the phase of the Gabor filter outputs, which in turn can be expressed as a function of the derivative of the Gabor filter outputs [60]. The algorithm in [60] uses a 5-point central difference to perform the derivative computation. However, we chose to perform the derivative computation by convolving the video sequence with filters that are derivatives of the Gabor kernels, denoted by $h'_x(\mathbf{i}), h'_y(\mathbf{i}), h'_t(\mathbf{i})$:

$$h'_x(\mathbf{i}) = h(\mathbf{i}) \left(\frac{-x}{\sigma^2} + jU_0 \right) \tag{31}$$

Similar definitions apply for the derivatives along y and t directions. This filter computes the derivative of the Gabor outputs more accurately and produced better optical flow estimates in our experiments.

Due to the aperture problem, each Gabor filter is only able to signal the component of motion that is normal to its own orientation. The Fleet and Jepson algorithm computes normal velocity estimates at each pixel for each Gabor filter. Given the normal velocities from the different Gabor outputs, a linear velocity model is fit to each local region using a least squares criterion to obtain a 2D velocity estimate at each pixel of the video sequence. A residual error in the least squares solution is also obtained at this stage. See [60], [82] for further details.

The original Fleet and Jepson algorithm uses just a single scale of filters. We found that using a single scale of filters was not sufficient, since optical flow was not computed in fast moving regions of the several video sequences due to temporal aliasing [60], [57]. We hence used 3 scales of filters to compute motion by extending the Fleet and Jepson algorithm 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 [57], [60]. 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. Note that more complex solutions such as coarse to fine warping methods have been proposed in the literature to combine flow estimates across scales [83], [84], [85]. We chose this approach for simplicity and found that reasonable results were obtained.

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. Instead, optical flow is only computed at pixels where there is sufficient information to do so. We set the optical flow to zero at all pixels where the flow was not computed.

REFERENCES

[1] Z. Wang and A. C. Bovik, *Image Quality Assessment*. New York: Morgan and Claypool Publishing Co., 2006.

[2] S. Winkler, *Digital Video Quality*. New York: Wiley and Sons, 2005.

[3] C. J. van den Branden Lambrecht, D. M. Costantini, G. L. Sicuranza, and M. Kunt, "Quality assessment of motion rendition in video coding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 5, pp. 766–782, 1999.

[4] 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.

[5] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception." *Journal Optical Society America A: Optics Image Science Vision*, vol. 24, no. 12, pp. B61–B69, Dec 2007.

[6] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 253–265, 2009.

[7] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 266–279, 2009.

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

[9] K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," in *IEEE Intl. Conf. Acoustics, Speech, and Signal Processing*, 2007.

[10] B. A. Wandell, *Foundations of Vision*. Sunderland, MA: Sinauer Associates Inc., 1995.

[11] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[12] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, 2002.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, April 2004.

[14] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[15] 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 Process.*, vol. 14, no. 12, pp. 2117–2128, 2005.

[16] H. R. Sheikh and A. C. Bovik, "An evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, November 2006.

[17] K. Seshadrinathan and A. C. Bovik, "An information theoretic video quality metric based on motion models," in *Third Intl. Workshop Video Processing and Quality Metrics for Consumer Electronics*, 2007.

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

[19] K. Seshadrinathan and A. C. Bovik, "Video quality assessment," in *The Essential Guide to Video Processing*, A. C. Bovik, Ed. Elsevier, 2009.

[20] J. M. Libert, C. P. Fenimore, and P. Roitman, "Simulation of graded video impairment by weighted summation: validation of the methodology," *Proc. SPIE*, vol. 3845, no. 1, pp. 254–265, Nov. 1999.

[21] C. Lee and O. Kwon, "Objective measurements of video quality using the wavelet transform," *Optical Engineering*, vol. 42, no. 1, pp. 265–272, Jan. 2003.

[22] C. Taylor and S. Dey, "Run-time allocation of buffer resources for maximizing video clip quality in a wireless last-hop system," in *Proc. IEEE Intl. Conf. Communications*, 2004.

[23] J. Mannos and D. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Trans. Inf. Theory*, vol. 20, no. 4, pp. 525–536, 1974.

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

[25] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. IEEE Intl. Conf. Image Processing*, 1994.

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

[27] D. M. Chandler, K. H. Lim, and S. S. Hemami, "Effects of spatial correlations and global precedence on the visual fidelity of distorted images," *Proc. SPIE*, vol. 6057, no. 1, p. 60570F, Feb 2006.

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

[29] G. E. Legge, "Sustained and transient mechanisms in human vision: Temporal and spatial properties," *Vision Research*, vol. 18, no. 1, pp. 69–81, 1978.

[30] J. J. Kulikowski and D. J. Tolhurst, "Psychophysical evidence for sustained and transient detectors in human vision." *Journal Physiology*, vol. 232, no. 1, pp. 149–162, Jul 1973.

[31] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatiotemporal model of the human visual system," *Proc. SPIE*, vol. 2668, no. 1, pp. 450–461, Mar 1996.

[32] S. Winkler, "Perceptual distortion metric for digital color video," *Proc. SPIE*, vol. 3644, no. 1, pp. 175–184, May 1999.

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

[34] M. Masry, S. S. Hemami, and Y. Sermadevi, "A scalable wavelet-based video distortion metric and applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 2, pp. 260–273, 2006.

[35] J. A. Movshon and W. T. Newsome, "Visual Response Properties of Striate Cortical Neurons Projecting to Area MT in Macaque Monkeys," *Journal Neuroscience*, vol. 16, no. 23, pp. 7733–7741, 1996.

[36] R. T. Born and D. C. Bradley, "Structure and function of visual area MT," *Annual Reviews Neuroscience*, vol. 28, pp. 157–189, 2005.

[37] M. A. Smith, N. J. Majaj, and J. A. Movshon, "Dynamics of motion signaling by neurons in macaque area MT," *Nature Neuroscience*, vol. 8, no. 2, pp. 220–228, Feb. 2005.

[38] J. A. Perrone, "A visual motion sensor based on the properties of V1 and MT neurons," *Vision Research*, vol. 44, no. 15, pp. 1733–1755, Jul. 2004.

[39] N. J. Priebe, S. G. Lisberger, and J. A. Movshon, "Tuning for spatiotemporal frequency and speed in directionally selective neurons of macaque striate cortex." *Journal Neuroscience*, vol. 26, no. 11, pp. 2941–2950, Mar 2006.

[40] J. Nachmias and R. V. Sansbury, "Grating contrast: Discrimination may be better than detection," *Vision Research*, vol. 14, no. 10, pp. 1039–1042, Oct. 1974.

[41] G. Legge and J. Foley, "Contrast masking in human vision," *Journal Optical Society America A: Optics Image Science Vision*, vol. 70, no. 12, pp. 1458–1471, Dec. 1980.

[42] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

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

[44] International Telecommunications Union, "Objective perceptual multimedia video quality measurement in the presence of a full reference," ITU-T Rec. J. 247, Tech. Rep., 2008.

[45] NTT. (2008) NTT News Release. [Online]. Available: http://www.ntt.co.jp/news/news08e/0808/080825a.html

[46] Opticom. [Online]. Available: http://www.opticom.de/technology/pevq_video-quality-testing.html

[47] M. Malkowski and D. Claben, "Performance of video telephony services in UMTS using live measurements and network emulation," *Wireless Personal Communications*, vol. 1, pp. 19–32, 2008.

[48] M. Barkowsky, J. Bialkowski, R. Bitto, and A. Kaup, "Temporal registration using 3D phase correlation and a maximum likelihood approach in the perceptual evaluation of video quality," in *IEEE Workshop Multimedia Signal Processing*, 2007.

[49] M. Yuen and H. R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Processing*, vol. 70, no. 3, pp. 247–278, Nov. 1998.

[50] K. Seshadrinathan and A. C. Bovik, "Motion-based perceptual quality assessment of video," in *Proc. SPIE - Human Vision and Electronic Imaging*, 2009.

[51] J. A. Movshon, I. D. Thompson, and D. J. Tolhurst, "Spatial summation in the receptive fields of simple cells in the cat's striate cortex." *Journal Physiology*, vol. 283, pp. 53–77, Oct 1978.

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

[53] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 55–73, Jan. 1990.

[54] D. J. Tolhurst and J. A. Movshon, "Spatial and temporal contrast sensitivity of striate cortical neurones," *Nature*, vol. 257, no. 5528, pp. 674–675, Oct. 1975.

[55] S. M. Friend and C. L. Baker, "Spatio-temporal frequency separability in area 18 neurons of the cat." *Vision Research*, vol. 33, no. 13, pp. 1765–1771, Sep 1993.

[56] M. C. Morrone, M. D. Stefano, and D. C. Burr, "Spatial and temporal properties of neurons of the lateral suprasylvian cortex of the cat." *Journal Neurophysiology*, vol. 56, no. 4, pp. 969–986, Oct 1986.

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

[58] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT." *Vision Research*, vol. 38, no. 5, pp. 743–761, Mar 1998.

[59] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion." *Journal Optical Society America A: Optics Image Science Vision*, vol. 2, no. 2, pp. 284–299, Feb 1985.

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

[61] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *First Intl. workshop video processing and quality metrics for consumer electronics*, 2005.

[62] R. Fox, "Visual masking," in *Handbook of Sensory Physiology. VIII. Perception*, R. Held, H. W. Leibowitz, and H. L. Teuber, Eds. Springer-Verlag, 1978.

[63] J. Foley, "Human luminance pattern-vision mechanisms: masking experiments require a new model," *Journal Optical Society America A: Optics Image Science Vision*, vol. 11, no. 6, pp. 1710–1719, Jun. 1994.

[64] A. Watson and J. Solomon, "Model of visual contrast gain control and pattern masking," *Journal Optical Society America A: Optics Image Science Vision*, vol. 14, no. 9, pp. 2379–2391, Sep. 1997.

[65] D. J. Heeger, "Normalization of cell responses in cat striate cortex." *Visual Neuroscience*, vol. 9, no. 2, pp. 181–197, Aug 1992.

[66] D. G. Albrecht and W. S. Geisler, "Motion selectivity and the contrast-response function of simple cells in the visual cortex." *Visual Neuroscience*, vol. 7, no. 6, pp. 531–546, Dec 1991.

[67] W. S. Geisler and D. G. Albrecht, "Cortical neurons: isolation of contrast gain control." *Vision Research*, vol. 32, no. 8, pp. 1409–1410, Aug 1992.
[68] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control." *Nature Neuroscience*, vol. 4, no. 8, pp. 819–825, Aug 2001.
[69] K. Seshadrinathan and A. C. Bovik, "Unifying analysis of full reference image quality assessment," in *IEEE Intl. Conf. Image Processing*, 2008.
[70] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells." *Journal Optical Society America A: Optics Image Science Vision*, vol. 4, no. 12, pp. 2379–2394, Dec 1987.
[71] J. A. Perrone and A. Thiele, "Speed skills: measuring the visual speed analyzing properties of primate MT neurons," *Nature Neuroscience*, vol. 4, no. 5, pp. 526–532, May 2001.
[72] N. C. Rust, V. Mante, E. P. Simoncelli, and J. A. Movshon, "How MT cells analyze the motion of visual patterns." *Nature Neuroscience*, vol. 9, no. 11, pp. 1421–1431, Nov 2006.
[73] (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
[74] A. Webster, "Progress and future plans for VQEG," in *ETSI STQ Workshop Multimedia Quality of Service*, 2008. [Online]. Available: http://portal.etsi.org/docbox/Workshop/2008/2008_06_STQWORKSHOP/VQEG_ArthurWebster.pdf
[75] S. H. Dai and M. O. Wang, *Reliability analysis in engineering applications*. Van Nostrand Reinhold, 1993.
[76] E. P. Simoncelli, "Statistical modeling of photographic images," in *Handbook of Image and Video Processing*, 2nd ed., A. C. Bovik, Ed. Academic Press, 2005.
[77] (2003) LIVE image quality assessment database. [Online]. Available: http://live.ece.utexas.edu/research/quality/subjective.htm
[78] A. K. Moorthy and A. C. Bovik, "Perceptually significant spatial pooling strategies for image quality assessment," in *Proc. SPIE - Human Vision and Electronic Imaging*, 2009.
[79] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "GAFFE: A gaze-attentive fixation finding engine," *IEEE Trans. Image Process.*, vol. 17, no. 4, pp. 564–573, 2008.
[80] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 193–201, 2009.
[81] S. S. Channappayya, A. C. Bovik, C. Caramanis, and R. W. Heath, Jr., "Design of linear equalizers optimized for the structural similarity index," *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 857–872, 2008.
[82] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *Intl. Journal Computer Vision*, vol. 12, no. 1, pp. 43–77, Feb. 1994.
[83] E. P. Simoncelli, "Distributed analysis and representation of visual motion," Ph.D. dissertation, MIT, 1993.
[84] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *Intl. Journal Computer Vision*, vol. 2, no. 3, pp. 283–310, Jan. 1989.
[85] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Intl. Joint Conf. Artificial Intelligence*, Vancouver, Canada, 1981.

Kalpana Seshadrinathan Kalpana Seshadrinathan received the B.Tech. degree from the University of Kerala, India in 2002 and the M.S. and Ph.D. degrees in Electrical Engineering from the University of Texas at Austin, in 2004 and 2008, respectively. She is currently a System Engineer with Intel Corporation in Phoenix, AZ. Her research interests include image and video quality assessment, computational aspects of human vision, motion estimation and its applications and statistical modeling of images and video. She is a recipient of the 2003 Texas Telecommunications Engineering Consortium Graduate Fellowship and the 2007 Graduate Student Professional Development Award from the University of Texas at Austin. She was Assistant Director of the Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin from 2005-2008. She is a member of the IEEE.

Alan Conrad Bovik Alan Conrad Bovik is the Curry/Cullen Trust Endowed Chair Professor at The University of Texas at Austin, where he is the Director of the Laboratory for Image and Video Engineering (LIVE). He is a faculty member in the Department of Electrical and Computer Engineering, the Department of Biomedical Engineering, and the Institute for Neuroscience. His research interests include image and video processing, computational vision, and visual perception. He has published over 500 technical articles in these areas and holds two U.S. patents. He is the author of *The Handbook of Image and Video Processing* (Academic Press, 2005), *Modern Image Quality Assessment* (Morgan & Claypool, 2006), and two new books, *The Essential Guide to Image Processing* and *The Essential Guide to Video Processing* (Academic Press).

Dr. Bovik has received a number of major awards from the IEEE Signal Processing Society, including: the Education Award (2008); the Technical Achievement Award (2005), the Distinguished Lecturer Award (2000); and the Meritorious Service Award (1998). He is also a recipient of the Hocott Award for Distinguished Engineering Research at the University of Texas at Austin; received the Distinguished Alumni Award from the University of Illinois at Champaign-Urbana (2008), the IEEE Third Millennium Medal (2000) and two journal paper awards from the international Pattern Recognition Society (1988 and 1993). He is a Fellow of the IEEE, a Fellow of the Optical Society of America, and a Fellow of the Society of Photo-Optical and Instrumentation Engineers. He has been involved in numerous professional society activities, including: Board of Governors, IEEE Signal Processing Society, 1996-1998; Editor-in-Chief, *IEEE Transactions on Image Processing*, 1996-2002; Editorial Board, *The Proceedings of the IEEE*, 1998-2004; Series Editor for Image, Video, and Multimedia Processing, Morgan and Claypool Publishing Company, 2003-present; and Founding General Chairman, *First IEEE International Conference on Image Processing*, held in Austin, Texas, in November, 1994.

Dr. Bovik is a registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial and academic institutions.