FLICKER SENSITIVE MOTION TUNED VIDEO QUALITY ASSESSMENT

Lark Kwon Choi and Alan C. Bovik

Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA
larkkwonchoi@utexas.edu, bovik@ece.utexas.edu

Abstract—From a series of human subjective studies, we have found that large motion can strongly suppress flicker visibility. Based on the spectral analysis of flicker videos in frequency domain, we propose a full reference video quality assessment (VQA) framework that incorporates flicker sensitive temporal visual masking. The framework predicts perceptually silenced flicker visibility using a model of the responses of primary visual cortex to video flicker, a motion energy model, and divisive normalization. By incorporating perceptual flicker visibility into motion tuned video quality measurements as in the MOVIE framework, we augment VQA performance with sensitivity to flicker. Results show that the proposed VQA framework correlates well with human results and is highly competitive with recent state-of-the-art VQA algorithms tested on the LIVE VQA database.

Keywords—temporal visual masking; flicker visibility; video quality assessment; motion perception; motion silencing.

I. INTRODUCTION

Global video traffic over networks is exponentially growing. Gigantic amounts of video content are available on mobile devices due to proliferating video technology, while users’ expectation for higher video quality is increasing. The Cisco Visual Networking Index [1] reports that mobile video immediately already impacts traffic, and will increasingly do so in the future generating more than 69 percent of mobile data by 2019. Developing more accurate, automatic VQA tools is important to provide more satisfactory levels of quality of experience to the end user by optimizing limited resources such as bandwidth and power consumption in end-to-end distortion vulnerable video delivery systems [2].

Understanding how humans perceive visual artifacts and modeling the visibility of video distortions are important for developing VQA algorithms, since humans are the ultimate arbiter of digital videos [3]. Based on substantial progress towards modeling low-level visual processing in the human visual system (HVS), a variety of successful VQA models have been proposed. Structural Similarity (SSIM) uses visual sensitivity to luminance (contrast) changes and to variations of structural information [4], while MOTion-based Video Integrity Evaluation (MOVIE) uses a model of extra-cortical Area Middle Temporal (MT) [5]. Models of multiscale and orientation properties, disruptions to natural scene statistics, and visual masking have been also widely used [4], [6], [7].

With regard to distortion visibility, the mere presence of spatial/temporal video distortions does not imply perceptual quality degradation, since distortion visibility can be strongly reduced by visual masking. As compared to spatial masking, temporal masking is not well-modeled although one type of temporal masking has been observed to occur near scene changes [8], and has been used in the development of early stage video compression algorithms [9], [10].

Although temporal masking is not yet well modeled, the phenomenon is very evident. Recently, Suchow and Alvarez [11] demonstrated a striking “motion silencing” illusion, where the salient temporal changes of objects in luminance, color, size, and shape appear to cease when objects move fast in collective motions. The motion silencing phenomenon implies that commonly occurring annoying temporal flicker distortions and motion in digital videos may be dramatically reduced by the presence of object motion. Plausible explanations have been proposed using human psychophysics [11-13], and a series of human studies have been executed on naturalistic videos to better understand motion silencing effects [14-16].

Motion plays a significant role in understanding temporal distortions. Hence, motion perception models in the HVS have been adopted in recent VQA algorithms [17]. MOVIE captures temporal distortions along computed motion trajectories using a motion-tuned spatiotemporal VQA framework. Specifically, the Temporal MOVIE Index computes the misaligned spectral signatures of local patches between reference and distorted videos using excitatory-inhibitory weights that mimic motion processing in Area MT [18], then evaluates motion-tuned temporal video integrity.

However, the weights in the Temporal MOVIE Index are defined only as a function of the distance from the motion tuning spectral plane of the reference video without regard to object velocity, where the same distance implies the same amount of temporal distortions. From a series of human subjective studies [14-16], we have found that large coherent object motions strongly suppress the visibility of flicker distortions in moving regions.

In this paper we propose a new VQA framework that represents flicker sensitive temporal visual masking. The framework predicts perceptually suppressed flicker visibility using models of the cortical responses of primary visual cortex to video flicker via Gabor linear decomposition, a motion energy model, and a divisive normalization stage. By injecting a perceptual flicker visibility index into the well-known MOVIE framework, the flicker sensitive framework not only measures motion-tuned video integrity, but also predicts temporal masking of flicker distortions, thereby substantially improving the prediction of video quality.

The remainder of this paper is organized as follows. Section II describes the flicker sensitive motion tuned VQA framework. We evaluate the performance of the framework in Section III and conclude the paper in Section IV.
II. Flicker Sensitive Motion Tuned Framework

A. Gabor Decomposition

The HVS promptly and efficiently encodes visual signals using multiscale, multi-orientation, multi-direction subband decomposition. The receptive field profiles of simple cells in primary visual cortex of each subband can be well-modeled as linear, bandpass Gabor filters [19], [20]. Hence, we used a multiscale spatiotemporally separable Gabor filter bank to model the responses of V1 neurons to videos. A space time 3D separable Gabor filter $h(x)$ is the product of a complex exponential with a Gaussian envelope:

$$h(x) = \frac{1}{(2\pi)^{1/2} \sqrt{|\Sigma|}} \exp \left( -\frac{x^T \Sigma^{-1} x}{2} \right) \exp(j U_0 x),$$

where $x = (x, y, t)$ is a spatiotemporal coordinate in a video sequence, $U_0 = (U_0, V_0, W_0)$ is the space-time center frequency of the Gabor filter, and $\Sigma$ is the covariance matrix of the Gaussian envelope.

In our flicker sensitive motion tuned framework, we first linearly decompose reference and test videos using a 3D Gabor filter bank, as illustrated in Figure 1. The Gabor filter bank is implemented similar to [5] and [21], but we use a wider range of possible speeds to more accurately measure the spatiotemporal cortical responses of V1 neurons, with filter bandwidths 0.45 octaves. Three scales of Gabor filters with 57 filters sample each scale on the surface of a sphere centered at the space-time frequency origin. The largest radial center frequency was 0.7π radians per sample and the filters are sampled out to a width of three standard deviations. A total of 171 filters are used: 10, 18, 15, 10, and 4 filters tuned to five different speeds, $s = \tan(\phi)$, where the vertical angle $\phi = 0$ 20, 40, 60 and 80 degrees and orientations $\theta$ at every 18, 20, 24, 36, and 90 degrees, respectively. The number of oriented filters was determined such that adjoining Gabor filters intersect at one standard deviation [21]. We also included a Gaussian filter centered at the frequency origin to obtain the low frequencies in the video, where the Gaussian filter intersects the coarsest scale of bandpass filters at one standard deviation.

B. Cortical Neuron Model

The outputs of quadrature pairs of linear Gabor filters are squared and summed to model the motion energy responses of V1 simple cells within each subband as follows [22]:

$$E(\phi, \theta) = |h_{\sin}(\phi, \theta) \ast I|^2 + |h_{\cos}(\phi, \theta) \ast I|^2,$$  \hspace{1cm} (2)

where $h_{\sin}(\phi, \theta)$ and $h_{\cos}(\phi, \theta)$ are sine and cosine Gabor filters at $\phi$ and $\theta$. $I$ is the luminance level of a video, while the symbol $\ast$ means convolution.

The individual responses are then divisive normalized to model the collective nonlinearity of adaptive gain control of V1 complex cells [23]. The divisive normalization process limits individual dynamic range of simple cell responses to combine all cortical neighborhoods that include cells tuned for the full range of orientations and directions. Specifically, the response of simple cell $S(\phi, \theta)$ is modeled by dividing an individual energy response by the sum of the neighboring energy responses. Then, the model V1 complex cell response $C(\phi, \theta)$ is obtained by averaging $S(\phi, \theta)$ along scales on constant space-time frequency orientations:

$$S(\phi, \theta) = K \frac{\sum_{\phi, \theta} E(\phi, \theta)}{\sum_{\phi, \theta} E(\phi, \theta) + \sigma^2},$$

$$C(\phi, \theta) = \sum_{m} c_m S_m(\phi, \theta),$$

where $K$ determines the maximum attainable response, and $\sigma$ is a semi-saturation constant. Here $K = 4$ and $\sigma = 0.2$ as was used in [23] in agreement with recorded physiological data. We used constant values $c_m (> 0)$ as weighting factors.

C. Spatial Video Quality

Spatial video quality is predicted using the Gabor responses from the reference and test videos as in the Spatial MOVIE framework [5], but we apply a wider range of frequency subbands than does Spatial MOVIE. We measure spatial errors from each subband Gabor response and the DC subband Gaussian filter output using divisive normalization. Next, the spatial error indices are averaged over a sliding $7 \times 7$ patch to obtain an error index at location $x$:

$$Q_s(x) = 1 - \frac{1}{M+1} \sum_{m \neq 0} \frac{Err_S(x, \phi, \theta) + Err_{DC}(x)}{M+1},$$  \hspace{1cm} (5)

where $Err_S$ and $Err_{DC}$ are the spatial errors from each subband and DC, respectively, while $M$ is the total number of Gabor filters. Next, the averaged spatial errors in a frame are converted into a single spatial quality index using the coefficient of variation (CoV) of $Q_s$ values in (5) for each frame, then the frame basis CoV values are averaged to obtain a final spatial video quality score [5].

D. Temporal Video Quality

To measure perceptual temporal video quality, flicker visibility is predicted and combined with the responses of the Temporal MOVIE framework. We first compute a weighted sum of the Gabor filter outputs using the Temporal MOVIE framework, where excitatory-inhibitory weights are assigned to each individual Gabor filter as a function of its distance...
from the motion-tuned spectral plane of a reference video as shown in Figure 2 [5]. Any misaligned spectrum from the spectral plane is penalized using inhibitory weights. Then the motion-tuned computed errors of a distorted video relative to the reference video serves to predict temporal video integrity as follows:

\[ Q_{\text{Motion}}(x) = 1 - \sum_{n=1}^{N} \gamma_n (v_n^r - v_n^d)^2, \]

where \( v_n^r \) and \( v_n^d \) are motion tuned responses to the reference and distorted videos, respectively, and \( \gamma \) is the unit volume Gaussian window of unit standard deviation [5].

Although the weighting procedure used in [5] is based on a motion processing model of Area MT [18], the weights do not take into account the speed of object motion (i.e., slope of the motion plane) which can affect temporal visual masking. For example, whenever \( \delta_1 = \delta_2 \) in Figure 2, the excitatory-inhibitory weight is the same, implying the same amount of temporal distortion. However, the results of a series of human subjective studies [14-16] show that large, coherent object motions strongly suppress the visibility of flicker distortions on naturalistic videos, where flicker distortions in static regions (e.g., corresponding to Figure 2(a)) were much more noticeable than in moving regions (e.g., corresponding to Figure 2(b)).

Regarding the spectral analysis of flicker videos in the frequency domain, we observed that a flicker video produces locally separated spectral signatures that lie parallel to the motion tuned plane of the no-flicker video, but at a distance from the reference spectral plane depending on the flicker frequency. We also observed that larger flicker yields larger model V1 responses on the flicker induced spectral signatures [24]. Based on these observations, we captured motion silenced perceptual flicker visibility by measuring locally shifted response deviations relative to those on the reference video at each spatiotemporal subband. Define a flicker sensitive temporal video quality index as the sum of deviations

\[ Q_{\text{Flicker}}(x) = 1 - \frac{1}{M} \sum_{n=1}^{N} \frac{\left| C^r(\varphi, \theta, x) - C^d(\varphi, \theta, x) \right|}{K}, \]

By multiplicatively combining the motion-tuning and flicker-sensitive measurements, define the flicker sensitive pointwise temporal quality index

\[ Q_T(x) = Q_{\text{Motion}}(x) \times Q_{\text{Flicker}}(x). \]
IV. CONCLUSION AND FUTURE WORK

We presented a new VQA framework that models flicker sensitive temporal visual masking. The framework augments the MOVIE Index by combining perceptual flicker visibility with motion-tuned video integrity scores. The results show that the perceptually driven VQA framework quite correlates well with human judgments of video quality and is also highly competitive with recent VQA models tested on the LIVE VQA database. As future work, this model might be extended to improve a temporally dynamic VQA model that incorporates flicker density, which will require a database of time-varying flickering video data similar to [32].

REFERENCES