Signal Processing: Image Communication 67 (2018) 182–198

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Video quality assessment accounting for temporal visual masking of local flicker

Lark Kwon Choi *, Alan Conrad Bovik

Laboratory for Image and Video Engineering (LIVE), Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78701, USA

ARTICLE INFO

Keywords:
Video quality assessment
Temporal visual masking
Motion silencing
Flicker visibility
Human visual system

ABSTRACT

An important element of the design of video quality assessment (VQA) models that remains poorly understood is the effect of temporal visual masking on the visibility of temporal distortions. The visibility of temporal distortions like local flicker can be strongly reduced by motion. Based on a recently discovered visual change silencing illusion, we have developed a full reference VQA model that accounts for temporal visual masking of local flicker. The proposed model, called Flicker Sensitive-MOtion-based Video Integrity Evaluation (FS-MOVIE), augments the well-known MOVIE Index by combining motion tuned video integrity features with a new perceptual flicker visibility/masking index. FS-MOVIE captures the separated spectral signatures caused by local flicker distortions, by using a model of the responses of neurons in primary visual cortex to video flicker, an energy model of motion perception, and a divisive normalization stage. FS-MOVIE predicts the perceptual suppression of local flicker by the presence of motion and evaluates local flicker as it affects video quality. Experimental results show that FS-MOVIE significantly improves VQA performance against its predecessor and is highly competitive with top performing VQA algorithms when tested on the LIVE, IVP, EPFL, and VQEGHD5 VQA databases.

1. Introduction

Digital videos have become pervasive in our daily life. Video streaming services such as Netflix and YouTube, video sharing in social media, and video calling using Skype have become commonplace. As mobile devices have become "smarter", video consumption is exponentially increasing [1]. Given the dramatic growth in purveyed video content and heightened user expectations of higher-quality videos, it is desirable to develop more accurate and automatic VQA tools that can be used to optimize video systems, towards providing satisfactory levels of quality of experience (QoE) to the end user [2].

To achieve optimal video quality under limited bandwidth, storage, and power consumption conditions, video encoding technologies commonly employ lossy coding schemes, which can cause compression artifacts that degrade perceptual quality [3]. Videos can also be degraded by transmission distortions (e.g., packet loss, playback interruption, and freezing) due to channel throughput fluctuations [4]. Hence, videos suffer not only from spatial distortions such as blocking, blurring, ringing, mosaic patterns, and noise, but also from temporal distortions such as motion compensation mismatches, flicker, mosquito effects, ghosting, jerkiness, smearing, and so forth [3].

Specifically, local flicker denotes the temporal fluctuation of spatially local luminance or chrominance in videos. Local flicker occurs mainly due to coarse quantization, mismatching of inter-frame blocks, improper deinterlacing, and dynamic rate changes in adaptive rate control [5]. Local flicker distortions, which are not well explained by current VQA models, frequently appear near moving edges and textures in compressed videos as well as in interlaced videos, producing annoying visual artifacts such as line crawling, interline flicker, and edge flicker [6,7].

Since humans are the ultimate arbiters of received videos, understanding how humans perceive visual distortions and modeling the visibility of distortions in digital videos have been important topics for developing successful quality assessment models [8]. Early human visual system (HVS) based VQA models include Mannos and Sakrison's metric [9], the Visual Differences Predictor (VDP) [10], Sarnoff Just Noticeable Differences (JND) Vision Model [11], Moving Pictures Quality Metric (MPQM) [12], the Perceptual Distortion Metric (PDM) [13], and the Digital Video Quality (DVQ) model [14]. Later models include Structural Similarity (SSIM) [15], Multiscale- SSIM (MS-SSIM) [16], motion-based SSIM [17], Visual Information Fidelity (VIF) [18], Visual Signal-to-Noise Ratio (VSNR) [19], Video Quality Metric (VQM) [20] and the Scalable Wavelet Based Video Distortion Index [21]. More recently, Ninassi et al. [22], TetraVQM [23], MOVIE [24], SpatioTemporal-Most Apparent Distortion (ST-MAD) [25], SpatioTemporal Reduced Reference Entropic

* Corresponding author.
E-mail addresses: larkkwonchoi@gmail.com (L.K. Choi), bovik@ece.utexas.edu (A.C. Bovik).

https://doi.org/10.1016/j.image.2018.06.009
Received 10 July 2017; Received in revised form 4 February 2018; Accepted 20 June 2018
Available online 23 June 2018

Differences (STRRED) [26], Video-BLind Image Integrity Notator using DCT-Statistics (V-BLIINDS) [27], and VQM-Variable Frame Delays (VQM-VFD) [28] are examples that include more sophisticated temporal aspects. In video streaming services, other factors impact the overall QoE such as initial loading delays, freezing, stalling, skipping, and video bitrate, all of which have been widely studied [29–31].

One potentially important aspect of the design of VQA models that remains poorly understood is the effect of temporal visual masking on the visibility of temporal distortions. The mere presence of spatial, temporal, or spatiotemporal distortions does not imply a corresponding degree of perceptual quality degradation, since the visibility of distortions can be strongly reduced or completely removed by visual masking [32]. Spatial visual masking, such as luminance masking and contrast/texture masking [33], is quite well-modeled in modern perceptual image quality assessment tools [10,15]. However, there remains significant scope to expand and improve computational models of temporal visual masking. While numerous related temporal aspects of visual perception have been studied, including change blindness [34], crowding [35], and global aspects of temporal masking [36–39], much less work has been done on masking of non-global temporal phenomenon, such as spatially localized flicker [40–42]. Masking that occurs near scene changes has been observed and used in algorithms [43], and some experimental visual masking devices have been applied in video compression [44–48] and JND modeling [11].

Recently, Suchow and Alvarez [39] demonstrated a striking "motion silencing" illusion, in the form of a powerful temporal visual masking phenomenon called change silencing, wherein the salient temporal changes of objects in luminance, color, size, and shape appear to cease in the presence of large, coherent object motions. This motion-induced failure to detect change not only shows a tight coupling between motion and object appearance, but also reveals that commonly occurring temporal distortions, such as local flicker, can be dramatically suppressed by the presence of motion.

Motivated by the visual change silencing phenomenon [40], we have investigated the nature of spatially localized flicker in natural digital videos, and the potential modeling of temporal visual masking of local flicker to improve VQA performance. We exploit a psychophysical model of temporal flicker masking on digital videos to create an improved VQA model. Specifically, we use the temporal flicker masking model to augment the well-known MOVIE Index. Using the results of a series of human subjective studies that we previously executed, we have developed a quantitative model of local flicker perception relating to motion silencing to more accurately predict video quality when there is flicker. We also have analyzed the influence of flicker on VQA in terms of compression bitrate, object motion, and temporal subsampling. This is an important step towards improving the performance of VQA models, by accounting for the effects of temporal visual masking on flicker distortions in a perceptually agreeable manner, and by further developing MOVIE in the temporal dimension related to flicker.

The proposed model, called Flicker Sensitive-Motion-based Video Integrity Evaluation (FS-MOVIE), computes bandpass filter responses on reference and distorted videos using a spatiotemporal Gabor filter bank [51,52], then deploys a model of the responses of V1 neurons through an energy model of motion perception [53] and a divisive normalization stage [54]. FS-MOVIE modifies MOVIE by responding to spectral separations caused by local flicker. This internal flicker visibility index is combined with motion-tuned measurements of video integrity, temporally pooled to produce a final video quality score. Our evaluation of the specific performance enhancements of FS-MOVIE, along with the overall comprehensive results, show that the video quality predictions produced by FS-MOVIE correlate quite highly with human subjective judgments of quality on distorted videos. Its performance is highly competitive with, and indeed exceeds, that of the most recent VQA algorithms tested on the LIVE [55], IVP [56], EPFL [57], and VQEGHD5 VQA databases [58]. The significant improvement of VQA performance attained by FS-MOVIE implies that temporal visual masking of local flicker is important.

The remainder of this paper is organized as follows. Section 2 explains the background concepts that motivate FS-MOVIE. The FS-MOVIE model is detailed in Section 3. We evaluate the performance of FS-MOVIE in Section 4. Section 5 concludes the paper with discussions of possible future work.

2. Background

2.1. Visual Masking

Visual masking is the decrease or elimination of the visibility of a stimulus, called the target, by the presence of another stimulus, called the mask, which is close to the target in space and/or time [32]. Visual masking typically occurs when the target and the mask have a similar orientation, spatiotemporal frequency, motion, color, or other attribute [8]. For example, local high-frequency energy in an image reduces the visibility of other high-frequency features such as noise, reducing the perceptual significance of the distortions. JPEG compression distortions and noise are highly visible on smooth luminance regions like faces or blue skies, whereas they can be nearly imperceptible on highly textured areas such as hair, grass, or flowers [59]. This is called contrast masking [33].

Spatial visual masking models of contrast/texture masking have been used to predict the perception of structural image degradations [15] and visible image differences [10]. Divisive normalization of the neuronal responses has also been shown to significantly reduce statistical dependencies between the responses and to affect distortion visibility, as well as providing an explanation of the contrast masking effect [54,60].

To understand temporal visual masking a large number of psychophysical experiments have been conducted using light flashes [36], sine wave-gratings [37], vernier stimuli [39], change blindness [34], crowding [35], and change silencing dots [40–42]. In video processing research, it has been found that human observers have difficulty perceiving a temporary reduction of spatial details in TV signals immediately before and after scene changes [43]. Netravali et al. [44] investigated the perception of quantization noise during luminance transitions. Haskell et al. [45] studied observers' tolerance to distortions of moving images. Puri et al. [46] designed an adaptive video encoder using the visibility of noise on flat areas, textures, and edges. Girod [47] highlighted the theoretical significance of temporal masking. Johnston et al. [48] built a non-linear video quantizer using temporal masking. A variety of ideas have been proposed to account for temporal masking in video compression algorithms [44–48]. For example, [49,50] used global frame-difference JND calculations to account for time-localized temporal masking caused by scene changes, while in a similar vein, [11] used global temporal flicker to mask spatial distortion.

2.2. Motion silencing of flicker distortions

A variety of methods for predicting flicker visibility have been proposed, including the sum of squared differences [5], [61]. However, these methods are largely content-dependent and require various thresholds between frames, are limited to block-wise accuracy, and do not take into account temporal visual masking of local flicker distortions.

To better understand temporal visual masking of local flicker distortions, we executed a series of human subjective studies on naturalistic videos about video quality, object motion, flicker frequency, and eccentricity [62], [63]. The results show that local flicker visibility is strongly reduced by the presence of large, coherent object motions, in agreement with human psychophysical experiments on synthetic stimuli [41], [42]. The impact of object motions on the visibility of flicker distortions is significant when the quality of a test video is poor. We have developed preliminary local flicker detector models [64], [65], which are used in our proposed flicker VQA model.

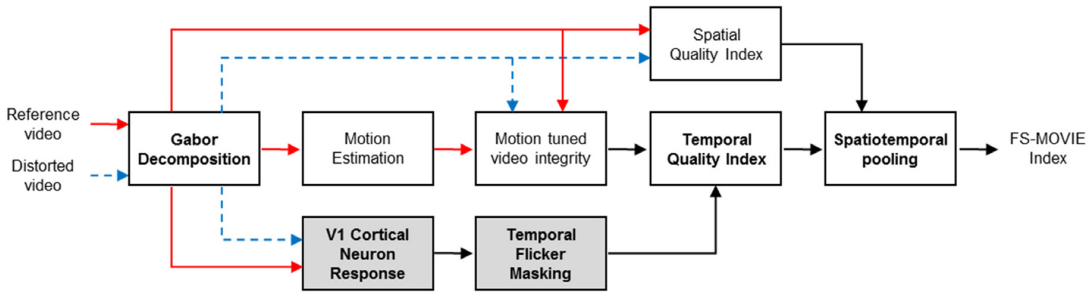


Fig. 1. Overall workflow of the proposed flicker sensitive motion tuned VQA model, FS-MOVIE.

2.3. Motion perception

Motion perception is the process of inferring the speed and direction of moving objects. Motion perception occurs from the retina through the lateral geniculate nucleus (LGN) and the primary visual cortex (V1) to the middle temporal (MT) visual area and beyond [66]. Visual signals are commonly modeled as spatially bandpass filtered by ganglion cells in the retina and temporally filtered at the LGN to reduce temporal entropy. In V1, a multiscale, multi-orientation decomposition of the visual data occurs over space and time. Further along, extra cortical area MT creates locally oriented, motion tuned spatiotemporal responses [67], [68].

Watson and Ahumada [69] proposed a model of velocity perception, where local motion was modeled as local translational motion. In the frequency domain, complex motions in video segments without scene changes can be analyzed using the spatiotemporally localized spectra of image patches assumed to be undergoing translation. Specifically, suppose $a(x, y, t)$ is an arbitrary space–time video patch at spatial coordinate (x, y) and time t. Let λ and ϕ denote the horizontal and vertical velocity components of a translating image patch. When an image patch translates at a constant velocity $[\lambda, \phi]$, the moving video sequence becomes $a(x - \lambda t, y - \phi t, t)$. The spectrum of a stationary image patch lies on the u, v plane, while the Fourier transform of a translating image patch shears into an oblique plane through the origin. Such a plane can be expressed

$$\lambda u + \phi v + w = 0, \tag{1}$$

where u, v, and w are spatial and temporal frequency variables corresponding to (x, y) and t, respectively [69]. The orientation of this plane indicates the speed and direction of motion.

2.4. From MOVIE to FS-MOVIE

The MOVIE Index incorporates motion perception models of cortical Area V1 and Area MT to predict video quality [24]. The Spatial MOVIE Index mainly captures spatial distortions using spatiotemporal Gabor filter responses. The Temporal MOVIE Index measures motion quality relevant to temporal distortions, whereby errors are computed between the motion tuned responses of reference and distorted video sequences to evaluate temporal video integrity [24].

The Temporal MOVIE framework computes motion tuned responses using excitatory and inhibitory weights on directional motion responses. However, the weights are defined by the *distance* of the local spectral plane from the Gabor filters, without considering the *speed* of object motions. When the distances are the same, the Temporal MOVIE Index predicts the same amount of temporal distortion even when the perceived distortions may be different due to motion.

Algorithm 1 Pseudocode for FS-MOVIE

Input: Reference video $r(\mathbf{x})$ and distorted video $d(\mathbf{x})$, where $\mathbf{x} = (x, y, t)$.
Output: FS-MOVIE Index.

```
Compute spatial and temporal quality indices Q_S(x) and Q_T(x) for frames:
for t = the 17th frame : the last 17th frame from the end
    Compute Gabor filter responses f(x,k) and g(x,k) by Eq. (2).
    Compute the Q_S(x) using Err_S(x,k) and Err_DC(x) by Eq. (5).
    Compute motion tuned video integrity using Err_Motion(x) by Eq. (12).
    Compute V1 cortical neuron responses C(φ,θ) by Eq. (15).
    Compute a temporal flicker masking index FS(x) by Eq. (21).
    Compute the Q_T(x) using Err_Motion(x) and FS(x) by Eq. (22).
end for
Apply spatiotemporal pooling for Q_S(x) and Q_T(x) by Eqs. (25) - (29):
    Compute the Spatial FS-MOVIE Index by Eq. (6).
    Compute the Temporal FS-MOVIE Index by Eq. (23).
    Compute the FS-MOVIE Index by Eq. (24).
```

From the results of a series of human subjective studies [62], [63], we have found that large object motions strongly suppress the visibility of local flicker. Since temporal masking of local flicker directly relates to cortical neuronal responses in cortical areas V1 and MT [40–42], and since MOVIE already models neuronal processes in these regions of the brain, we viewed the MOVIE Index as an ideal candidate for enhancement. By incorporating the new temporal flicker visibility/masking index into the MOVIE model, we hypothesized that it would be possible to more accurately predict perceptual video quality. Therefore, FS-MOVIE modifies MOVIE in ways that align with both sensible video engineering and vision science.

3. Flicker sensitive motion tuned VQA model

We now detail FS-MOVIE. Reference and test videos are decomposed into spatiotemporal bandpass channels using a 3D Gabor filter bank. The outputs of the Gabor filter bank are used as in Spatial MOVIE to obtain a spatial quality index. The temporal flicker masking index is embedded into Temporal MOVIE, yielding a flicker-sensitive temporal quality index. The complete FS-MOVIE Index is then obtained using a spatiotemporal pooling strategy. Overall workflow is shown in Fig. 1, and the schematics are summarized in Algorithm 1. In Fig. 1, steps in MOVIE processing that are modified in FS-MOVIE are in bold face, while grayed boxes indicate new processing steps in FS-MOVIE not present in MOVIE.

3.1. Gabor decomposition

Natural images and videos are inherently multiscale and multi-orientation, and objects move multi-directionally at diverse speeds. To efficiently encode natural visual signals, the vision system decomposes the visual world over multi-scales, orientations, directions, and speeds. Since the receptive field profiles of V1 simple cells are well-modeled as linear bandpass Gabor filters [51], [52], Gabor decompositions are widely used to model simple cell responses to video signals [24], [70]. Likewise, we use a 3D spatiotemporal separable Gabor filter

$$h(\mathbf{x}) = \frac{1}{(2\pi)^{3/2}|\sum|^{1/2}} \exp\left(-\frac{\mathbf{x}^T \sum^{-1} \mathbf{x}}{2}\right) \exp\left(j\mathbf{U}_0\mathbf{x}\right), \tag{2}$$

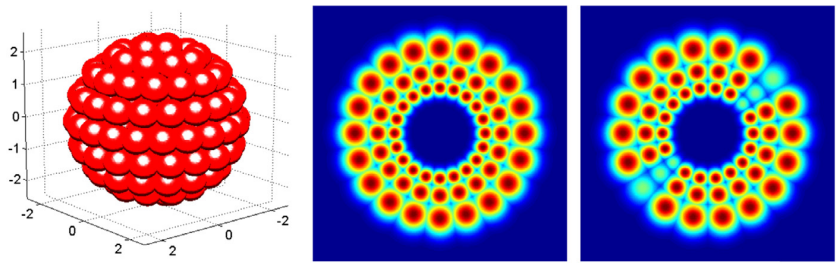


Fig. 2. Gabor filter bank in the frequency domain. (a) Geometry of the Gabor filter bank. (b) A slice of the Gabor filter bank along the plane of zero temporal frequency. (c) A slice of the Gabor filter bank along the plane of zero vertical spatial frequency.

where $\mathbf{x} = (x, y, t)$ are space–time video coordinates, $\mathbf{U_0} = (U_0, V_0, W_0)$ is the center frequency, and $\sum$ is the covariance matrix.

We implemented a Gabor filter bank similar to [24], but we used a wider range of filters to accommodate temporal masking of local flicker. In the Gabor filter bank, three scales ($P = 3$) of Gabor filters were deployed, with 57 filters at each scale on the surface of a sphere centered at the space–time frequency origin as shown in Fig. 2. The filters were implemented with octave bandwidths of 0.45 octaves, measured at one standard deviation of the Gabor frequency response. In the implementation, the largest radial center frequency (i.e., the finest scale of filters) was 0.7π radians per sample, and the filters were sampled out to a width of three standard deviations. For a frequency bandwidth bw (in octaves) and central frequency cf_0 about which the filters should be tuned, the standard deviation of the Gaussian envelope in frequency space was determined as $cf_0 \times (2^{bw} - 1)/(2^{bw} + 1)$ [70]. Following [70], the smallest radial center frequency (i.e., the coarsest scale of filters) was 0.375π ($=0.7\pi/(2^{0.45})^2$) radians per sample, with a standard deviation of 5.49 ($=1/[0.375\pi \times (2^{0.45} - 1)/(2^{0.45} + 1)]$) pixels (frames), supporting 33 pixels along the spatial axis and 33 frames along the temporal axis.

A total of 171 ($=57 \times 3$) filters comprised the filter bank: 10, 18, 15, 10, and 4 filters were tuned to five different speeds tan (φ) over corresponding vertical angles $\varphi = 0$, 20, 40, 60 and 80 degrees and orientations $\theta = 18$, 20, 24, 36, and 90 degrees, respectively. The orientations of Gabor filters were chosen such that adjacent filters intersected at one standard deviation following [70]. For example, 10 filters were tuned to a temporal frequency of 0 radians per sample corresponding to no motion, where filters were chosen to be multiples of 18° in the range [0, 180°). Fig. 2(b) and (c) show the slices of the Gabor filter bank used in FS-MOVIE along the plane of zero temporal frequency and along the plane of zero vertical spatial frequency. We also included a Gaussian filter centered at the frequency origin to capture low frequencies, as in [24]. The standard deviation of the Gaussian filter was chosen so that it intersected the coarsest scale of bandpass filters at one standard deviation, supporting 7 pixels and frames along the spatial and temporal axes.

3.2. Spatial FS-MOVIE index

Let $r(\mathbf{x})$ and $d(\mathbf{x})$ denote the reference and distorted videos respectively. Then $r(\mathbf{x})$ and $d(\mathbf{x})$ are passed through the Gabor filter bank described in Section 3.1 to obtain bandpass filtered videos. Let $\mathbf{f}(\mathbf{x_0}, k)$ and $\mathbf{g}(\mathbf{x_0}, k)$ be the magnitudes of the complex Gabor channel responses of a Gabor filter at $k = 1, 2, \ldots, K$, where $K = 171$, contained within a 7×7 window B centered at arbitrary coordinate $\mathbf{x_0}$ of the reference and distorted videos, respectively. Then $\mathbf{f}(\mathbf{x_0}, k)$ is a vector of dimension N ($=49$), where $\mathbf{f}(\mathbf{x_0}, k) = [f_1(\mathbf{x_0}, k), f_2(\mathbf{x_0}, k), \ldots, f_N(\mathbf{x_0}, k)]$. Similar definitions apply for $\mathbf{g}(\mathbf{x_0}, k)$. We used the 3D spatiotemporal Gabor filter bank, but only applied the 2D 7×7 window B in the spatial plane when computing the Spatial FS-MOVIE Index [24].

We do not alter the method of spatial quality prediction used in MOVIE [24]. However, Spatial FS-MOVIE computes spatiotemporal bandpass responses over a wider range of frequency subbands than does Spatial MOVIE, as described in Section 3.1. Spatial FS-MOVIE measures weighted spatial errors from each sub-band Gabor response and from the DC sub-band Gaussian filter output, respectively, as follows:

$$Err_S(\mathbf{x}_0, k) = \frac{1}{2}\sum_{n=1}^{N}\gamma_n\left[\frac{f_n(\mathbf{x}_0, k) - g_n(\mathbf{x}_0, k)}{\max\left(\sqrt{\sum_{n=1}^{N}\gamma_n|f_n(\mathbf{x}_0, k)|^2}, \sqrt{\sum_{n=1}^{N}\gamma_n|g_n(\mathbf{x}_0, k)|^2}\right) + A_1}\right]^2, \tag{3}$$

$$Err_{\mathrm{DC}}(\mathbf{x}_0) = \frac{1}{2}\sum_{n=1}^{N}\gamma_n \times\left[\frac{f_n(\mathrm{DC}) - g_n(\mathrm{DC})}{\max\left(\sqrt{\sum_{n=1}^{N}\gamma_n|f_n(\mathrm{DC}) - \mu_{\mathbf{f}}|^2}, \sqrt{\sum_{n=1}^{N}\gamma_n|g_n(\mathrm{DC}) - \mu_{\mathbf{g}}|^2}\right) + A_2}\right]^2, \tag{4}$$

where $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \ldots, \gamma_N\}$ is a unit-volume ($\sum_{n=1}^{N}\gamma_n = 1$) Gaussian window of unit standard deviation sampled out to a width of three standard deviations. In our implementation, $N = 49$. The Gaussian window is used to avoid blocking or ringing artifacts caused by using a square window. We fixed $A_1 = 0.1$ and $A_2 = 1$ as in [24] to prevent numerical instability when the denominator is small. Then $\mathbf{f}(\mathrm{DC})$ and $\mathbf{g}(\mathrm{DC})$ are vectors of dimension N extracted from the Gaussian filtered versions of $r(\mathbf{x})$ and $d(\mathbf{x})$, while $\mu_{\mathbf{f}}$ and $\mu_{\mathbf{g}}$ are the averages of $\mathbf{f}(\mathrm{DC})$ and $\mathbf{g}(\mathrm{DC})$ over N samples, respectively.

Next, the error indices $Err_S(\mathbf{x}, k)$ and $Err_{\mathrm{DC}}(\mathbf{x})$ are converted into a spatial quality index [24]:

$$Q_S(\mathbf{x}_0) = 1 - \frac{\frac{P}{K}\sum_{k=1}^{K}Err_S(\mathbf{x}_0, k) + Err_{\mathrm{DC}}(\mathbf{x}_0)}{P + 1}, \tag{5}$$

where K ($=171$) is the total number of Gabor filters, and P ($=3$) is the total number of scales.

Finally, define the Spatial FS-MOVIE Index using a spatio-temporal pooling strategy. The coefficient of variation (CoV) of Q_S values in (5) is obtained as a single score on each frame. Then apply a temporal pooling strategy to achieve a single score for each video as follows:

$$\text{Spatial FS-MOVIE} = TP\left(\frac{\sigma_{Q_S(x,y,t)}}{\mu_{Q_S(x,y,t)}}\right), \tag{6}$$

where σ_{Q_S} and μ_{Q_S} are the standard deviation and the mean of Q_S, respectively. TP is a temporal pooling function on the frame CoV values. The details of TP are described in Section 3.8. The CoV is motivated by

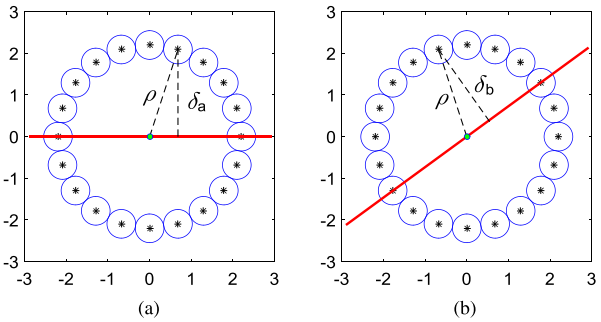


Fig. 3. The motion tuned spectral planes relative to a slice through the Gabor filter bank at one scale: (a) at a static region and (b) at a moving region. The horizontal and vertical axes are spatial and temporal frequency, respectively. The red solid line indicates a spectral plane, while blue small circles represent Gabor filters. The centers of each Gabor filter are marked. ρ is the radius of the sphere along which the center frequency of the Gabor filters lies. δ_a and δ_b are the distances of the center frequency of one Gabor filter from the spectral plane at static and moving regions, respectively.

the fact that larger values of σ_{Qs} indicate a broader spread of both high and low quality regions, yielding lower overall perceptual quality [24].

3.3. Motion tuned video integrity

When temporal distortions are present, Temporal MOVIE penalizes the shifted spectrum of the distorted video lying along a different orientation than the reference video, by computing a weighted sum of the Gabor filter outputs. The weight assigned to each individual Gabor filter is determined by its distance from the spectral plane of the reference video. The motion- tuned error of a distorted video relative to the reference video serves to evaluate temporal video integrity [24].

Let λ_n and ϕ_n indicate the horizontal and vertical velocity components of an image patch on the reference video, where $n = 1, 2, \ldots, N$ elements of the flow field spanned by a 7×7 local window B centered on $\mathbf{x_0}$. λ_n and ϕ_n are obtained using [70]. Define a sequence of distance vectors $\delta(k)$, $k = 1, 2, \ldots, K$ of dimension N (=49). Each element of this vector indicates the perpendicular distance of the center frequency of the kth Gabor filter from the plane containing the spectrum of the reference video in a window centered on $\mathbf{x_0}$ extracted using B, as shown in Fig. 3. 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

$$\delta_n(k) = \left| \frac{\lambda_n u_0(k) + \phi_n v_0(k) + w_0(k)}{\sqrt{\lambda_n^2 + \phi_n^2 + 1}} \right|, n = 1, 2, \ldots, N. \tag{7}$$

A set of excitatory and inhibitory weights are derived as a function of the distance $\delta(k)$ in (7). First, assign a maximum weight to the filters that intersect the spectral plane and a minimum weight to the filter lying at the greatest distance, using the weighting function [24]

$$\alpha_n'(k) = \frac{\rho(k) - \delta_n(k)}{\rho(k)}, \tag{8}$$

where $\rho(k)$ is the radius of the sphere along which the center frequency of the filters lies. Excitatory and inhibitory weights [24] are obtained by shifting the weights in (8) to be zero mean, and by normalizing them so that the maximum weight is unity:

$$\alpha_n(k) = \frac{\alpha_n'(k) - \mu_\alpha}{\max_{k=1,2,\ldots,K/P}[\alpha_n'(k) - \mu_\alpha]}, \tag{9}$$

where μ_α is the average value of $\alpha_n'(k)$ at each scale. Similar definitions apply for other scales.

Motion tuned responses from the reference and distorted videos sequences are computed, respectively [24], as

$$v_n^r(\mathbf{x}_0) = \frac{[f_n(DC) - \mu_f]^2 + \sum_{k=1}^K \alpha_n(k) f_n(\mathbf{x}_0, k)^2}{[f_n(DC) - \mu_f]^2 + \sum_{k=1}^K f_n(\mathbf{x}_0, k)^2 + A_3}, \tag{10}$$

$$v_n^d(\mathbf{x}_0) = \frac{[g_n(DC) - \mu_g]^2 + \sum_{k=1}^K \alpha_n(k) g_n(\mathbf{x}_0, k)^2}{[g_n(DC) - \mu_g]^2 + \sum_{k=1}^K g_n(\mathbf{x}_0, k)^2 + A_3}, \tag{11}$$

where $A_3 = 100$ to stabilize (10) and (11), as described in [24].

Define a motion error index [24] at $\mathbf{x}_0$ to capture deviations between the local motions of the reference and distorted videos:

$$Err_{\text{Motion}}(\mathbf{x}_0) = \sum_{n=1}^N \gamma_n [v_n^r(\mathbf{x}_0) - v_n^d(\mathbf{x}_0)]^2, \tag{12}$$

where $\boldsymbol{\gamma}$ is the same unit volume Gaussian window of unit standard deviation sampled out to a size of 7×7. The metric (12) takes value 0 when the reference and test videos are identical.

3.4. Modeling V1 cortical neuron responses

The responses of Area V1 neurons were modeled using the spatiotemporal energy model [53] with divisive normalization [54]. The motion energy within a spatiotemporal frequency band was extracted by squaring the responses of quadrature Gabor filter components and by summing them as follows:

$$E_i(\varphi, \theta) = \left[h_{sin,i}(\varphi, \theta) * I\right]^2 + \left[h_{cos,i}(\varphi, \theta) * I\right]^2, \tag{13}$$

where $h_{\text{sin},i}(\varphi, \theta)$ and $h_{\text{cos},i}(\varphi, \theta)$ are sine and cosine Gabor filters at φ, θ, and a scale i, respectively. I is the luminance level of a video, while the symbol * means convolution.

The quantity (13) models the response of an individual neuron to a specific band of spatiotemporal frequencies. To agglomerate the combined responses of all cortical neighborhoods that include cells tuned over the full range of orientations and directions, the response of each neuron was normalized to limit its dynamic range of responses without altering the relative responses of neurons in the pool [54]. The response of a modeled simple cell $S_i(\varphi, \theta)$ is computed by dividing each individual energy response by the sum of its all neighboring (i.e., all φ and θ) energy responses for each scale i:

$$S_i(\varphi, \theta) = R \frac{E_i(\varphi, \theta)}{\sum_{\varphi, \theta} E_i(\varphi, \theta) + \sigma^2}, \tag{14}$$

where R determines the maximum attainable response, and σ is a semi-saturation constant. Here $R = 4$ and $\sigma = 0.2$ was used in agreement with recorded physiological data [54]. Note that summing the energy outputs in (14) over all φ and θ yields the total Fourier energy of the stimulus [54], where the normalization could also be computed locally by summing over a limited region of space and a limited range of frequencies [54]. The model V1 complex cell responses $C(\varphi, \theta)$ are obtained by averaging the responses (14) along scales on constant space–time frequency orientations:

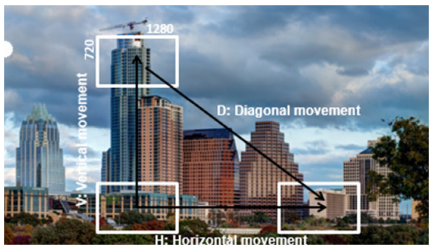
$$C(\varphi, \theta) = \sum_{i=1}^3 c_i S_i(\varphi, \theta), \tag{15}$$

where $c_i = 1/3$ are weighting factors. We used constant values although they could be Gaussian weighted by distance [54].

3.5. Temporal flicker masking

3.5.1. The nature of flicker

Although the excitatory–inhibitory weighting procedure used in the measurements of motion tuned temporal distortions in MOVIE is based on a model of Area MT [68], the weights are defined only in terms of the *distance* from the motion tuning plane without considering the *speed* of object motion (i.e., slope of the motion plane). In MOVIE, whenever $\delta_a =$



(a)

Unit: pixels per frame

	Horizontal movement				Vertical movement				Diagonal movement			
x	0	1	2	3	0	0	0	0	0	1	2	3
y	0	0	0	0	0	1	2	3	0	1	2	3

(b)

Fig. 4. Simulated translational motion video. A total of 12 extracted videos were used to study local spectral motion signatures. (a) Exemplar regions translated in the horizontal, vertical, and diagonal directions. (b) Translation speeds: x (horizontal), y (vertical) pixels per frame.

δ_b in Fig. 3, the excitatory–inhibitory weight is the same, predicting the same amount of temporal distortions. However, humans may perceive temporal distortions differently since large object motions strongly suppress the visibility of flicker distortions, where flicker distortions on static regions (e.g., Fig. 3a) are much more noticeable than on moving regions (e.g., Fig. 3b).

Here, we study the nature of flicker and propose a new model of temporal flicker masking. To demonstrate the distribution of the spectral signatures of flicker distorted videos, we simulated translational motion on a very large 15619 × 2330 static image by shifting a frame-size window (1280 × 720 pixels) at constant speeds (e.g., 0, 1, 2, and 3 pixels per frame) in the horizontal, vertical, and diagonal directions, as illustrated in Fig. 4. Then, we induced quantization flicker by compressing the videos using a H.264 codec, alternating every 3 frames using different Quantization Parameter (QP) pairs (e.g., between QP26 and QP44; between QP26 and QP38; between QP26 and QP32) as used in [64]. We estimated V1 responses for each condition listed in Fig. 4b without flicker and with flicker separately. We applied this process to natural videos, as shown in Fig. 4, since our goal is to understand and model temporal masking of local flicker on natural videos. We accomplish this in simulations by controlling the velocity of motion, and the levels of flicker distortions caused by video compression in the form of quantization flicker.

We observed that a flicker video produces bifurcated local spectral signatures that lie parallel to the motion tuned plane of the no-flicker video but at a distance from the reference spectral plane determined by the flicker frequency, as illustrated in Fig. 5. This phenomenon might be explained as follows: consider a video sequence modeled as a frame translating at a constant velocity $[\lambda, \phi]$ and flickering. It may be expressed

$$o(x,y,t) = a(x-\lambda t, y-\phi t, t) \times \frac{[1+b(x,y,t)]}{2}, \tag{16}$$

where $a(x, y, t)$ is an arbitrary space–time image patch, while $b(x, y, t)$ is a bounded periodic function of period $2L$ (e.g., $-1 \le b(x, y, t) \le 1$). Then, assuming that $b(x, y, t)$ is sufficiently regular, it may be represented by the Fourier series

$$b(x,y,t) = \frac{\eta_0}{2} + \sum_{n=1}^{\infty} \kappa_n \cos\left[\frac{n\pi(x,y,t)}{L} + \psi_n\right], \tag{17}$$

where $\kappa_n = (\eta_n^2 + \zeta_n^2)^{0.5}$, $\eta_n = \kappa_n \cos\psi_n$, $\zeta_n = -\kappa_n \cos\psi_n$, and $\psi_n = \tan^{-1}(-\zeta_n/\eta_n)$. Although actual flicker may not be truly periodic, our approach assumes a local "near-periodicity" in space and time due to the nature of the processes that cause flicker, such as video compression. Denote $\cos[n\pi(x, y, t)/L + \psi_n]$ as $\cos(\Omega_n t)$ for simplicity. The 3D space–time Fourier transform of this translating and flickering video can then be written as:

$$O(u,v,w) = (1+\frac{\eta_0}{2})\pi A(u,v)\delta(\Gamma) + \frac{\kappa_n}{2}\pi A(u,v)\sum_{n=1}^{\infty}\left[\delta(\Gamma+\frac{\Omega_n}{2\pi}) + \delta(\Gamma-\frac{\Omega_n}{2\pi})\right], \tag{18}$$

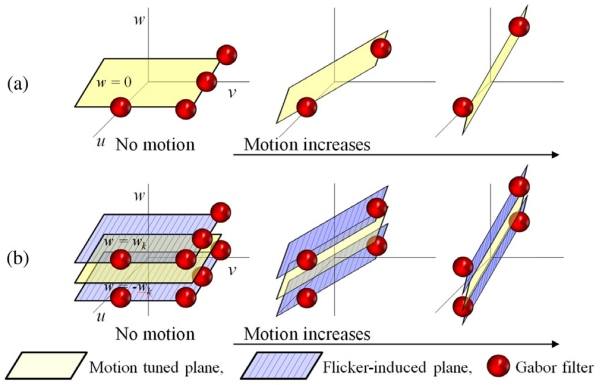


Fig. 5. Spectral signatures that constitute motion tuned planes: (a) Reference video. (b) Flicker video.

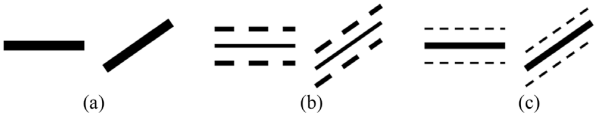


Fig. 6. Schematic illustration of spectral signatures constituting motion tuned planes: (a) No flicker video, (b) flicker video with large flicker magnitude, and (c) flicker video with small flicker magnitude. The solid line is a reference motion tuned plane, while the dashed line is a flicker-induced plane. The thickness of the lines shows the magnitude of the V1 responses. From left to right, videos are static and moving, respectively.

where $O(u, v, w)$ and $A(u, v)$ denote the Fourier Transforms of $o(x, y, t)$ and $a(x, y, t)$, $\Gamma = \lambda u + \phi v + w$, and $O(u, v, w)$ consists of multiple replicas (harmonics) of $A(u, v)$ oriented and mapped onto a 2D plane in the 3D Fourier domain. The multiple planes are defined by the following equations:

$$\Gamma = 0, \Gamma + \frac{\Omega_n}{2\pi} = 0, \text{ and } \Gamma - \frac{\Omega_n}{2\pi} = 0. \tag{19}$$

The first term shears into an oblique plane through the origin in the frequency domain, while the second and the third terms show two planes shifted in the negative and positive temporal frequency directions, by amount $\Omega_n/2\pi$, as shown in Fig. 5.

We also observed that larger flicker magnitudes (caused by larger QP level alternations, e.g., between QP26 and QP44, rather than between Q26 and QP32) produced larger model V1 responses on the induced spectral signatures, as illustrated in Fig. 6. We observed similar results when we executed the same spectral analysis on naturalistic videos obtained from the LIVE VQA database by inducing quantization flicker distortions.

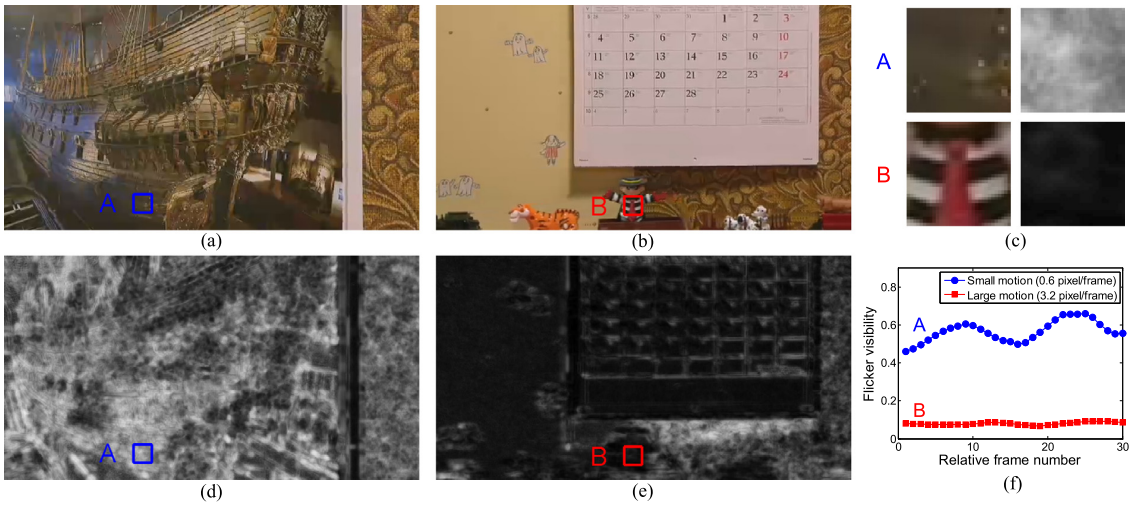


Fig. 7. Illustration of the perceptual flicker visibility index. (a) Frame 48 from the H.264 compressed video containing small motions. (b) Frame 464 from the H.264 compressed video containing large motions. (c) Segments A and B of (a) and (b) as well as (d) and (e). (d) The flicker visibility map of (a). (e) The flicker visibility map of (b). Note: Brighter regions indicate larger predicted flicker visibility. (f) Average flicker visibility at segments A and B along 30 frames. Test videos and corresponding perceptual flicker visibility map videos are available at http://live.ece.utexas.edu/research/flicker/flicker_visibility.html.

3.5.2. Perceptual flicker visibility index

Our new perceptual flicker visibility index is based on the way flicker changes the local spectral signatures of videos and how motion influences the resulting V1 neuron responses. Shifted or separated spectral signatures not present in a reference video might be associated with flicker distortions. Therefore, we devised an approach to capture temporally masked, perceptual flicker visibility by measuring locally shifted energy deviations relative to those on the reference video at each spatiotemporal frequency. This approach has advantages over other flicker prediction methods [5], [61]. The proposed method provides a pixel-wise accurate flicker visibility index map on both static and moving regions without content-dependent thresholds [65].

Let $C^r(\varphi, \theta, \mathbf{x})$ and $C^d(\varphi, \theta, \mathbf{x})$ model the V1 neuron responses to the reference and to the distorted videos in (15), respectively. Define a temporally masked perceptual flicker visibility index

$$FV(\mathbf{x}) = \sum_{\varphi,\theta} \left| C^r(\varphi,\theta,\mathbf{x}) - C^d(\varphi,\theta,\mathbf{x}) \right|. \tag{20}$$

To restrict the range of FV to [0, 1], average FV along φ and θ after normalizing FV by R, then define the flicker sensitive index

$$FS(\mathbf{x}) = \frac{P}{K} \sum_{\varphi,\theta} \frac{\left| C^r(\varphi,\theta,\mathbf{x}) - C^d(\varphi,\theta,\mathbf{x}) \right|}{R}, \tag{21}$$

where K is the total number of Gabor filters, and P is the number of scales. $R = 4$ as used in (14). The FS value in (21) is 0 when the reference and test videos are identical. Fig. 7 shows the predicted flicker visibility index on local scenes at small motions (Fig. 7a) and at large motions (Fig. 7b) in the H.264 compressed video "Mobile and Calendar" from the LIVE VQA database. Brighter regions denote larger flicker visibility. The predicted flicker visibility index represents the suppression of flicker distortions well in the presence of large object motions.

3.6. Temporal FS-MOVIE index

The Temporal FS-MOVIE Index predicts temporal video quality by combining Temporal MOVIE with a new temporal visual masking model of flicker visibility over a wider range of possible speeds.

We first define a pointwise flicker sensitive temporal quality index from $Err_{\text{Motion}}(\mathbf{x})$ in (12) and $FS(\mathbf{x})$ in (21) as follows,

$$Q_T(\mathbf{x}) = 1 - \left[Err_{\text{Motion}}(\mathbf{x}) \times FS(\mathbf{x}) \right]. \tag{22}$$

Next, define the Temporal FS-MOVIE Index as the square root of the coefficient of variation (CoV) of Q_T to obtain a single score for each frame, then apply a temporal pooling strategy on the frame-based square root of the CoV values to achieve a single score for each video as follows:

$$\text{Temporal FS-MOVIE} = TP\left(\sqrt{\frac{\sigma_{Q_T(x,y,t)}}{\mu_{Q_T(x,y,t)}}} \right), \tag{23}$$

where σ_{Q_T} and μ_{Q_T} are the standard deviation and the mean of Q_T, respectively, and TP is a temporal pooling function. Details of TP are described in Section 3.8. We used the square root of CoV values, as in [24], since the range of Temporal FS- MOVIE scores is smaller than that of the Spatial FS-MOVIE scores, due to the divisive normalization in (10) and (11).

3.7. FS-MOVIE index

We first compute the product of the CoV of Q_S in (8) and the square root of the CoV of Q_T in (22) on each frame, then apply the temporal pooling function TP on the product. The product of the CoV of Q_S and the square root of the CoV of Q_T makes FS-MOVIE respond equally to like percentage changes in either the Spatial or Temporal FS-MOVIE Indices. Hence, the ultimate FS-MOVIE Index is defined as

$$\text{FS-MOVIE} = TP\left(\frac{\sigma_{Q_S(x,y,t)}}{\mu_{Q_S(x,y,t)}} \times \sqrt{\frac{\sigma_{Q_T(x,y,t)}}{\mu_{Q_T(x,y,t)}}} \right). \tag{24}$$

3.8. Temporal pooling

The perceptual sensitivity to flicker distortions can be affected by prolonged exposure to flickering stimuli [71]. When an observer is exposed to large flicker distortions over a longer period of time (e.g., 100 ms), flicker visibility may be affected by "visual persistence", [72] whereby a visual stimulus is retained for a period of time beyond the termination of the stimulus. Conversely, when small flicker distortions are prolonged, the HVS dynamically controls the flicker sensitivity, and allocates a finite range of neural signaling, so an observer's flicker sensitivity may be attenuated [71]. Similar accumulation and adaptation processes may contribute to observers' responses to time-varying video quality as a "recency effect" (whereby more recent distortions are more accessible to memory recall) [73], or "temporal

Table 1
List of modified or new model parameters of FS-MOVIE compared with MOVIE [24] including Temporal Hysteresis (TH) pooling [74].

Parameter	Description	MOVIE (+ TH)	FS-MOVIE	Related equation number
φ	Vertical angle of the Gabor filter bank	0°, 30°, and 60°	0°, 20°, 40°, 60° and 80°	(3), (5), (7)–(11), (13)–(15), (20), (21)
θ	Orientation of the Gabor filter bank	20°, 22°, and 40°	18°, 20°, 24°, 36°, and 90°	
R	Maximum attainable response of the simple cell	N/A	4	(14), (21)
σ	Semi-saturation constant of the simple cell	N/A	0.2	(14)
c_i	Weighting factors of the complex cell	N/A	1/3	(15)
τ	Duration of the memory effect	2 s	0.8 s	(25), (26)
β	Combination factor between memory and retention components	0.8	0.1	(28)

hysteresis" [74]. To account for these processes in our VQA model, we used a temporal hysteresis (TH) model of temporal pooling [74], rather than simply averaging the products of the CoV and the square root of the CoV values, as in MOVIE [24]. Other pooling strategies such as weighted summation [75], asymmetric mapping [76], [77], percentile pooling [78], temporal variation of spatial distortions [22], and machine learning [79] might also be of interest. In addition, we enforce the TH model to account for flicker events that occur over short time periods in FS-MOVIE, since we are interested in capturing the effects of local transient flicker events. In the TH pooling model [74], predictions of human judgments follow a smooth trend, drop sharply with predictions of poor video quality, but do not increase as sharply with predictions of improved video quality. Although [74] assumed a memory effect of a longer duration (e.g., 2 s), FS-MOVIE considering local flicker assumes a shorter memory duration (e.g., 0.8 s) and uses all CoV values.

Let $q(t_i)$ represent the time varying score (e.g., the CoV of Q_S in (5) and the square root of CoV of Q_T in (22)) for each frame at a specific time t_i, and let τ indicate the memory duration. First, define a memory component over the previous $t = \tau$ seconds to reflect intolerance due to poor quality video events:

$$l(t_i) = \begin{cases} q(t_i), & t_i = 1 \\ \max[q(t)], t = \max(1, t_i - \tau), t_i - 1, & t_i > 1 \end{cases} \tag{25}$$

Note that a larger CoV value indicates worse quality. Next, a hysteresis component over the next $t = \tau$ seconds is defined to include the hysteresis effect (humans respond quickly to drops in quality but do not increase as sharply when the measured video quality returns to higher quality). Then, sort the time varying scores over the next τ seconds in descending order and combine them using a Gaussian weighting function [80]. Let $\upsilon = \{\upsilon_1, \upsilon_2, \ldots, \upsilon_j\}$ be the sorted elements, and $\omega = \{\omega_1, \omega_2, \ldots, \omega_j\}$ be a descending half Gaussian weighting function that sums to 1. The standard deviation of the half Gaussian weighting function was set to be $(2J - 1)/12$, where J was the total number of sorted elements in υ [74]. For example, the standard deviation of the half Gaussian weighting function was 3.25 and 6.5833 for the 25 frame-per-second (fps) and 50 fps videos, respectively on the LIVE VQA database. We fixed $\tau = 0.8$ s in FS-MOVIE (See Section 4.3). Then

$$\upsilon = sort[q(t)], t = \{t_i + 1, \min(t_i + \tau, T)\}, \tag{26}$$

$$m(t_i) = \sum_{j=1}^{J} \upsilon_j \omega_j, j = \{1, 2, \ldots, J\}, \tag{27}$$

where T is the length of a test video. We then linearly combine the memory and the retention components in (25) and (27) to obtain time varying scores that account for the hysteresis effect. The final video quality is computed as the average of the time varying scores as follows [74]:

$$q'(t_i) = \beta m(t_i) + (1 - \beta) l(t_i), \tag{28}$$

$$TP_{\text{VIDEO}} = \frac{1}{T} \sum_{t=1}^{T} q'(t), \tag{29}$$

where β is a linear combination factor. The criteria for selecting the parameter values of τ and β are detailed in Section 4.3.

3.9. Implementation details

We implemented FS-MOVIE using the public C++ MOVIE code [24], using Microsoft Visual Studio on an x64 platform build. The temporal hysteresis pooling was implemented in MATLAB using the authors' original code [74]. Table 1 shows the list of modified or new model parameters of FS-MOVIE compared with MOVIE. The values of parameters φ, θ, τ, and β were modified to account for flicker events that occur over short time periods, while the new parameters R, σ, and c_i were introduced to model temporal visual masking of local flicker distortions. Other parameter values used in FS-MOVIE are the same as in MOVIE. Lastly, FS-MOVIE computes quality maps on all frames, to better capture flicker events, while MOVIE calculates quality maps every 16th (or 8th) frame.

4. Performance evaluation

4.1. Test setup

We tested the FS-MOVIE Index against human subjective quality scores on the LIVE [55], IVP [56], EPFL [57], and VQEGHD5 [58] VQA databases. The LIVE VQA database [55] consists of 150 distorted videos and 10 reference videos, where six videos contain 250 frames at 25 fps, one video contains 217 frames at 25 fps, and three videos contain 500 frames at 50 fps. They are natural scenes with resolutions of 768 × 432 pixels in YUV 4:2:0 format. Each of the reference videos were simulated to generate 15 distorted videos using four types of distortions: MPEG-2 compression, H.264 compression, transmission over a wireless network, and transmission over IP networks. Difference mean opinion scores (DMOS) were generated from the 38 subjects. The IVP VQA database [56] includes 10 reference videos and 128 distorted videos with resolutions of 1920 × 1088 pixels at 25 fps. Each video has duration of about 10 s. Distorted videos were generated using MPEG-2 compression, Dirac wavelet compression, H.264 compression, and IP network packet loss on H.264 compressed videos. DMOS were obtained from 42 observers. The EPFL VQA database [57] has 12 reference videos and 156 distorted videos, encoded with a H.264 (12 videos) codec, then corrupted by packet loss over an error-prone network (144 videos). It contains one set of 78 videos at CIF resolution (352 × 288) and another set of 78 videos at 4CIF resolution (704 × 576). The videos are of lengths 10 s at 30 fps. Mean opinion scores (MOS) were obtained from the 17 subjects. The VQEGHD5 VQA database [58] consists of 13 reference videos and 155 distorted videos. Each video is of duration 10 s at 25 fps, and of resolution is 1080p. There are two datasets: the specific set contains 144 (9 reference + 9×15 distorted) videos, while the common set includes 24 (4 reference + 4×5 distorted) videos. The distortions include MPEG-2 and H.264 compressions only (bitrate: 2–16 Mbps) and compression plus transmission errors (slicing error and freezing error) caused by burst packet loss. Only 11 (7+4) reference and 125 (7×15 + 4×5) distorted videos are publicly available in the Consumer Digital Video Library (CDVL) [81]. DMOS were obtained from the 24 subjects on these.

To compare the performance of FS-MOVIE against other VQA methods, we tested the following VQA models: PSNR, MS-SSIM [16], VSNR [19], VQM [20], VQM-VFD [28], ST-MAD [25], STRRED [26], and MOVIE [24]. Frame-based VQA algorithms such as PSNR, MS-SSIM,

Table 2
Comparison of VQA algorithm performances on the LIVE [55], IVP [56], EPFL [57], and VQEGHD5 [58] VQA databases: (A) Spearman Rank Ordered Correlation Coefficient (SROCC) and (B) Pearson Linear Correlation Coefficient (PLCC) between the algorithm prediction scores and the MOS or DMOS.

Algorithm	LIVE					IVP					EPFL	VQEGHD5
	Wireless	IP	H.264	MPEG-2	All	Dirac	H.264	MPEG-2	IP	All	All	All
(A)												
PSNR	0.6574	0.4167	0.4585	0.3862	0.5398	0.8532	0.8154	0.6974	0.6284	0.6470	0.7440	0.5120
MS-SSIM	0.7289	0.6534	0.7313	0.6684	0.7364	0.8100	0.8004	0.6503	0.3426	0.5736	0.9222	0.6341
VSNR	0.7019	0.6894	0.6460	0.5915	0.6755	0.7976	0.8670	0.6387	0.6426	0.6663	0.8949	0.6412
VQM	0.7214	0.6383	0.6520	0.7810	0.7026	0.8870	**0.8891**	**0.8625**	0.5835	0.7925	0.9210	0.4606
VQM-VFD	0.7510	**0.7922**	0.6525	0.6361	0.7354	0.8687	0.8471	0.7188	0.6853	0.8071	0.8868	**0.9005**
ST-MAD	0.8099	0.7758	**0.9021**	0.8461	0.8251	0.7228	0.7338	0.7633	0.3207	0.6614	0.8902	0.6963
STRRED	0.7857	0.7722	0.8193	0.7193	0.8007	0.8527	0.8614	0.6774	0.6650	0.7374	0.9380	0.7207
Spatial MOVIE	0.7927	0.7046	0.7066	0.6911	0.7270	0.9057	0.7764	0.8198	0.5835	0.6582	0.9081	0.6961
Temporal MOVIE	0.8114	0.7192	0.7797	0.8170	0.8055	0.8945	0.8430	0.8287	0.7521	0.7956	0.9111	0.7467
MOVIE	0.8109	0.7157	0.7664	0.7733	0.7890	0.9083	0.8400	0.8518	0.7285	0.7668	0.9267	0.7556
Spatial MOVIE with dense Gabor filters	0.7837	0.6979	0.7103	0.6849	0.7337	0.8523	0.7814	0.7873	0.6448	0.6938	0.9098	0.7008
Temporal MOVIE with dense Gabor filters	0.8077	0.7362	0.7606	0.8221	0.8173	0.7798	0.7902	0.7344	**0.7980**	0.7814	0.9227	0.8290
MOVIE with dense Gabor filters	0.8081	0.7522	0.7591	0.7784	0.7957	0.8109	0.7983	0.7704	0.7302	0.7608	0.9354	0.8050
Temporal MOVIE with dense Gabor filters and flicker visibility	0.7961	0.7544	0.7752	0.8489	0.8209	0.7855	0.7889	0.7237	0.7849	0.7916	0.9193	0.8415
MOVIE with dense Gabor filters and flicker visibility	0.8075	0.7562	0.7711	0.8046	0.8094	0.8194	0.7955	0.7722	0.7614	0.7689	**0.9391**	0.8312
Spatial MOVIE with temporal hysteresis pooling [a]	0.8002	0.7731	0.7850	0.7664	0.7919	**0.9128**	0.7679	0.8394	0.6519	0.7467	0.9097	0.7936
Temporal MOVIE with temporal hysteresis pooling [a]	0.7807	0.6983	0.8039	0.8761	0.8127	0.8785	0.8199	0.8305	0.6738	0.8129	0.9073	0.7533
MOVIE with temporal hysteresis pooling [a]	0.8051	0.7664	0.8032	0.8420	0.8296	0.9083	0.8302	0.8452	0.7028	0.8154	0.9278	0.7936
Spatial FS-MOVIE	**0.8255**	0.7802	0.8191	0.7523	0.8074	0.8492	0.7811	0.7704	0.7302	0.7508	0.9297	0.7978
Temporal FS-MOVIE	0.7897	0.7184	0.8148	**0.8898**	0.8413	0.7980	0.7698	0.7415	0.7975	**0.8156**	0.9245	0.8667
FS-MOVIE	0.8139	0.7722	0.8490	0.8609	**0.8482**	0.8300	0.7831	0.7637	0.7575	0.8067	0.9381	0.8408
(B)												
PSNR	0.7058	0.4767	0.5746	0.3986	0.5645	0.8952	0.8741	0.6431	0.5861	0.6453	0.7428	0.5492
MS-SSIM	0.7184	0.7764	0.7420	0.6222	0.7470	0.8708	0.8358	0.7368	0.4444	0.5896	0.9218	0.6389
VSNR	0.7191	0.7541	0.6295	0.6793	0.6983	0.8184	0.8933	0.6732	0.6770	0.6758	0.8993	0.6625
VQM	0.7548	0.6666	0.6660	0.8132	0.7301	**0.9268**	0.8978	0.9145	0.6557	0.7860	0.9200	0.4937
VQM-VFD	0.8144	**0.8616**	0.7403	0.7172	0.7763	0.9038	0.8765	0.8110	0.6547	0.8119	0.8866	**0.9020**
ST-MAD	0.8591	0.8065	**0.9155**	0.8560	0.8332	0.8084	0.7864	0.8076	0.4579	0.6702	0.8911	0.7122
STRRED	0.8053	0.8527	0.8141	0.7570	0.8119	0.8676	0.8836	0.8122	0.6568	0.7336	0.9398	0.7380
Spatial MOVIE	0.8232	0.7590	0.7702	0.7130	0.7520	**0.9268**	0.8093	0.9204	0.4842	0.6622	0.9113	0.6966
Temporal MOVIE	0.8431	0.7782	0.8133	0.8410	0.8264	0.8973	0.8891	0.8213	0.7753	0.7955	0.9112	0.7417
MOVIE	0.8475	0.7657	0.8143	0.7983	0.8134	0.9077	**0.9043**	0.8648	0.6821	0.7577	0.9260	0.7485
Spatial MOVIE with dense Gabor filters	0.8177	0.7658	0.7733	0.7160	0.7572	0.8644	0.8382	0.8230	0.6688	0.6932	0.9176	0.6960
Temporal MOVIE with dense Gabor filters	0.8467	0.8042	0.8028	0.8496	0.8398	0.7996	0.8334	0.7487	0.7931	0.7793	0.9355	0.8369
MOVIE with dense Gabor filters	0.8517	0.7857	0.8027	0.7986	0.8230	0.8205	0.8325	0.7361	0.7400	0.7550	0.9425	0.8037
Temporal MOVIE with dense Gabor filters and flicker visibility	0.8495	0.7995	0.8082	0.8672	0.8443	0.8122	0.8390	0.7800	0.7588	0.7868	0.9328	0.8498
MOVIE with dense Gabor filters and flicker visibility	0.8401	0.8016	0.7947	0.8258	0.8324	0.8224	0.8348	0.7306	0.7337	0.7672	0.9474	0.8324
Spatial MOVIE with temporal hysteresis pooling [a]	0.8414	0.8401	0.8073	0.7718	0.8118	0.9232	0.8043	**0.9208**	0.6100	0.7520	0.9211	0.7930
Temporal MOVIE with temporal hysteresis pooling [a]	0.8607	0.7485	0.8266	0.8864	0.8281	0.8748	0.8631	0.8351	0.6684	**0.8199**	0.9041	0.7562
MOVIE with temporal hysteresis pooling [a]	0.8563	0.8032	0.8414	0.8482	0.8470	0.9003	0.8752	0.8695	0.7750	0.8130	0.9274	0.7932
Spatial FS-MOVIE	0.8683	0.8598	0.8450	0.7692	0.8176	0.8572	0.8449	0.8015	0.7182	0.7502	0.9316	0.8029
Temporal FS-MOVIE	**0.8735**	0.7899	0.8512	**0.9160**	**0.8673**	0.8075	0.8272	0.7223	**0.8046**	0.8144	0.9397	0.8783
FS-MOVIE	0.8599	0.8009	0.8765	0.8721	0.8636	0.8321	0.8313	0.7305	0.7917	0.8053	**0.9504**	0.8544

[a] The scores are obtained from the most updated code provided by the authors of [74].

and VSNR were applied frame-by-frame and the average score across all frame used as a final measure of quality. We used the Metrix Mux toolbox [82] implementations of PSNR, MS-SSIM, and VSNR. The video frames of the reference and distorted videos were correctly aligned. For other VQA algorithms, we used the source code provided by the authors' webpages. Furthermore, we studied the performance of FS-MOVIE against various relevant configurations of MOVIE, including MOVIE with dense Gabor filters, MOVIE with dense Gabor filters and flicker visibility, and MOVIE with temporal hysteresis pooling. This was done to isolate the effects of the differences between FS-MOVIE and MOVIE.

4.2. Algorithm performance

4.2.1. Correlation with human subjective judgments

We used the Spearman Rank Order Correlation Coefficient (SROCC) and the Pearson Linear Correlation Coefficient (PLCC) after nonlinear regression in [83] between the human scores and the model indices to compare the performances between the VQA models. We linearly rescaled VQA model scores to ensure numerical convergence as used in [24].

Table 2 shows the algorithm performances using SROCC and PLCC for each distortion type and over all videos on the tested VQA databases. In each column, the bold font highlights the top performing algorithm. Overall, ST-MAD, STRRED, MOVIE, and FS-MOVIE yielded better performances on the LIVE VQA database, while VQM, VQM-VFD, MOVIE, and FS-MOVIE achieved better performances on the IVP database. MS-SSIM, VQM, STRRED, MOVIE, FS-MOVIE delivered good performances on the EPFL VQA database, while VQM- VFD, MOVIE, and FS-MOVIE, yielded superior performances on the VQEGHD5 VQA database. MS-SSIM was better than PSNR on the LIVE, EPFL, and VQEGHD5 VQA databases, but oddly, not on the IVP VQA database. VQM-VFD achieved excellent performance on the VQEGHD5 VQA database, which simulates frame delays such as freezing errors.

It is clear from the results that although MOVIE effectively predicts spatial and temporal video distortions, outperforming PSNR, MS-SSIM, VSNR, VQM, and VQM-VFD on the LIVE VQA database, FS-MOVIE strongly improves the performance of MOVIE by accounting for temporal visual masking of local flicker distortions. In addition, the enhanced versions of MOVIE using the various elements of FS-MOVIE present progressive improvements in performance. For example, the SROCC 0.7890 achieved by MOVIE is improved to 0.7957 when MOVIE is augmented with dense Gabor filters, and to 0.8094 when augmented with dense Gabor filters and flicker visibility, respectively. When MOVIE with dense Gabor filters and flicker visibility is combined with the hysteresis temporal pooling model, thereby creating the full FS-MOVIE model, the SROCC is improved to 0.8482 exceeding performances of all the tested VQA models. The superior performance of Temporal FS-MOVIE shown in Table 2 highlights the perceptual efficacy accounting for temporal masking of local flicker distortions.

On the IVP VQA database, MOVIE performance was also noticeably improved by the FS-MOVIE enhancements, where the respective SROCC values were 0.7668 and 0.8067, and the PLCC values were 0.7577 and 0.8053, respectively. On the EPFL VQA database, STRRED and FS-MOVIE exhibited similar higher monotonicity (SROCC, STRRED: 0.9380, FS-MOVIE: 0.9381), while FS-MOVIE yielded the best linearity (PLCC, STRRED: 0.9398, FS-MOVIE: 0.9504) as shown in Table 2. This result can be also observed in the scatter plots in Fig. 8 between the algorithm scores and the MOS on the EPFL VQA database. On the VQEGHD5 VQA dataset, similar to other VQA databases, (Temporal) MOVIE performance was significantly improved by (Temporal) FS-MOVIE, where the respective PLCC values were (0.7417) 0.7485 and (0.8783) 0.8544, respectively.

Regarding the model performances across distortion types, FS-MOVIE delivered stable results, although FS-MOVIE performed a little better on the MPEG-2 and H.264 compressed videos on the LIVE VQA database. Across VQA algorithms, Spatial FS-MOVIE, VQM-VFD, ST-MAD, and Temporal FS-MOVIE yielded better performance on the Wireless, IP, H.264, and MPEG-2 distortion types, respectively, in terms of SROCC on the LIVE VQA database, while Spatial MOVIE with TH, VQM, MOVIE, and Temporal FS-MOVIE performed better on the Dirac, H.264, MPEG-2, and IP distortion types, respectively, in terms of SROCC on the IVP VQA database.

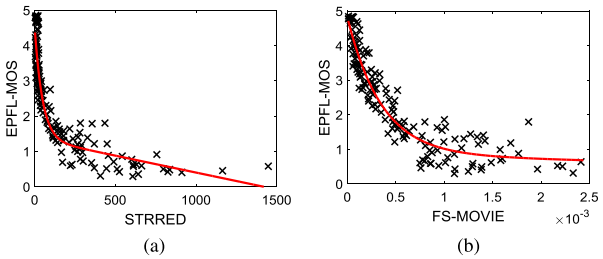


Fig. 8. Scatter plots of the objective VQA scores against MOS for all videos in the EPFL VQA database: (a) STRRED and (b) FS-MOVIE.

4.2.2. Statistical significance

We also tested the statistical significance of the results presented in Section 4.2.1, using an F-test based on the residuals between the averaged human ratings (e.g., DMOS) and the model predictions. Statistical significance test shows whether the performance of one objective model was statistically superior to that of a competing objective model. The residual error between the quality predictions of an objective VQA model and the DMOS values on the VQA databases was used to test the statistical superiority of one model over another [55], [58], [84], [85]. An F-test was performed on the ratio of the variance of the residual error from one objective model to that of another objective model at the 95% significance level. The null hypothesis states that variances of the error residuals from the two different objective models were equal. The F-ratio is always formed by placing the objective model with the larger residual error variance in the numerator. Threshold F-ratios can be determined based on the number of video sequences in each database and on the significance level. For example, on the LIVE VQA database, the total number of video sequences is 150, and the threshold F-ratio is 1.31 at the 95% significance level. An F-ratio larger than the threshold indicates that the performance of the VQA model in the numerator of the F-ratio is statistically inferior to that of the VQA model in the denominator. We executed an F-test using the MATLAB function `vartest2` at the 95% significance level ($\alpha = 0.05$) with the tail option 'right' and 'left' separately, then obtained the final statistical significance.

The F-test assumes that the residuals are independent samples from a normal Gaussian distribution [86]. To validate the assumption, we used the Kurtosis-based criterion for Gaussianity in [83]: if the residuals have a kurtosis between 2 and 4, they are taken to be Gaussian. We verified that the residuals were almost normally distributed, and that the means of the residuals were almost zero for the tested models.

Specifically, 100%, 86%, 100%, and 95% of the residuals have the Gaussian normal distribution on the LIVE, IVP, EPFL, and VQEGHD5 VQA databases, respectively. VQM, VQM-VFD, and S-MOVIE + TH on the IVP VQA database as well as VQM on the VQEGHD5 VQA database did not satisfy Kurtosis-based criterion for Gaussianity. Fig. 9 shows the histogram of residuals between the quality predictions of the objective model and the DMOS values on the LIVE VQA database, with mean and kurtosis values.

The results of the statistical significance test are shown in Table 3. Each entry in the table is a code-word consisting of four symbols, which correspond to the LIVE, IVP, EPFL, and VQEGHD5 datasets, in that order. The symbol '1' in the table indicates that the row algorithm

Table 3
Statistical analysis of VQA algorithm performances on the LIVE, IVP, EPFL, and VQEGHD5 VQA databases. The symbol '1' in the table indicates that the row (algorithm) is statistically better than the column (algorithm), while the symbol '0' indicates that the row is worse than the column; the symbol '-' indicates that the row and column are not significantly different. The symbol '×' denotes that the statistical significance could not be determined since the Gaussianity was not satisfied in the F-test. In each cell, entities denote performance on the LIVE, IVP, EPFL, and VQEGHD5 VQA databases, in that order.

	PSNR	MS-SSIM	VSNR	VQM	VQM-VFD	ST-MAD	STRRED	S-MOVIE	T-MOVIE	MOVIE	Dense-S-MOVIE	Dense-T-MOVIE	Dense-MOVIE	Dense-Flicker-T-MOVIE	Dense-Flicker-MOVIE	S-MOVIE + TH	T-MOVIE + TH	MOVIE + TH	S-FS-MOVIE	T-FS-MOVIE	FS-MOVIE
PSNR	- - - -	0–00	0000	0×0×	0×00	0–00	0000	0–00	0000	0000	0–00	0000	0000	0000	0000	0×00	0000	0000	0000	0000	0000
MS-SSIM	1–11	- - - -	-01-	-×-×	0×10	0-1-	000-	- -10	00–0	00–0	-0-0	0000	0000	0000	0000	0×-0	00–0	0000	0000	0000	0000
VSNR	1111	-10-	- - - -	-×0×	0×-0	01- -	0-0-	010-	000-	0–00	0-0-	0000	0000	0000	0000	0×00	0000	0000	0000	0000	0000
VQM	1×1×	-×-×	- ×1×	-×-×	0×1×	0×1×	0×0×	-×1×	0×-×	0×-×	0×-×	0×0×	0×0×	0×0×	0×0×	0×-×	0×-×	0×-×	0×0×	0×0×	0×0×
VQM-VFD	1×11	1×01	1× -1	1×0×	-×- -	-×-1	-×01	1×01	-×01	-×01	1×01	-×01	-×01	0×01	0×01	-×01	-×01	0×01	-×01	0×0-	0×0-
ST-MAD	1–11	1-0-	10- -	1×0×	-×-0	- - - -	-00-	1- -0	-000	-000	1000	-000	-000	0000	-000	-×00	-000	0000	-000	0000	0000
STRRED	1111	111-	1-1-	1×1×	-×10	-11-	- - - -	111-	-01-	- -10	1-1-	-000	-000	00–0	0000	-×-0	-010	00–0	- - -0	0000	0000
S-MOVIE	1–11	- - 01	101-	-×0×	0×10	0- -1	000-	- - - -	00- -	000-	-00-	0000	0000	0000	0000	0×00	00- -	0000	0000	0000	0000
T-MOVIE	1111	11–1	111-	1×-×	-×10	-111	-10-	11- -	- - - -	-10-	11- -	- -00	-100	- -00	- -00	1×00	- - - -	- -00	-100	0000	0-00
MOVIE	1111	11–1	1–11	1×-×	-×10	-111	- -01	111-	-01-	- - - -	1- - -	-000	-000	0000	-000	-×-0	-01-	00–0	- -00	0000	0000
Dense-S-MOVIE	1–11	-1-1	1-1-	1×-×	0×10	0111	0-0-	-11-	00- -	0- - -	- - - -	0000	0000	0000	0000	0×-0	00- -	00–0	0000	0000	0000
Dense-T-MOVIE	1111	1111	1111	1×1×	-×10	-111	-111	1111	- -11	-111	1111	- - - -	-1-1	- - - -	- - - -	1×11	-011	- -11	-111	00–0	0-0-
Dense-MOVIE	1111	1111	1111	1×1×	-×10	-111	-111	1111	-011	-111	1111	-0-0	- - - -	-0-0	- - -0	-×1-	-011	0-1-	- - - -	00–0	0000
Dense-Flicker-T-MOVIE	1111	1111	1111	1×1×	1×10	1111	11–1	1111	- -11	1111	1111	- - - -	-1-1	- - - -	-10-	1×11	- -11	- -11	11–1	0- -0	- -0-
Dense-Flicker-MOVIE	1111	1111	1111	1×1×	1×10	-111	1111	1111	- -11	-111	1111	- - - -	- - -1	-01-	- - - -	1×11	-011	- -11	- -11	00–0	0000
S-MOVIE+TH	1×11	1×-1	1×11	1×-×	-×10	-×11	-×-1	1×11	0×11	-×-1	1×-1	0×00	-×0-	0×00	0×00	-×- -	0×1-	0×- -	-×0-	0×00	0×00
T-MOVIE+TH	1111	11–1	1111	1×-×	-×10	-111	-101	11- -	- - - -	-10-	11- -	-100	-100	- -00	-100	1×0-	- - - -	-10-	-10-	0–00	0-00
MOVIE+TH	1111	1111	1111	1×-×	1×10	1111	11–1	1111	- -11	11–1	11–1	- -00	1-0-	- -00	- -00	1×- -	-01-	- - - -	110-	-000	- -00
S-FS-MOVIE	1111	1111	1111	1×1×	-×10	-111	- - -1	1111	-011	- -11	1111	-000	- - - -	00–0	- -00	-×1-	-01-	001-	- - - -	0000	0000
T-FS-MOVIE	1111	1111	1111	1×1×	1×1-	1111	1111	1111	1111	1111	1111	11–1	11–1	1- -1	11–1	1×11	1–11	-111	1111	- - - -	- -0-
FS-MOVIE	1111	1111	1111	1×1×	1×1-	1111	1111	1111	1–11	1111	1111	1-1-	1111	- -1-	1111	1×11	1–11	- -11	1111	- -1-	- - - -

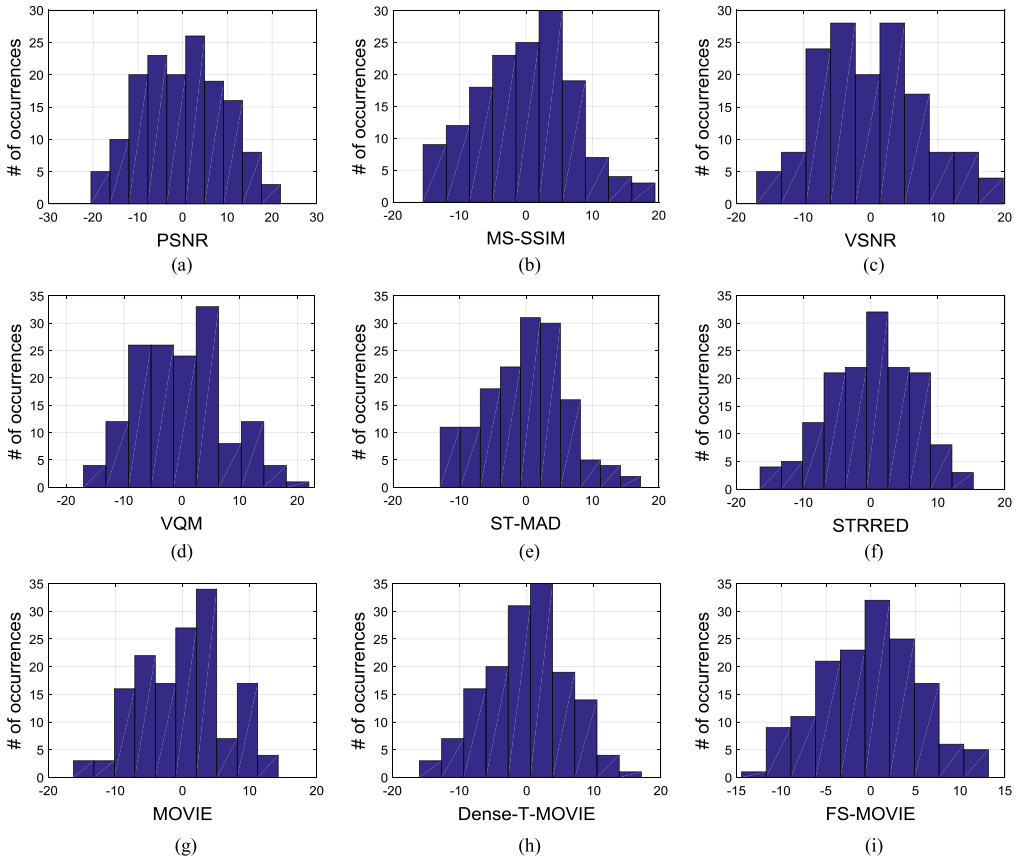


Fig. 9. The histogram of the residuals between the quality predictions of an objective model and the DMOS values on the LIVE VQA database. When the residuals have a kurtosis between 2 and 4, they are taken to be Gaussian. (a) PSNR (mean = 0, kurtosis = 2.3), (b) MS-SSIM (mean = 0, kurtosis = 2.7), (c) VSNR (mean = −0.0048, kurtosis = 2.5), (d) VQM (mean = 0, kurtosis = 2.7), (e) ST-MAD (mean = 0, kurtosis = 2.8), (f) STRRED (mean = 0, kurtosis = 2.7), (g) MOVIE (mean = 0, kurtosis = 2.6), (h) Temporal MOVIE with dense Gabor filters, Dense-T-MOVIE (mean = 0, kurtosis = 2.9), and (i) FS-MOVIE, (mean = 0, kurtosis = 2.6).

is statistically better than the column algorithm, while the symbol '0' indicates that the row is statistically worse than the column. The symbol '-' denotes that the row and column are not statistically different. The symbol 'x' denotes that statistical significance could not be determined because the criterion for Gaussianity was not satisfied. For example, the first symbol value of '1' at the second row and at the first column in Table 3 means that MS-SSIM is statistically better than PSNR on the LIVE VQA database. From Table 3, it is obvious that Temporal FS-MOVIE and FS-MOVIE were either statistically superior or competitive with the other tested objective VQA algorithms, including the predecessor MOVIE, on the LIVE, IVP, EPFL, and VQEGHD5 VQA databases. The results imply that the temporal visual masking factor is important to VQA improvement.

4.2.3. Bitrate influence

It is of great interest to study how video compression bitrates influence the visibility of flicker distortions. We produced two categories of "low bitrate videos" and "high bitrate videos" from the H.264 and MPEG-2 compressed videos on the LIVE VQA database. Specifically, a total of 20 low bitrate videos (the lowest bitrate videos per content and compression type) and a total of 20 high bitrate videos (the highest bitrate videos per content and compression type) were tested. The low compression bitrates were about 700 kbps for MPEG-2 and 200 kbps for H.264 compression, respectively, whereas the high compression bitrates were about 4 Mbps for MPEG-2 and 5 Mbps for H.264 compressions [55]. Table 4A shows the PLCC between the algorithm scores and the DMOS values for the VQA algorithms on the low bitrate videos and high bitrate videos on the LIVE VQA database. The third column of Table 4A shows the result when both the low and high bitrate videos are tested. In each column, the bold font highlights the top performing VQA algorithm.

The VQA models largely performed better on the low bitrate videos than on the high bitrate videos. VQM-VFD, ST-MAD, and Temporal FS-MOVIE achieved higher PLCC (over 0.8) on the low bitrate videos, while ST-MAD obtained higher PLCC (over 0.7) on the high bitrate videos. When all low and high bitrate videos were tested on, FS-MOVIE achieved the best performance (PLCC: 0.9208) among the tested methods. Although MOVIE and Temporal MOVIE performed better on the low bitrate videos against FS-MOIVE and Temporal FS-MOVIE, FS-MOVIE (PLCC: 0.7616 on the high bitrate videos and 0.9208 on the all videos) and Temporal FS-MOVIE (PLCC: 0.8778 and 0.9205) strongly improved the performance of MOVIE (PLCC: 0.6472 and 0.8618) and Temporal MOVIE (PLCC: 0.7410 and 0.8721) on both the high bitrate and all videos, respectively. This result demonstrates that FS-MOVIE captured the perceptually suppressed flicker distortion visibility well when large object motions exist and when the quality of the test video sequence is poor, in agreement with the results of the human studies in [62].

4.2.4. Motion influence

To analyze motion influences on the performance of VQA models, we categorized all videos on the LIVE VQA database into two subsets of videos: those with small motions and those with large motions. Using the optical flow values obtained in Section 3.3, we averaged the velocity

Table 4
Comparison of the VQA algorithm performances influenced by (A) the bitrate and by (B) the motion on the LIVE VQA database. PLCC between the VQA algorithm scores and the DMOS is shown. The low bitrates were about 700 kbps for MPEG-2 and 200 kbps for H.264 compression, while the high bitrates were about 4 Mbps for MPEG-2 and 5 Mbps for H.264 compression [55].

Algorithm	(A)			(B)		
	Low Bitrate	High Bitrate	All	Small Motion	Large Motion	All
PSNR	0.4896	0.1063	0.5708	0.6419	0.5215	0.5645
MS-SSIM	0.5183	0.5633	0.8003	0.8085	0.7115	0.7470
VSNR	0.4784	0.5354	0.7068	0.7833	0.6673	0.6983
VQM	0.6895	0.5444	0.7909	0.6594	0.8378	0.7301
VQM-VFD	0.8064	0.6494	0.7849	0.7212	0.8464	0.7763
ST-MAD	0.8070	**0.7303**	0.8960	0.8017	0.8660	0.8332
STRRED	0.6993	0.5376	0.8691	0.7388	0.8391	0.8119
Spatial MOVIE	0.7145	0.6679	0.8342	0.7589	0.7605	0.7520
Temporal MOVIE	0.7410	0.6744	0.8721	0.7851	0.8833	0.8264
MOVIE	0.6472	0.5963	0.8618	0.7936	0.8571	0.8134
Spatial MOVIE with dense Gabor filters	0.7225	0.6562	0.8338	0.7639	0.7665	0.7572
Temporal MOVIE with dense Gabor filters	0.7183	0.6266	0.8898	0.7698	0.9125	0.8398
MOVIE with dense Gabor filters	0.6638	0.5822	0.8630	0.7870	0.8673	0.8230
Temporal MOVIE with dense Gabor filters and flicker visibility	0.7885	0.6603	0.8904	0.7706	0.9201	0.8443
MOVIE with dense Gabor filters and flicker visibility	0.6741	0.4133	0.8753	0.7783	0.8922	0.8324
Spatial MOVIE with temporal hysteresis pooling	0.7568	0.6211	0.8787	0.8139	0.8253	0.8118
Temporal MOVIE with temporal hysteresis pooling	0.7757	0.6659	0.8979	0.7789	0.8933	0.8281
MOVIE with temporal hysteresis pooling	0.7468	0.4986	0.9070	0.8140	0.8952	0.8470
Spatial FS-MOVIE	0.7553	0.5990	0.8955	**0.8223**	0.8420	0.8176
Temporal FS-MOVIE	**0.8778**	0.6160	0.9205	0.8081	**0.9271**	**0.8673**
FS-MOVIE	0.7616	0.5185	**0.9208**	0.8105	0.9231	0.8636

Table 5
Comparison of VQA algorithm performances as influenced by temporal subsampling on the LIVE VQA database. (A) SROCC and (B) PLCC results over all videos sampled at every 8, 4, 2, and 1 frames are shown.

Algorithm	Temporal subsampling at every			
	8 frames	4 frames	2 frames	1 frame
(A)				
Spatial MOVIE with dense Gabor filters	0.7330	0.7339	0.7340	0.7337
Temporal MOVIE with dense Gabor filters	0.8156	0.8169	0.8171	0.8173
MOVIE with dense Gabor filters	0.7926	0.7949	0.7952	0.7957
Spatial FS-MOVIE	0.8114	0.8203	0.8237	0.8074
Temporal FS-MOVIE	0.8156	0.8234	0.8354	0.8413
FS-MOVIE	**0.8307**	**0.8391**	**0.8463**	**0.8482**
(B)				
Spatial MOVIE with dense Gabor filters	0.7565	0.7572	0.7572	0.7572
Temporal MOVIE with dense Gabor filters	0.8359	0.8389	0.8395	0.8398
MOVIE with dense Gabor filters	0.8177	0.8215	0.8230	0.8230
Spatial FS-MOVIE	0.8299	0.8415	0.8389	0.8176
Temporal FS-MOVIE	0.8337	0.8419	0.8535	**0.8673**
FS-MOVIE	**0.8436**	**0.8534**	**0.8615**	0.8636

magnitude on a frame-by-frame basis, then the average magnitude over all frames was used as a final measure of motion. For the fair comparison of motions between the 25 fps videos and 50 fps videos, we summed the velocity magnitude for every two consecutive frames, then averaged the sum on the 50 fps videos. Among the 10 contents, "st", "rh", "mc", "sf", and "pr" were categorized as having small motions, while "bs", "pa", "sh", "tr", and "rb" were considered as having large motions.

PSNR, MS-SSIM, and VSNR generally performed better on the small motion videos rather than on the large motion videos, as shown in Table 4B. In contrast VQM, VQM-VFD, ST-MAD, STRRED, MOVIE, and FS-MOVIE performed better on the large motion videos rather than on the small motion videos. FS-MOVIE and Temporal FS-MOVIE show the best performances on the large motion videos, which suggests that FS-MOVIE effectively accounts for perceptually suppressed flicker visibility when large object motions exist.

We observed that the compressed versions of "rb", "tr", and "mc" contain relatively higher flicker, and they were accurately assessed by FS-MOVIE with small regression errors. Note that the "mc" video sequence includes both small motions and large motions, although the overall amount of motion is small. We illustrate the flicker visibility index on small motions and large motions in Fig. 7. We also observed that FS-MOVIE relatively does not perform well on the "st", "rh", and "sf" videos with larger regression errors compared to the "tr" and "mc" videos. This might be due to the small amount of motion in those videos.

4.2.5. *Influence of temporal subsampling*

To understand how temporal subsampling affects the performance of MOVIE and FS-MOVIE, we compared the SROCC and PLCC results over all videos where the VQA values were sampled at every 8, 4, 2, and 1 frames on the LIVE VQA database. We applied the same Gabor filter configurations described in Section 3.1.

Table 5A and 5B show the algorithm performances in terms of SROCC and PLCC, respectively. In each column, the bold font highlights the top performing algorithm. FS-MOVIE yielded increasingly better

Table 6
Comparison of performances using various temporal pooling methods on the LIVE VQA database. (A) SROCC and (B) PLCC are shown when spatial, temporal, and spatiotemporal quality scores obtained by the dense Gabor filters and flicker visibility model were fed into the tested temporal pooling methods.

	Simple average	Percentile pooling	Asymmetric mapping	Temporal hysteresis
(A)				
Spatial quality scores	0.7337	0.7882	0.6712	**0.8074**
Temporal quality scores	0.8209	0.8064	0.7192	**0.8413**
Spatiotemporal quality scores	0.8094	0.8201	0.7350	**0.8482**
(B)				
Spatial quality scores	0.7572	0.8106	0.7004	**0.8176**
Temporal quality scores	0.8443	0.8199	0.7510	**0.8673**
Spatiotemporal quality scores	0.8324	0.8359	0.7643	**0.8636**

Table 7
Computational complexity analysis: (A) VQA algorithms. (B) MOVIE and FS-MOVIE for each step per frame on the Tractor (768 × 432) video.

(A)

Algorithm	Tractor (768 × 432 pixels, 25 fps, 10 s)
PSNR	3.09 s
MS-SSIM	11.34 s
VSNR	41.82 s
VQM	35.87 s
VQM-VFD	74.82 s
ST-MAD	335.90 s
STRRED	54.94 s
MOVIE	5.73 h
FS-MOVIE	74.70 h

(B)

Item	MOVIE	FS-MOVIE
Reading a frame	0.06 s	0.06 s
Gabor filtering per frame	268.63 s	462.24 s
Computing optical flows per frame	416.44 s	712.03 s
Computing indices per frame	27.27 s	60.34 s
The total amount of runtime per frame	712.40 s	1234.47 s

performances, while MOVIE delivered similar performances when more frames were included. For example, the SROCC 0.7926 obtained by MOVIE with dense Gabor filters subsampled every 8 frames remained at 0.7957 when all frames were used, whereas the SROCC 0.8307 obtained by FS-MOVIE improved to 0.8482 when all frames were used. Temporal FS-MOVIE achieved the most significant increase in performance (SROCC increased from 0.8156 to 0.8413) when all frames were used instead of just every 8th frame. These results suggest that flicker distortions that occur over very short time periods are more effectively captured by the (Temporal) FS-MOVIE when more temporal frames are used.

4.2.6. Influence of temporal pooling

We also analyzed the influence of temporal pooling on VQA performance. Four pooling approaches were tested on the LIVE VQA database: the average [24], percentile pooling [20], [78], asymmetric mapping [76], [77], and the temporal hysteresis used in FS-MOVIE, as described in Section 3.8. Spatial, temporal, and spatiotemporal quality scores achieved using dense Gabor filters and the flicker visibility model were fed into the four temporal pooling methods. Specifically, let $q(t_i)$ denote the time varying score (e.g., the CoV of Q_S in (5) and the square root of CoV of Q_T in (22)) for each frame at a time t_i. Percentile pooling [20], [78] weights lower quality regions heavily. For each video, the final percentile pooling score q_p is obtained using the lowest 6% of the quality scores with $w_r = 4000$ [78] as follows,

$$q_p = \frac{w_r \times q(t_i)_{\downarrow 6\%} + 1 \times q(t_i)_{\uparrow 94\%}}{w_r \times N_{\downarrow 6\%} + 1 \times N_{\uparrow 94\%}}, \tag{30}$$

where $q(t_i)_{\downarrow 6\%}$ and $q(t_i)_{\uparrow 94\%}$ are the $q(t_i)$ values belonging to the lowest 6% and the highest 94% of quality scores, while $N_{\downarrow 6\%}$ and $N_{\uparrow 94\%}$ are the number of quality scores belonging to the lowest 6% and the highest 94%, respectively. Asymmetric mapping of $q_a(t_i)$ is calculated by a causal, low-pass function of $q(t_i)$ [76],

$$q_a(t_i) = \begin{cases} q_a(t_{i-1}) + \alpha_- \Delta(t_i), \text{ if } \Delta(t_i) \le 0, \\ q_a(t_{i-1}) + \alpha_+ \Delta(t_i), \text{ if } \Delta(t_i) > 0, \end{cases} \tag{31}$$

where $\Delta(t_i) = q(t_i) - q_a(t_i)$, and the values of α_+ and α_- were 0.0030 and 0.0044, respectively as in [76]. These different weights α_+ and α_- were used to measure asymmetrical responses to sustained increases and decreases in frame-level quality over time. For each video, the overall video quality was computed as the average asymmetrical mapping $q_a(t_i)$.

Table 6 shows the SROCC and PLCC results of the tested four temporal pooling methods on the LIVE VQA database. In each row, the bold font denotes the top performing pooling method. When compared with simple average pooling, percentile pooling gave good performances on spatial quality scores, but not on temporal quality scores. Among tested temporal pooling strategies, temporal hysteresis pooling adjusting the flicker masking in FS-MOVIE showed the best performance and significantly improved the simple average pooling (corresponding to MOVIE with dense Gabor filters and flicker visibility). These results imply the importance of flicker accumulation and adaptation as well as the recency and temporal hysteresis effects in VQA. Percentile pooling gave better performance on spatial quality scores, but not on temporal quality scores, as compared with simple average pooling.

4.2.7. Computational complexity

Table 7A tabulates the runtime results required by each VQA model when predicting quality on "Tractor" with 768 × 432 pixels, 25 fps, and 10 s from the LIVE VQA database. Source code was obtained from each author's web sites. A Windows 7 PC with Intel® Core™ i7-6700 K CPU @4.0 GHz processor and 32 GB of RAM was used. Table 7B details the runtime at each step of MOVIE and FS-MOVIE per frame on the "Tractor" video.

Table 7B shows that Gabor decomposition and optical flow computation dominated the complexity of both MOVIE and FS-MOVIE, taking about 38% and 58% of all computation processes. The increased time to compute FS-MOVIE, relative to MOVIE, largely results from the increased number of Gabor filters, yielding an increase of about 1.7× per frame. MOVIE computes indices at every 8th frame of a test video, while FS-MOVIE computes indices on all frames, causing another 8-fold increase in runtime per video. When multiple videos are assessed with the same reference video, the results of Gabor filtering and optical flow computation of the reference video can be reused reducing a significant amount of runtime.

Our non-optimized implementation of FS-MOVIE has a high computational load, since it involves a large number of serial processes of Gabor filtering and optical flow computation. Better hardware based programming, using a GPU-accelerated NVIDIA CUDA implementation that enables a large amount of parallel processing and a specialized memory hierarchy might significantly reduce the computational loads, as shown in the recent GPGPU based implementation [87]. Fast Gabor

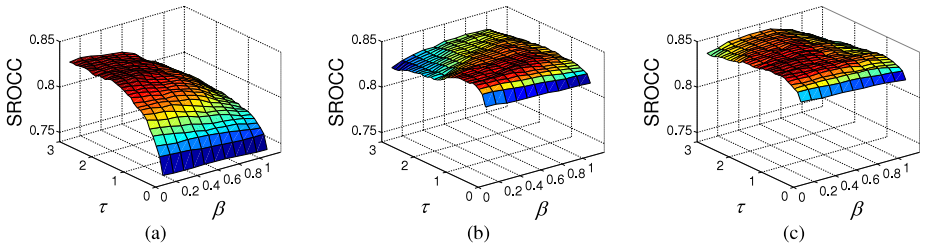


Fig. 10. SROCC performance of FS-MOVIE as functions of the duration of the memory effect τ (seconds) and the linear combination factor β on the LIVE VQA database. (a) Spatial FS-MOVIE. (b) Temporal FS-MOVIE. (c) FS-MOVIE. Hot colors indicate better SROCC performance. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

filtering and fast optical flow estimation methods also might further reduce the computational complexity.

4.3. Parameter variation of temporal pooling

We tested FS-MOVIE with different values of the hysteresis parameters such as the duration of the memory effect τ and the linear combination factor β. Fig. 10 demonstrates SROCC performance plotted against τ and β values on the LIVE VQA database. As shown in Fig. 10, shorter memory durations yielded better results, especially for Temporal FS-MOVIE, since flicker is a transient phenomenon. Spatial FS-MOVIE yielded better performance for larger values of τ. FS-MOVIE achieved stable SROCC values over the ranges $0.5 \leq \tau \leq 1.5$ s and $0 \leq \beta \leq 0.6$. We chose $\tau = 0.8$ s and $\beta = 0.1$ in FS-MOVIE, and applied the same parameter values in all of the performance evaluations on the tested LIVE, IVP, EPFL, and VQEGHD5 VQA databases.

5. Discussions and conclusion

We proposed a new VQA model called FS-MOVIE that accounts for temporal visual masking of local flicker in distorted videos by augmenting the MOVIE framework. We described how a simple neural model of Gabor decomposition, an energy model of motion, and divisive normalization can be used to quantify the local spectral signatures of local flicker distortions in a video which can be used to predict perceptual flicker visibility. Predicting suppressed local flicker distortions significantly improves VQA performance. Results show that FS-MOVIE correlates quite well with human judgments of video quality and is competitive with modern VQA algorithms. Although the proposed FS-MOVIE is motivated by a recently discovered visual change silencing phenomenon on synthetic stimuli such as moving dots [40], temporal visual masking of local flicker also occurs on natural videos.

Although the range of compression parameters on the LIVE, IVP, EPFL, and VQEGHD5 VQA databases does not generate many visually obvious local flicker distortions, FS-MOVIE nevertheless achieves a significant improvement in VQA performance relative to MOVIE. More severe flicker does occur in practice, so it would be of interest to conduct a large human subjective study that includes more severe flickers and diverse types of flickers such as counter-phase flickers and edge flickers. Although we tested bitrate effects on algorithm performance, due to the lack of enough test videos of diverse resolutions (assuming the same content, the same QP values, the same display size, and the same viewing distance), we could not test spatial resolution effects on flicker distortions. It would be interesting to study this as future work. Building a database of time-varying flickering video data like [85] would also help in the design of better flicker-sensitive VQA models. We think that the use of higher frame rates during video recording, or motion compensated frame insertion, could be also helpful to reduce the perception of quantization flicker. Future work may aim at developing temporal visual masking models of other temporal distortions (e.g., strobing artifacts).

It is also important to note that motion on the retina, not in space, is responsible for the motion silencing phenomena, as shown in [40]. The current version of FS-MOVIE does not account for the effects of relative motion with respect to gaze shift, which is a limitation worth considering in future implementations. To analyze the impact of relative motion on temporal flicker masking effects, one could employ gaze tracking data that would distinguish between sequences where fixed gaze points were recorded, and sequences where an observer's eye movements occurred. In addition, since motion silencing is a function of eccentricity [42], [63], it would be worthwhile to consider the effects of eccentricity on temporal flicker masking in the VQA context.

Some of the predominant temporal artifacts in modern video delivery systems involve stalling, freezing, and skipping. When a video stalls, freezes, or skips, viewers may not perceive temporary reductions of spatial details or artifacts. Although we tested freezing videos on the VQEGHD5 VQA database, it would also be interesting to explore temporal flicker masking and video quality impacts caused by abrupt stalls, freezes, and skips.

Understanding change patterns in spectral signatures arising from multiple distortions could also be useful when predicting distortion specific or generalized distortion visibility on videos, and might lead to the development of better video scene statistics models and no-reference VQA algorithms. We believe that perceptual temporal flicker masking as a form of temporal visual masking will play an increasingly important role in modern models of objective VQA.

Acknowledgments

This work was supported by Intel and Cisco Corporations under the VAWN program and by the National Science Foundation under Grants IIS-0917175 and IIS-1116656.

References

[1] Cisco Corporation, Cisco Visual Networking index: Global mobile data traffic forecast update, 2015-2020. [Online]. Available: http://www.cisco.com/c/dam/m/en_in/innovation/enterprise/assets/mobile-white-paper-c11-520862.pdf.
[2] L.K. Choi, Y. Liao, A.C. Bovik, Video QoE metrics for the compute continuum, IEEE Commun. Soc. Multimed. Tech. Comm. (MMTC) E-Lett. 8 (5) (2013) 26–29.
[3] M. Yuen, H. Wu, A survey of hybrid MC/DPCM/DCT video coding distortions, Signal Process. 70 (3) (1998) 247–278.
[4] C. Chen, L.K. Choi, G. de Veciana, C. Caramanis, R.W. Heath Jr, A.C. Bovik, Modeling the time-varying subjective quality of http video streams with rate adaptations, IEEE Trans. Image Process. 23 (5) (2014) 2206–2221.
[5] X. Fan, W. Gao, Y. Lu, D. Zhao, Flickering reduction in all intra frame coding, in: Proc. JVT-E070, Joint Video Team of ISO/IEC MPEG & ITU-T VCEG Meeting, 2002.
[6] A.M. Tekalp, Digital Video Processing, Prentice-Hall PTR, Upper Saddle River, NJ, 1995.
[7] S. Daly, N. Xu, J. Crenshaw, V. Zunjarrao, A psychophysical study exploring judder using fundamental signals and complex imagery, in: Proc. SMPTE Annual Technical Conference & Exhibition, vol. 2014, 2014, pp. 1–14.
[8] A.C. Bovik, Automatic prediction of perceptual image and video quality, Proc. IEEE 101 (9) (2013) 2008–2024.
[9] J. Mannos, D. Sakrison, The effects of a visual fidelity criterion of the encoding of images, IEEE Trans. Inform. Theory 20 (4) (1974) 525–536.
[10] S.J. Daly, Visible differences predictor: An algorithm for the assessment of image fidelity, in: Proc. SPIE Human Vis. Visual Process. and Digital Display III, 1992, pp.2–15.

[11] J. Lubin, D. Fibush, Sarnoff JND vision model, in: T1A1.5 Working Group Document #97-612, ANSI T1 Standards Committee, 1997.
[12] C.J. van den Branden Lambrecht, O. Verscheure, Perceptual quality measure using a spatiotemporal model of the human visual system, Proc. SPIE 2668 (1) (1996) 450–461.
[13] S. Winkler, Perceptual distortion metric for digital color video, Proc. SPIE 3644 (1) (1999) 175–184.
[14] A.B. Watson, J. Hu, J.F. McGowan III, Digital video quality metric based on human vision, J. Electron. Imaging 10 (1) (2001) 20–29.
[15] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[16] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: Proc. IEEE Asilomar Cof. Sig., Syst. Comput., vol. 2, 2003, pp. 1398–1402.
[17] K. Seshadrinath, A.C. Bovik, A structural similarity metric for video based on motion models, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, ICASSP, 2007, pp. 869–872.
[18] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.
[19] D.M. Chandler, S.S. Hemami, VSNR: A wavelet-based visual signal-to-noise ratio for natural images, IEEE Trans. Image Process. 16 (9) (2007) 2284–2298.
[20] M.H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Trans. Broadcast. 10 (3) (2004) 312–322.
[21] M. Masry, S.S. Hemami, Y. Sermadevi, A scalable wavelet-based video distortion metric and applications, IEEE Trans. Circuits Syst. Video Technol. 16 (2) (2006) 260–273.
[22] A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, Considering temporal variations of spatial visual distortions in video quality assessment, IEEE J. Sel. Top. Signal Process. 3 (2) (2009) 253–265.
[23] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, A. Kaup, Temporal trajectory aware video quality measure, IEEE J. Sel. Top. Signal Process. 3 (2) (2009) 266–279.
[24] K. Seshadrinathan, A.C. Bovik, Motion-tuned spatio-temporal quality assessment of natural videos, IEEE Trans. Image Process. 19 (2) (2010) 335–350.
[25] P.V. Vu, C.T. Vu, D.M. Chandler, A spatiotemporal most apparent distortion model for video quality assessment, in: Proc. IEEE Int. Conf. Image Process., ICIP, 2011, pp. 2505–2508.
[26] R. Soundararajan, A.C. Bovik, Video quality assessment by reduced reference spatio-temporal entropic differencing, IEEE Trans. Circuits Syst. Video Technol. 23 (4) (2013) 684–694.
[27] M.A. Saad, A.C. Bovik, Blind prediction of natural video quality, IEEE Trans. Image Process. 23 (3) (2014) 1352–1365.
[28] M.H. Pinson, L.K. Choi, A.C. Bovik, Temporal video quality model accounting for variable frame delay distortions, IEEE Trans. Broadcast. 60 (4) (2014) 637–649.
[29] M.N. Garcia, D. Dytko, A. Raake, Quality impact due to initial loading, stalling, and video bitrate in progressive download video services, in: Proc. 6th Int. Workshop Quality of Multimedia Experience, QoMEX, 2014, pp. 129–134.
[30] D. Ghadiyaram, J. Pan, A.C. Bovik, A subjective and objective study of stalling events in mobile streaming videos, IEEE Trans. Circuits Syst. Video Technol. (2017). http://dx.doi.org/10.1109/TCSVT.2017.2768542.
[31] D. Ghadiyaram, J. Pan, A.C. Bovik, Learning a continuous-time streaming video QoE model, IEEE Trans. Image Process. 27 (5) (2018) 2257–2271.
[32] B. Breitmeyer, H. Ogmen, Visual Masking: Time Slices Through Conscious and Unconscious Vision, Oxford University Press, New York, NY, USA, 2006.
[33] G.E. Legge, J.M. Foley, Contrast masking in human vision, J. Opt. Soc. Amer. 70 (12) (1980) 1458–1470.
[34] D.J. Simons, R.A. Rensink, Change blindness: Past, present, and future, Trends Cogn. Sci. 9 (1) (2005) 16–20.
[35] D.M. Levi, Crowding—an essential bottleneck for object recognition: A mini-review, Vis. Res. 48 (2008) 635–654.
[36] G. Sperling, Temporal and spatial visual masking. I. Masking by impulse flashes, J. Opt. Soc. Amer. A 55 (5) (1965) 541–559.
[37] B.E. Rogowitz, Spatial/temporal interactions: Backward and forward metacontrast masking with sine-wave gratings, Vis. Res. 23 (10) (1983) 1057–1073.
[38] J.T. Enns, V. Di Lollo, What's new in visual masking?, Trends Cogn. Sci. 4 (9) (2005) 345–352.
[39] F. Hermens, G. Luksys, W. Gerstner, M. Herzog, U. Ernst, Modeling spatial and temporal aspects of visual backward masking, Psychol. Rev. 225 (1) (2008) 83–100.
[40] J.W. Suchow, G.A. Alvarez, Motion silences awareness of visual change, Curr. Biol. 21 (2) (2011) 140–143.
[41] L.K. Choi, A.C. Bovik, L.K. Cormack, Spatiotemporal flicker detector model of motion silencing, Perception 43 (12) (2014) 1286–1302.
[42] L.K. Choi, A.C. Bovik, L.K. Cormack, The effect of eccentricity and spatiotemporal energy on motion silencing, J. Vis. 16 (5) (2016) 1–13.
[43] A.J. Seyler, Z. Budrikis, Detail perception after scene changes in television image presentations, IEEE Trans. Inform. Theory 11 (1) (1965) 31–43.
[44] A.N. Netravali, B. Prasada, Adaptive quantization of picture signals using spatial masking, Proc. IEEE 65 (4) (1977) 536–548.
[45] B.G. Haskell, F.W. Mounts, J.C. Candy, Interframe coding of videotelephone pictures, Proc. IEEE 60 (1972) 792–800.
[46] A. Puri, R. Aravind, Motion-compensated video with adaptive perceptual quantization, IEEE Trans. Circuits Syst. Video Technol. 1 (1991) 351–378.
[47] B. Girod, The information theoretical significance of spatial and temporal masking in video signals, in: Proc. SPIE Human Vis. Visual Process. and Digital Display, 1989, pp.178–187.
[48] J.D. Johnston, S.C. Knauer, K.N. Matthews, A.N. Netravali, E.D. Petajan, R.J. Safranek, P.H. Westerink, Adaptive Non-Linear Quantizer, 1992, U.S. Patent 5,136,377.
[49] C.H. Chou, C.W. Chen, A perceptually optimized 3-D subband codec for video communication over wireless channels, IEEE Trans. Circuits Syst. Video Technol. 6 (2) (1996) 143–156.
[50] Z. Chen, C. Guillemot, Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model, IEEE Trans. Circuits Syst. Video Technol. 20 (6) (2010) 806–819.
[51] J.G. Daugman, Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, J. Opt. Soc. Amer. A 2 (7) (1985) 1160–1169.
[52] A.C. Bovik, M. Clark, W.S. Geisler, Multichannel texture analysis using localized spatial filters, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1) (1990) 55–73.
[53] E.H. Adelson, J.R. Bergen, Spatiotemporal energy models for the perception of motion, J. Opt. Soc. Amer. A 2 (2) (1985) 284–299.
[54] D.J. Heeger, Normalization of cell responses in cat striate cortex, Visual Neurosci. 9 (2) (1992) 181–197.
[55] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, L.K. Cormack, Study of subjective and objective quality assessment of video, IEEE Trans. Image Process. 19 (6) (2010) 1427–1441.
[56] F. Zhang, S. Li, L. Ma, Y.C. Wong, K.N. Ngan, IVP subjective quality video database, (2011) [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/.
[57] F.D. Simone, M. Nacari, M. Tagliasacchi, F. Dufaux, S. Tubaro, T. Ebrahimi, Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel, in: Proc. 1st Int. Workshop Quality of Multimedia Experience, QoMEX, 2009, pp. 204–209.
[58] Video Quality Experts Group, VQEG, Report on the validation of video quality models for high definition video content, in: Tech. Rep., 2010, [Online]. Available: https://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx.
[59] Z. Wang, A.C. Bovik, Reduced- and no-reference image quality assessment, IEEE Signal Process. Mag. 28 (6) (2011) 29–40.
[60] P.C. Teo, D.J. Heeger, Perceptual image distortion, in: Proc. SPIE Human Vision, Visual Process., and Digital Display V, vol. 2179, 1994, pp. 127–141.
[61] E. Gelasca, T. Ebrahimi, On evaluating video object segmentation quality: A perceptually driven objective metric, IEEE J. Sel. Top. Signal Process. 3 (2) (2009) 319–335.
[62] L.K. Choi, L.K. Cormack, A.C. Bovik, On the visibility of flicker distortions in naturalistic videos, in: Proc. 5th Int. Workshop Quality of Multimedia Experience, QoMEX, 2013. pp. 164–169.
[63] L.K. Choi, L.K. Cormack, A.C. Bovik, Eccentricity effect of motion silencing on naturalistic videos, in: Proc. IEEE 3rd Global Conf. Sig. and Inf. Process., GlobalSIP, 2015, pp. 1190–1194.
[64] L.K. Choi, L.K. Cormack, A.C. Bovik, Motion silencing of flicker distortions on naturalistic videos, Signal Process., Image Commun. 39 (2015) 328–341.
[65] L.K. Choi, A.C. Bovik, Perceptual flicker visibility prediction model, in: Proc. IS&T Human Vision and Electronic Imaging, HVEI, 2016, pp. 108:1–6.
[66] R. Blake, R. Sekuler, Perception, fifth ed., McGraw-Hill, New York, NY, USA, 2006.
[67] M. Carandini, J.B. Demb, V. Mante, D.J. Tolhurst, Y. Dan, B.A. Olshausen, J.L. Gallant, N.C. Rust, Do we know what the early visual system does?, J. Neurosci. 25 (46) (2005) 10577–10597.
[68] E.P. Simoncelli, D.J. Heeger, A model of neuronal responses in visual area MT, Vis. Res. 38 (5) (1998) 743–761.
[69] A.B. Watson, A.J. Ahumada, Model of human visual-motion sensing, J. Opt. Soc. Amer. A 2 (2) (1985) 322–342.
[70] D. Fleet, A. Jepson, Computation of component image velocity from local phase information, Int. J. Comput. Vis. 5 (1) (1990) 77–104.
[71] S. Shady, D.I.A. MacLeod, H.S. Fisher, Adaptation from invisible flicker, Proc. Natl. Acad. Sci. USA 101 (14) (2004) 5170–5173.
[72] R.W. Bowen, J. Pola, L. Matin, Visual persistence: Effects of flash luminance, duration and energy, Vis. Res. 14 (4) (1974) 295–303.
[73] J.R. Andersion, M. Matessa, A production system theory of serial memory, Psychol. Rev. 104 (4) (1997) 728–748.
[74] K. Seshadrinath, A.C. Bovik, Temporal hysteresis model of time varying subjective video quality, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, ICASSP, 2011, pp. 1153–1156.
[75] Z. Wang, L. Lu, A.C. Bovik, Video quality assessment based on structural distortion measurement, Signal Process., Image Commun. 19 (2) (2004) 121–132.
[76] M.A. Masry, S.S. Hemami, A metric for continuous quality evaluation of compressed video with severe distortions, Signal Process., Image Commun. 19 (2004) 133–146.
[77] K.T. Tan, M. Ghanbari, D.E. Pearson, An objective measurement tool for MPEG video quality, Signal Process. 70 (1998) 279–294.
[78] A.K. Moorthy, A.C. Bovik, Visual importance pooling for image quality assessment, IEEE J. Sel. Top. Signal Process. 3 (2) (2009) 193–201.

[79] M. Narwaria, W. Lin, A. Liu, Low-complexity video quality assessment using temporal quality variations, IEEE Trans. Multimed. 14 (3) (2012) 525–535.
[80] H.G. Longbotham, A.C. Bovik, Theory of order statistic filters and their relationship to linear FIR filters, IEEE Trans. Acoust. Speech Signal Process. 37 (2) (1989) 275–287.
[81] CDVL [Online]. Available: http://www.cdvl.org/.
[82] M. Gaubatz, Metrix Mux Visual Quality Assessment Package, [Online] Available: https://github.com/sattarab/image-quality-tools/tree/master/metrix_mux.
[83] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans. Image Process. 15 (11) (2006) 3440–3451.
[84] Final VQEG Report on the Validation of Objective Models of Video Quality Assessment, The Video Quality Experts Group, 2003, [Online]. Available: https://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-ii/frtv-phase-ii.aspx.
[85] A.K. Moorthy, L.K. Choi, A.C. Bovik, G. de Veciana, Video quality assessment on mobile devices: Subjective, behavioral and objective studies, IEEE J. Sel. Top. Signal Process. 6 (6) (2012) 652–671.
[86] D.C. Howell, Statistical Methods for Psychology, Wadsworth, Belmont, CA, 2007.
[87] A. Yadav, S. Sohoni, D. Chandler, GPGPU based implementation of a high performing No Reference, NR-IQA algorithm, BLIINDS-II, in: Proc. IS&T Image Quality and System Performance XIV, 2017, pp. 21–25.