FLICKER SENSITIVE MOTION TUNED VIDEO QUALITY ASSESSMENT

Lark Kwon Choi and Alan C. Bovik

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

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

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

I. INTRODUCTION

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

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

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

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

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

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

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

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

978-1-4673-9919-7/16/$31.00 ©2016 IEEE

II. Flicker Sensitive Motion Tuned Framework

A. Gabor Decomposition

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

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

where $\mathbf{x} = (x, y, t)$ is a spatiotemporal coordinate in a video sequence, $\mathbf{U}_0 = (U_0, V_0, W_0)$ is the space-time center frequency of the Gabor filter, and Σ is the covariance matrix of the Gaussian envelope.

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

B. Cortical Neuron Model

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

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

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

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

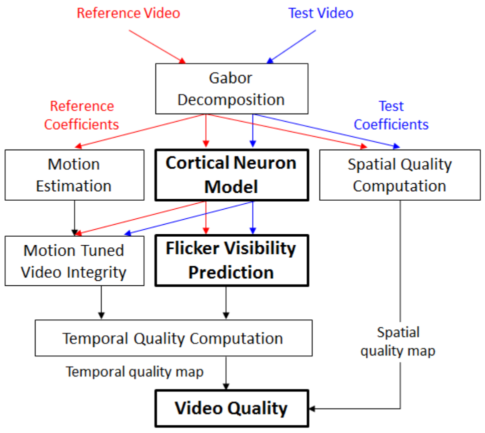


Figure 1. Block diagram of the proposed flicker sensitive motion tuned framework for video quality assessment.

$$S(\varphi,\theta) = K \frac{E(\varphi,\theta)}{\sum_{\varphi,\theta} E(\varphi,\theta) + \sigma^2}, \tag{3}$$

$$C(\varphi,\theta) = \sum_m c_m S_m(\varphi,\theta), \tag{4}$$

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

C. Spatial Video Quality

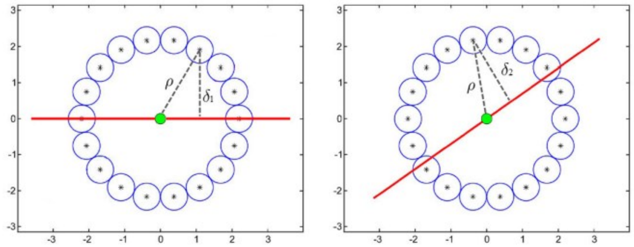
Spatial video quality is predicted using the Gabor responses from the reference and test videos as in the Spatial MOVIE framework [5], but we apply a wider range of frequency subbands than does Spatial MOVIE. We measure spatial errors from each subband Gabor response and the DC subband Gaussian filter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

$$Q_S(\mathbf{x}) = 1 - \frac{\frac{1}{M}\sum_{\varphi,\theta} Err_S(\mathbf{x},\varphi,\theta) + Err_{\mathrm{DC}}(\mathbf{x})}{M+1}, \tag{5}$$

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

D. Temporal Video Quality

To measure perceptual temporal video quality, flicker visibility is predicted and combined with the responses of the Temporal MOVIE framework. We first compute a weighted sum of the Gabor filter outputs using the Temporal MOVIE framework, where excitatory-inhibitory weights are assigned to each individual Gabor filter as a function of its distance



(a) (b)

Figure 2. Motion tuned spectral planes relative to a slice through the Gabor filter bank at one scale: (a) in a static region and (b) in a moving region. The horizontal axis is spatial frequency, while the vertical axis denotes temporal frequency. The red solid line indicates a spectral plane, while blue small circles represent Gabor filter pass bands.

from the motion-tuned spectral plane of a reference video as shown in Figure 2 [5]. Any misaligned spectrum from the spectral plane is penalized using inhibitory weights. Then the motion-tuned computed errors of a distorted video relative to the reference video serves to predict temporal video integrity as follows:

$$Q_{\text{Motion}}(\mathbf{x}) = 1 - \sum_{n=1}^{N} \gamma_n (v_n^r - v_n^d)^2, \qquad (6)$$

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

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

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

$$Q_{\text{Flicker}}(\mathbf{x}) = 1 - \frac{1}{M} \sum_{\varphi=1,\theta=1}^{M} \frac{\left| C^r(\varphi,\theta,\mathbf{x}) - C^d(\varphi,\theta,\mathbf{x}) \right|}{K}. \qquad (7)$$

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

$$Q_T(\mathbf{x}) = Q_{\text{Motion}}(\mathbf{x}) \times Q_{\text{Flicker}}(\mathbf{x}). \qquad (8)$$

We obtain a single score for each frame by using the CoV of Q_T. The overall temporal quality is then defined by averaging the frame level of CoV values. Since the range of temporal video quality scores is smaller than that of spatial video quality scores, due to the divisive normalization, we used the square root of the temporal scores, similar to [5].

E. *Video Quality*

Finally, flicker sensitive motion tuned video quality is achieved using a simple spatiotemporal pooling method. We product the spatial and temporal quality indices as follows;

$$\text{Video Quality} = \text{Spatial Quality} \times \text{Temporal Quality}. \qquad (9)$$

III. Performance Evaluation

To evaluate the performance of the proposed framework, we compute correlation coefficients between the predicted video quality scores and the human subjective quality scores on the LIVE VQA database [25]. Then, we compared the framework performance with recent objective VQA model performances: Peak Signal-to-Noise Ratio (PSNR), Multiscale SSIM (MS-SSIM) [26], Visual Signal-to-Noise Ratio (VSNR) [27], Video Quality Metric (VQM) [28], VQM-Variable Frame Delay (VQM-VFD) [29], MOVIE [5], Spatiotemporal Most Apparent Distortion (ST-MAD) [30], and Spatiotemporal Reduced Reference Entropic Difference (STRRED) [6] are compared. PSNR, MS-SSIM, and VSNR were applied on a frame-by-frame basis and the average score across all frames were used as a final quality score. Note that STRRED is a reduced reference VQA model, while the other methods are full reference algorithms.

Tables I-II show the performance of all tested algorithms in terms of the Spearman Rank Order Correlation Coefficient (SROCC) and the Pearson Linear Correlation Coefficient (PLCC) after nonlinear regression in [31], respectively, for each distortion type and for the entire LIVE VQA database. The bold font indicates the top performing model for each column. PSNR provides a baseline of comparison of VQA models. It is clear from the results that although MOVIE outperforms PSNR, MS-SSIM, VSNR, VQM, and VQM-VFD, the proposed flicker sensitive motion tuned framework further improves the performance of MOVIE by predicting perceptual flicker visibility, while accounting for temporal visual masking. The superior performance of the flicker sensitive temporal quality model highlights the perceptual importance of temporal masking of flicker distortions.

The proposed flicker sensitive motion tuned VQA framework quite correlates well with human judgments of video quality on the LIVE VQA database, achieving highly competitive results against state-of-the-art VQA models. The proposed model generally shows stable performance across specific distortion types achieving the best results on the entire LIVE VQA database, although ST-MAD performed better on H.264 compression and MPEG 2 compression artifacts and VQM-VFD achieved better performance on transmission distortions over IP networks among the tested algorithms.

TABLE I. SPEARMAN RANK ORDER CORRELATION COEFFICIENT

	Wireless	IP	H.264	MPEG-2	All
PSNR	0.6574	0.4167	0.4585	0.3862	0.5398
MS-SSIM	0.7289	0.6534	0.7313	0.6684	0.7364
VSNR	0.7019	0.6894	0.6460	0.5915	0.6755
VQM	0.7214	0.6383	0.6520	0.7810	0.7026
VQM-VFD	0.7510	**0.7922**	0.6525	0.6361	0.7354
ST-MAD	0.8060	0.7686	**0.9043**	**0.8478**	0.8242
STRRED	0.7857	0.7722	0.8193	0.7193	0.8007
Spatial MOVIE	0.7927	0.7046	0.7066	0.6911	0.7270
Temporal MOVIE	**0.8114**	0.7192	0.7797	0.8170	0.8055
MOVIE	0.8109	0.7157	0.7664	0.7733	0.7890
Proposed spatial quality	0.7940	0.6930	0.7720	0.6909	0.7416
Proposed temporal quality	0.7826	0.7895	0.8276	0.8059	**0.8304**
Proposed video quality	0.7949	0.7513	0.8265	0.7671	0.8061

TABLE II. PEARSON LINEAR CORRELATION COEFFICIENT

	Wireless	IP	H.264	MPEG-2	All
PSNR	0.6695	0.4689	0.5330	0.3986	0.5604
MS-SSIM	0.7157	0.7267	0.7020	0.6640	0.7379
VSNR	0.6992	0.7341	0.6216	0.5980	0.6896
VQM	0.7325	0.6480	0.6459	0.7860	0.7236
VQM-VFD	0.8144	**0.8616**	0.7403	0.7172	0.7763
ST-MAD	0.8123	0.7900	**0.9097**	**0.8422**	0.8299
STRRED	0.8039	0.8020	0.8228	0.7467	0.8062
Spatial MOVIE	0.7883	0.7378	0.7252	0.6587	0.7451
Temporal MOVIE	0.8371	0.7383	0.7920	0.8252	0.8217
MOVIE	0.8386	0.7622	0.7902	0.7595	0.8116
Proposed spatial quality	0.8092	0.7301	0.8135	0.7220	0.7670
Proposed temporal quality	0.8233	0.8197	0.8436	0.8346	**0.8480**
Proposed video quality	**0.8533**	0.8193	0.8624	0.7973	0.8278

IV. CONCLUSION AND FUTURE WORK

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

REFERENCES

[1] Cisco Corporation, Cisco Visual Networking index: Global mobile data traffic forecast update, 2014-2019.

[2] L. K. Choi, Y. Liao, and A. C. Bovik, "Video QoE metrics for the compute continuum," *IEEE Commun. Soc. Multimed. Tech. Comm. (MMTC) E-Lett.*, vol. 8, no. 5, pp. 26-29, 2013.

[3] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, Sep. 2013.

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

[5] K. Seshadrinathan and A. C. Bovik, "Motion-tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335-350, Feb. 2010.

[6] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol*., vol. 23, pp. 684-694, Apr. 2013.

[7] M. A. Saad and A. C. Bovik, "Blind prediction of natural video quality," *IEEE Trans. Image Process*., vol. 23, no. 3, pp. 1352-1365, Mar. 2014.

[8] A. J. Seyler and Z. Budrikis, "Detail perception after scene changes in television image presentations," *IEEE Trans. Inf. Theory*, vol.11, no.1, pp.31-43, Jan. 1965.

[9] A. N. Netravali and B. Prasada, "Adaptive quantization of picture signals using spatial masking," *Proc. IEEE*, vol. 65, no. 4, pp. 536-548, Apr. 1977.

[10] A. Puri and R. Aravind, "Motion-compensated video with adaptive perceptual quantization," *IEEE Trans. Circuits Syst. Video Technol*., vol. 1, pp. 351-378, Dec. 1991.

[11] J. W. Suchow and G. A. Alvarez, "Motion silences awareness of visual change," *Curr. Biol.*, vol. 21, no. 2, pp.140-143, Jan. 2011.

[12] M. Turi and D. Burr, "The motion silencing illusion results from global motion and crowding," *J. Vis.*, vol. 13, no. 5, Apr. 2013.

[13] L. K. Choi, A. C. Bovik, and L. K. Cormack, "Spatiotemporal flicker detector model of motion silencing," *Perception*, vol. 43, no. 12, pp. 1286-1302, Dec. 2014.

[14] L. K. Choi, L. K. Cormack, and A. C. Bovik, "On the visibility of flicker distortions in naturalistic videos," in *Proc. IEEE Int. Workshop Qual. Multimedia Exper.,* Jul. 2013, pp. 164-169.

[15] L. K. Choi, L. K. Cormack, and A. C. Bovik, "Motion silencing of flicker distortions on naturalistic videos," *Signal Process. Image Commun*., vol. 39, pp. 328-341, Mar. 2015.

[16] L. K. Choi, L. K. Cormack, and A. C. Bovik, "Eccentricity effect of motion silencing on naturalistic videos," in *Proc. IEEE 3rd Global Conf. Sig. and Inf. Process.*, Dec. 2015.

[17] K. Seshadrinath and A. C. Bovik, "A structural similarity metric for video based on motion models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*., Apr. 2007, pp. 869-872.

[18] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT," *Vis. Res*., vol. 38, no. 5, pp. 743-761, Mar. 1998.

[19] 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*, vol. 2, pp. 1160-1169, 1985.

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

[21] D. Fleet and A. Jepson, "Computation of component image velocity from local phase information," *Int. J. Comput. Vis.*, vol. 5, no. 1, pp. 77-104, 1990.

[22] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Amer. A*, vol. 2, no. 2, pp. 284-299, Feb. 1985.

[23] D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Vis. Neurosci*., vol. 9, no. 2, pp. 181-197, Aug. 1992.

[24] L. K. Choi and A. C. Bovik, "Perceptual flicker visibility prediction model," in *Proc. Human Vision and Electronic Imaging*, Feb. 2016.

[25] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process*., vol.19, no.6, pp.1427-1441, Jun. 2010.

[26] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Cof. Sig., Syst. Comput.*, Nov. 2003, vol. 2, pp. 1398-1402.

[27] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual singal-to-noise ratio for natural images," *IEEE Trans. Image Process*., vol.16, no.9, pp.2284-2298, Sep. 2007.

[28] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast*., vol. 10, no. 3, pp. 312-322, Sep. 2004.

[29] M. H. Pinson, L. K. Choi, and A. C. Bovik, "Temporal video quality model accounting for variable frame delay distortions," *IEEE Trans. Broadcast*., vol. 60, no. 4, pp. 637-649, Dec. 2014.

[30] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most apparent distortion model for video quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2505-2508.

[31] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440-3451, 2006.

[32] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Sel. Topics Signal Process*., vol. 6, no. 6, pp. 652-671, Oct. 2012.