Objective Video Quality Assessment Based on Perceptually Weighted Mean Squared Error

Sudeng Hu, Lina Jin, Hanli Wang, *Senior Member, IEEE*, Yun Zhang, *Senior Member, IEEE*, Sam Kwong, *Fellow, IEEE*, and C.-C. Jay Kuo, *Fellow, IEEE*

***Abstract*— Object quality assessment for compressed video is critical to various video compression systems that are essential in the video delivery and storage. Although mean squared error (MSE) is computationally simple, it may not be accurate to reflect the perceptual quality of compressed videos, which are also affected dramatically by the characteristics of the human visual system (HVS), such as contrast sensitivity, visual attention, and masking effect. In this paper, a video quality metric is proposed based on perceptually weighted MSE. A low-pass filter is designed to model the contrast sensitivity of the HVS with the consideration of visual attention. The imperceptible distortion is adaptively removed in the salient and nonsalient regions. To quantitatively measure the masking effect, the randomness of video content is proposed in both the spatial and temporal domains. Since the masking effect highly depends on the regularity of structure and motion in the spatial and temporal directions, the video signal is modeled as a linear dynamic system, and the prediction error of future frames from previous frames is used as randomness to measure the significance of masking. The relation is investigated between MSE and perceptual quality scores across various contents, and a masking modulation model is proposed to compensate the impact of the masking effect on the MSE. The performance of the proposed quality metric is validated on three video databases with various compression distortions. The experimental results demonstrate that the proposed algorithm outperforms other benchmark quality metrics.**

***Index Terms*— Human visual system (HVS), low-pass filter, masking effect, video quality assessment, visual attention.**

Manuscript received March 27, 2015; revised August 15, 2015, November 14, 2015, and January 24, 2016; accepted April 7, 2016. Date of publication April 20, 2016; date of current version September 5, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61472281 and Grant 61471348; in part by the Shu Guang Project of Shanghai Municipal Education Commission through the Shanghai Education Development Foundation under Grant 12SG23; in part by the Program for Professor of Special Appointment (Eastern Scholar) within the Shanghai Institutions of Higher Learning under Grant GZ2015005; in part by the Shenzhen Overseas High-Caliber Personnel Innovation and Entrepreneurship Project under Grant KQCX20140520154115027; and in part by the Guangdong Special Support Program for Youth Science and Technology Innovation Talents under Grant 2014TQ01X345. This paper was recommended by Associate Editor P. Le Callet.

S. Hu, L. Jin, and C.-C. J. Kuo are with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: sudenghu@gmail.com; linajin.j@gmail.com; cckuo@sipi.usc.edu).

H. Wang is with the Key Laboratory of Embedded System and Service Computing, Department of Computer Science and Technology, Ministry of Education, Tongji University, Shanghai 200092, China (e-mail: hanliwang@tongji.edu.cn).

Y. Zhang is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: yun.zhang@siat.ac.cn).

S. Kwong is with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: cssamk@cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2016.2556499

I. Introduction

With the development of video technologies, video has become one of the most important electronic media in our daily lives. Original videos take a very large space and need to be compressed before transmission or storage, but the compression might degrade the video quality. Since humans are the final receivers of videos, in the sense of quality of experience, it is highly desired to precisely predict a human's perception on compressed videos. Due to the large time and human resource consumption of the subjective video quality assessment, great efforts have been dedicated to developing various objective video quality metrics.

A number of video quality metrics have been designed to simulate the characteristics of the human visual system (HVS). Contrast sensitivity is one of the most important properties of the HVS, which varies to different spatial and temporal frequencies, and has been psychophysically studied and modeled in the contrast sensitivity function (CSF) [1]–[7]. Video quality metrics employ the CSF to analyze the visibility of impairs [8], [9]. In [8], the video is preprocessed with separable filters in the temporal and spatial domains. A low-pass and a bandpass filter are used for temporal filtering, whereas spatial filtering is implemented in the discrete wavelet transform (DWT) domain. However, as stated in [3] and [10], separate processing temporal and spatial frequency is not possible. In [9], distortion is decoupled into detail losses and additive impairments with DWT, and the sensitivity of the distortion is analyzed through a comprehensive spatial–temporal CSF, and the weighting factors are calculated to adjust the distortion according to the sensitivity at different DWT frequencies. In these CSF models, the contrast sensitivity is modeled only as a function of frequency, without taking the visual attention into consideration.

Actually, the contrast sensitivity is not uniformly distributed over the video content. Instead, it peaks at the gazed region and decreases away from it. While static images might give viewers enough time to watch the details in different regions, videos release tremendous information within a very short time, which makes the HVS unable to receive all of it. Therefore, visual attention plays an important role in quality assessment and has been a concern in recent studies [11]–[15]. In [11], the difference of wavelet coefficients between an undistorted image and its distorted version is weighted with the foveation error sensitivity, according to the visual attention. In [12], a video presentation is transferred from its original Cartesian coordinate to the curvilinear coordinate by a foveation filtering operation and then the distortion is calculated with

weighted signal-to-noise ratio. In [13], various quality metrics are modified by weighing the original metrics with a saliency map derived from the eye-tracking data of visual attention, and improvements in performance were observed compared with the metrics without visual attention. An overview of applying visual attention in quality assessment is given in [14]. In these methods, it simply gives greater weights to the distortion in the attended areas at the pooling stage, and the weight is usually designed intuitively. Therefore, it is difficult to justify and develop a proper and accurate weighting scheme that can work the same way as the HVS in balancing the attended and unattended distortions. In [15], in addition to spatial and temporal CSF properties, visual attention is considered when the critical frequency is modeled. Then the critical frequency is integrated with the wavelet-based distortion visibility model.

Another important characteristic to consider in video quality is the masking effect, which refers to a human's reduced ability to detect a stimulus on a spatially or temporally complex background. The traditional way to measure the masking effect is by using a divisive gain control method, which decomposes the video into multiple channels and analyzes the masking effect among the channels by divisive gain normalization [16], [17]. However, the mechanism of gain control mostly remains unknown. In addition, since only a simple masker, such as sinusoidal gratings or white noise, is used in the experiments to search for optimal parameters to fit the gain control model, there is no guarantee that these models are applicable to natural images [18]. In [19] and [20], it is pointed out that the masking effect highly depends on the level of randomness created by the background. Usually the regular background contains predictable content and the stimulus will become distinct from the neighborhood when it is different from a human's expectation of its position. In the random background, the content is unpredictable, and thus any change on it will be less noticed. Therefore, there is higher masking in the random background than the regular background. In [19], the concept of entropy masking is proposed to measure the masking effect of the background using zero order entropy. However, it measures masking only in the spatial domain for videos, which is obviously inadequate, because the temporal activities will also affect the visibility of distortion significantly. Usually distortion is highly masked in the massive and random motions, while less masked in regular and smooth motions. In [21], the mismatch between two consecutive frames is used to measure temporal activities. However, it may not reflect the regularity of motion precisely, since smooth and regular motion can also produce a large mismatch. Therefore, it is desired to develop the method that could measure the regularity of motion and thus measure the masking effect of videos.

On other hand, although the mean squared error (MSE) has been criticized for the low correlation to the HVS due to its low computational cost, it is still widely used in practice. The inaccuracy of the MSE in perceptual quality prediction comes from the lack of psychophysical designs in HVS, like counting the imperceptible distortions. In this paper, we revise the MSE by incorporating important HVS characteristics. First, to remove the imperceptible distortion from the MSE, a low-pass filter is designed based on the CSF and visual attention. Since the contrast sensitivity is affected both by frequency and visual attention, visual saliency is introduced to adjust the cutoff frequency in the CSF so that the developed low-pass filter could adaptively remove the imperceptible distortion according to the location that is attended or not. In this way, the problem of nonuniform sampling of visual acquisition is solved naturally by removing less high frequency distortion in salient regions and more in nonsalient regions. In addition, the masking modulation is applied afterward to reduce the imperceptible distortion covered by masking. Because smooth and regular motions will hide less distortion than massive and irregular motions, we first propose a method to measure the randomness of video with a dynamic model. Since video content is easier to predict with regular motion than random motion, the prediction error actually reflects the randomness of video and can be used as the measurement of randomness to indicate how much the background could mask the noise. Furthermore, we investigate the model of masking modulation, which quantitatively analyzes how the modified MSE should be compensated according to the proposed randomness. The analysis is performed based on the relation between the modified MSE and the perceptual quality scores across different video contents.

The rest of this paper is organized as follows. In Section II, the foveated low-pass filter is proposed. The masking modulation model is introduced in Section III. In Section IV, the experimental results are given to compare the performance of the proposed video quality metric with other benchmarks. Finally, Section V concludes this paper.

II. Foveated Low-Pass Filter

The initial visual signal processing in HVS includes two steps. In the first step, the visual signal goes through the eyes optics, forming an image on the retina. Because of the diffraction and other imperfections in the eye, such processing would blur the passed image. In the second step, the image will be filtered by neural filters as it is received by photoreceptor cells on the retina and then passed on to the lateral geniculate nucleus and the primary visual cortex. These processes are more like low-pass filtering and will hide considerable high-frequency information from perception.

A. Low-Pass Filtering With Spatiotemporal CSF

The CSF, which is defined as the inverse of the contrast threshold of detectable contrast at a given frequency, provides a comprehensive measure of vision. Although it is not exactly equivalent to the modulation transfer function (MTF), it reflects the same trend as the modulation gain. For instance, higher sensitivity at particular frequencies always means higher modulation gain at the corresponding frequencies and vice versa. Therefore, many researchers have treated the CSF as MTF, and used it to define characteristics of initial processing in HVS [22]–[24]. Constrast sensitivity of HVS peaks at certain spatial and temporal frequency and drops sharply after that along both spatial and temporal frequencies. The traditional CSF model from [3] modified in [1] considers the contrast sensitivity as a function of both the spatial and

temporal frequencies, which can be expressed as

$$\mathrm{CSF}(\omega, v_r) = c_0(k_1 + k_2|\log(\varepsilon \cdot v/3)|^3) \cdot v_r \cdot \omega^2 \cdot \exp(-c_1 \cdot \omega \cdot (\varepsilon \cdot v_r + 2)/k_3) \quad (1)$$

where $\omega/2\pi$ is the spatial frequency in cycles per degree, and v_r is the retinal image velocity, implicitly denoting the temporal frequency. k_1, k_2, and k_3 are empirical constants set as 6.1, 7.3, and 23 in [3]. c_0 and c_1 are used to control the magnitude and the bandwidth of a CSF curve. Note that such a model is developed for near-threshold distortion; to simplify the problem, we assume it also applies to suprathreshold distortion.

According to [1] and [25], object velocity jointly with eye movement determines temporal frequency, i.e., retinal velocity v_r. There are three types of eye movements: smooth-pursuit eye movement, natural drift eye movement, and saccadic eye movement. The exact eye movement is affected by the moving objects and the subject's ability to track them under some psychological constraints [26], [27]. In this paper, the eye movement estimation in [1], [6], and is used to calculate retinal velocity from object velocity that is measured by optical flow and viewing distance.

The processed visual signal after passing through the initial part of HVS can be modeled as

$$I' = F^{-1}(\mathrm{CSF}(\omega)) * I \quad (2)$$

where I' and I are the processed and original visual signal, respectively; F^{-1} is the inverse Fourier transform; and $*$ is the convolution operation.

B. Foveated Low-Pass Filter

Our gaze is mainly driven to follow the most salient regions, and the distortions that occur outside the salient areas are assumed to have a lower impact on the overall quality. This is because the photoreceptor cells are not equally distributed, but they are dense in the fovea and sparse on the peripheral retina. Therefore, the gazed regions on an image have better visual resolution in the HVS, and consequently it is less blurred, whereas the regions outside foveation will lose many more details. Since the contrast sensitivity changes with the location of the image projected onto the retina, the filter should be adaptively changed rather than using one that is constant.

In [28], the contrast threshold is modeled based on the spatial frequency of the visual signal and its retinal eccentricity to the fixation. Since the contrast sensitivity is the inverse of the contrast threshold, the corresponding CSF can be expressed as

$$\mathrm{CSF}(\omega, e) = \frac{1}{\mathrm{CT}_0} \cdot \exp\left(-\mu \cdot \omega \cdot \frac{e + e_2}{e_2}\right), \quad f > 0 \quad (3)$$

where $\omega/2\pi$ is the spatial frequency, e is the retinal eccentricity, CT_0 is a constant presenting the minimum contrast threshold, e_2 is the half-resolution eccentricity, and μ is the spatial frequency decay constant. The retinal eccentricity e is the angle between the fixation and the location of the signal, and it is related to the distance between the two points and the viewing distance. Compared with (1), (3) does not consider the temporal factor and approximate the spatial properties in a monotonically decreasing curve. However, by modulating the spatial frequency with retinal eccentricity, it includes the consideration of unequal distribution of sensitivity over the whole retina.

To develop a comprehensive CSF that considers both spatiotemporal frequencies and the foveated vision mechanism, the temporal factor model in (1) is integrated into (3), and the foveated CSF is developed as

$$\mathrm{CSF}(\omega, v_r, e) = c_0(k_1 + k_2|\log(\varepsilon \cdot v_r/3)|^3) \cdot v_r \cdot \exp(-c_1 \cdot \omega \cdot (e + e_2)/e_2 \cdot (\varepsilon \cdot v_r + 2)/k_3). \quad (4)$$

In this model, contrast sensitivity monotonically decreases with spatial frequencies. Note that as the luminance of videos change over time, our visual function, such as the pupil, will adapt to accommodate these changes [29]–[32]. Since the video clips in our experiments are only 10 s or less than 10 s, to simplify the problem, we assume that the luminances of videos are maintained over the whole sequence. By transforming the CSF in (4) into the spatial domain as proved in the Appendix, we have the impulse response of the initial processing system in the HVS as

$$h(d_F, e, v_r) = \frac{1}{\pi} \cdot \frac{a \cdot b}{a^2 + d_F^2} \quad (5)$$

where a and b relate to retinal velocity and retinal eccentricity, $b = c_0(k_1 + k_2|\log(\varepsilon \cdot v_r/3)|^3) \cdot v_r$ and $a = c_1 \cdot (e + e_2)/e_2 \cdot (\varepsilon \cdot v_r + 2)/k_3$; d_F is the distance from the filter center, i.e., $d_F = (x^2 + y^2)^{1/2}$. Since the parameters a and b are changing with the retinal velocity and retinal eccentricity, the filter in (5) becomes adaptive to their factors.

C. Computational Model of Eccentricity

Since the visual acuity varies on the different location of a video, the accurate prediction of visual attention is critical. Recording eye movements is so far the most reliable means for studying human visual attention, and it provides the ground truth of the fixation locations on videos. It is highly desirable to incorporate this information into the developed foveated low-pass filter. However, recording such data requires extra equipment like eye-tracking devices, and the experiments are expensive and time consuming. More importantly, since humans are involved in the process, it is impossible to develop it into objective quality metrics whereby each component should be automatic. An alternative way is using saliency detection algorithms. In general, saliency is defined as what attracts human perceptual attention. Computational visual attention models trying to predict the gaze location of humans with features from images or videos can be generally classified into two categories: a bottom-up approach [33]–[37] and a top-down approach [38]. The actual mechanisms of visual attention are much more complicated and involve many factors. Usually top-down approaches use both the low-level and high-level features. For example, high-level features can be faces, people, and text, whereas low-level features could be color, edge, etc. Top-down approaches highlight the

importance of high-level and semantic features, but they may not be general enough to include all situations; for example, they often fail to detect salient objects for which they have not been trained. As stated in [39], both stimulus features and task demands affect visual attention. However, videos in our problem are used for general purpose and not for specific tasks. Thus, without clear task demands, the high-level features may not be helpful in detecting saliency. In addition, bottom-up approaches usually consume less time than top-down. We adopt [34] in this paper to estimate saliency map.

The saliency map quantifies the possibility of the locations being the gazed. A location with a large value in the saliency map is more likely to be gazed, and hence the eccentricity of that location projected on the retina will be small, and vice versa. Therefore, the retinal eccentricity of a location increases as its visual saliency value decreases. In [13], the saliency value is assumed to be Gaussian distributed around the fixation as $s = \exp(-d_E^2/\sigma^2)$, where d_E is the distance from fixation, and σ is the model parameter. Since our saliency map is generated by computational saliency models and the actual distribution depends on the employed computational saliency models, instead of using Gaussian distribution, we apply a more general distribution as

$$s = \exp\left(-\frac{d_E^{\theta}}{\sigma^2}\right) \tag{6}$$

where θ is the model parameter depending on different saliency detection algorithms, and in our experiments, $\theta = 4$. The location with the maximum saliency value is assumed to be the gaze location and based on (6), it is straightforward to use the visual saliency value to approximate the retinal eccentricity as

$$\begin{aligned} e(i,j) &= \arctan\frac{(-\sigma^2 \ln(s(i,j)))^{\vartheta}}{L} \\ &\approx \gamma \cdot \ln(1/s(i,j))^{\vartheta} \end{aligned} \tag{7}$$

where $s(i,j)$ is the visual saliency value at position (i,j) and L is the viewing distance. $\gamma = \sigma^{2\vartheta}/L$, $\vartheta = 1/\theta$. The values of $s(i,j)$ within each frame are normalized into the range of [0, 1].

D. Blockwise Filtering

Since the contrast sensitivity is different in positions, the low-pass filtering that simulates the initial processing of the HVS can be applied with adaptive filters based on (3) and (7). For the constant filters, it is equivalent to apply filtering in the frequency domain or spatial domain. However, since the proposed low-pass filter changes spatially, the spatial information will be lost in the Fourier frequency domain; it can be implemented only in the spatial domain as

$$\Delta I_f = h(e, v_r) * (I_d - I_o) = h(e, v_r) * \Delta I \tag{8}$$

where I_d and I_o are distorted and original frames, respectively. $h(e, v_r)$ is the low-pass filter in (5). Equation (8) is computationally heavy, since for each pixel we have to generate a new filter according to the corresponding saliency values and retinal velocities. Usually the saliency map is continuous and smooth, and thus we can assume that the saliency value within a neighborhood is similar. Low-pass filtering can be processed block by block with block size $N \times N$ and a larger N can reduce the computational complexity but with coarser eccentricity estimation, while a smaller N can provide finer estimation but with higher computational cost. In our experiments, block size is set to 32×32 for a good balance between accuracy and computational complexity. For the kth block B_k, the average eccentricity of the block

$$\bar{e}_k = \frac{1}{N^2} \sum_{(m,n)\in B_k}^{N} e(m,n) \tag{9}$$

is used to present to visual attention. Similarly, the average retinal velocity of the block is used for the entire block as

$$\bar{v}_{rk} = \frac{1}{N^2} \sum_{(m,n)\in B_k}^{N} v_r(m,n) \tag{10}$$

where $v_r(m,n)$ can be estimated by optical flow and viewing distance. Thus, a constant filter is applied within a block as

$$\Delta I_f(i,j) = h(\bar{e}_k, \bar{v}_{rk}) * \Delta I(i,j) \tag{11}$$

where $(i,j) \in B_k$.

The visual illustration of foveated low-pass filtering is shown in Fig. 1. We can see that in Fig. 1(b), the high frequency signals are equally removed across the content, even in the regions that we are interested in. However, in Fig. 1(d), they are removed adaptively according to the saliency map shown in Fig. 1(c), and high frequencies remain in the salient regions.

After the adaptive low-pass filtering, MSE_f is calculated as the mean of the sum of the squared difference between the original and compressed video sequences as

$$\mathrm{MSE_f} = \frac{1}{WHL} \sum_{t=1}^{L} \sum_{i=1,j=1}^{WH} \Delta I'_f(i,j,t)^2 \tag{12}$$

$$D = \ln(\mathrm{MSE_f}) \tag{13}$$

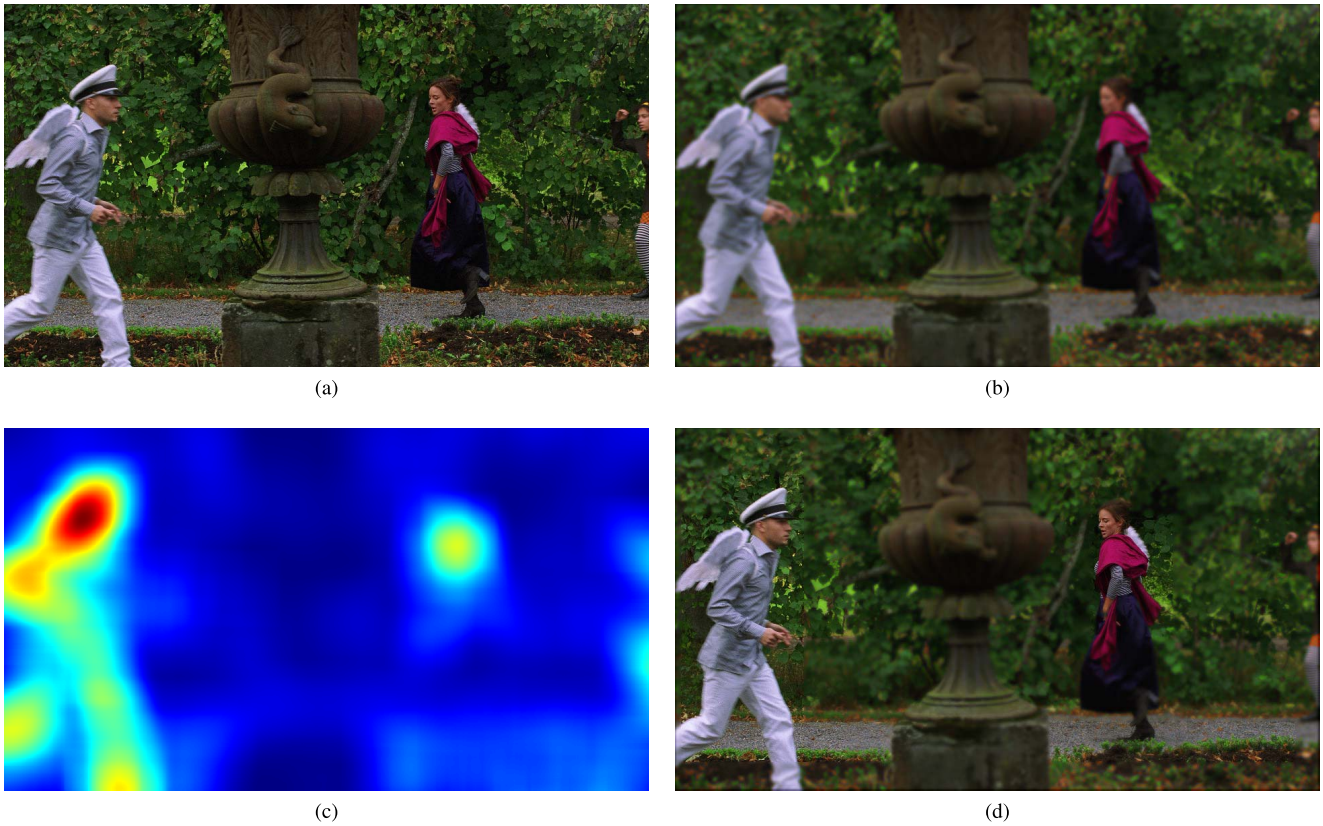
where W, H, and L are the width, height, and duration of the video sequences. Here MSE is analyzed in logarithmic scale as presented in (13) because in the logarithmic scale, the difference of quality curves among different contents is more obvious and clearer than in the linear scale.

III. Perceptual Modulation

The visibility of distortion highly depends on the content of the background. Usually a strong masking effect can prevent the distortion from being observed and thus reduce the distortion perceptually. Therefore, it is important to measure the masking effect. In [19], it is pointed out that the masking effect highly depends on the level of randomness created by the background. For videos, randomness should be measured in both spatial and temporal domains.

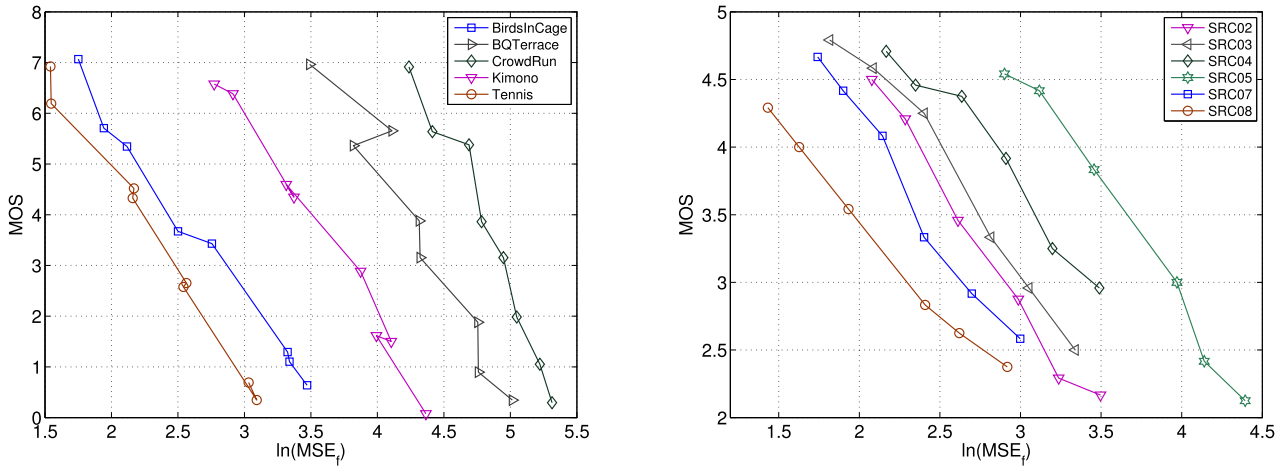
A. Displacement of Metric Curves

The relationship between the mean opinion score (MOS) and D in (13) is shown in Fig. 2 for various sequences from different databases. Each point corresponds to a distorted video



(a) (b) (c) (d)

Fig. 1. Visual illustration of foveated low-pass filtering. (a) Original image. (b) Filtered with constant low-pass filter. (c) Saliency map. (d) Filtered with foveated low-pass filter.



(a) (b)

Fig. 2. Relation of MOS and $\ln(\mathrm{MSE_f})$ for different video sequences. (a) On the MCLV database [49]. (b) On the VQEG database [50].

sequence, and metric curves are formed by connecting the points that share the same original video. In other words, the connected points in Fig. 2 are video sequences compressed from the same original sequence but with different compression levels. Under the same video content, D is a good predictor of perceptual quality (i.e., MOS), since the MOS monotonically decreases with D.

However, such a relation cannot be applied to distorted videos with different contents. As we can observe in Fig. 2, there are different horizontal displacements for the metric curves of different video contents. Such a difference in horizontal displacement mainly comes from the different masking effect of various video contents. Given the same MOS, the points of metric curves on the right side have more actual distortion, i.e., $\mathrm{MSE_f}$, than on left, as shown in Fig. 2, which means the video in the right metric curve has more masking, which produces the same perceptual quality as the videos on the left side. Therefore, the videos with a strong masking effect

TABLE I
SLOPES AND GOODNESS OF FITTING

	SeqName	BB	BC	BQ	CR	DK	EA	EB	FB	KM	OT	SK	TN	Ave
MCLV	Slope	-3.268	-2.177	-3.176	-4.097	-3.213	-3.765	-5.192	-3.341	-3.079	-2.912	-3.456	-2.992	-3.388
	R^2	0.934	0.983	0.981	0.983	0.989	0.981	0.991	0.986	0.982	0.988	0.957	0.981	0.978
	SeqName	SRC01	SRC02	SRC03	SRC04	SRC05	SRC06	SRC07	SRC08	SRC09				
VQEG	Slope	-1.002	-1.412	-1.321	-1.331	-1.492	-2.201	-1.489	-1.012	-1.423				-1.409
	R^2	0.983	0.986	0.986	0.982	0.991	0.944	0.980	0.981	0.972				0.978

are more likely to have metric curves on the right side, and the displacement of these curves with respect to the left side reflects the significance of the masking effect.

To quantitatively analyze the masking effect, we assume that the shapes of the curves in Fig. 2 are identical by neglecting the small differences among them. The points of the same contents are fitted with linear curves, and the slopes of different curves are presented in Table I as well as the goodness of fit R^2. We can see that within each database, the slopes of most video sequences are close to each other, which means that the shapes of these curves are almost the same. R^2 describes how well the linear model fits to the actual data and the closer to 1 its value is, the better the mode is. Although the values of R^2 in Table I are all so close to 1, which means that the linear model is accurate, it is not necessary to limit the model to linear. Instead, as long as the shape of these curves are the same, we can generalize the relation of D and MOS as

$$\widehat{\text{MOS}} = F(D - P) \tag{14}$$

where P is the horizontal displacement depending on the video content, and $F(\cdot)$ can be a linear function or other monotonic decreasing function representing the shape of these curves. P reflects the masking effect of the video content. A strong masking effect always results in large P values. Since $F(\cdot)$ is fixed in (14), an accurate estimation of P is critical to the MOS prediction. Due to the difference of the masking effect, P varies significantly from sequence to sequence.

B. Temporal and Spatial Randomness

To measure the masking effect of video content, the regularity of video content is analyzed quantitatively in both spatial and temporal domains. As an important characteristic of the video, motion information is highly related to masking activities. Usually distortion is highly masked in the massive and random motions, while less masked in regular and smooth motions.

For regular motion, the future frames can be predicted from the past frames by learning the temporal behavior of a short video clip in the past. Thus, the prediction error reflects the randomness of motion. To capture the temporal activities of the past video, the video sequence can be modeled as a discrete-time dynamic system [40]. To simplify the problem, the video signal is modeled as a linear dynamic system as in [41]. Let $Y_k^l = [y(k), \ldots, y(l)] \in \mathbb{R}^{m\times(l-k)}$ denote a short sequence from the kth frame to the lth frame, and each frame is rearranged into a column vector $y \in \mathbb{R}^m$, where m equals the number of pixels within a frame, i.e., $m = W \times H$. The motion in the video is simulated as the evolution process of a dynamic system, described as

$$\begin{cases} Y_k^l = CX_k^l + W_k^l \\ X_k^l = AX_{k-1}^{l-1} + V_k^l \end{cases} \tag{15}$$

where $X_k^l = [x(k), \ldots, x(l)]$ and $X_{k-1}^{l-1} = [x(k-1), \ldots, x(l-1)] \in \mathbb{R}^{n\times(l-k)}$ are the state sequences of Y_k^l and Y_{k-1}^{l-1}, respectively, and $m > n$. $A \in \mathbb{R}^{n\times n}$ is the state transition matrix that encodes the regular motion information, and $V_k^l \in \mathbb{R}^{n\times(l-k)}$ is the sequence of motion noise that cannot be represented by the regular information A. $C \in \mathbb{R}^{m\times n}$ is the observation matrix encoding the shapes of objects within the frames, and $V_k^l \in \mathbb{R}^{n\times(l-k)}$ is the sequence of observation noise that cannot be represented by the regular shape information C. Given the video sequence Y_k^l, the model parameters A, C and the state sequence X_k^l are not unique. There are infinite choices of these matrices that can give exactly the same video sequence Y_k^l. An efficient method was proposed in [42], which employs a singular value decomposition and keeps the n largest singular values as

$$Y_k^l = U\Sigma V^T + W_k^l \tag{16}$$

where $\Sigma = \text{diag}[\sigma_1, \ldots, \sigma_n]$ contains the n largest singular values and $U \in \mathbb{R}^{m\times n}$, $V \in \mathbb{R}^{(l-k)\times n}$ are corresponding decomposition vectors. By setting $X_k^l = \Sigma V^T$ and $C(l) = U$, we can determine the state sequence and the model parameter C. Since the redundancy in Y_k^l is removed by reducing the dimension from m to n, X_k^l is the compact representation of Y_k^l with a loss of information W_k^l.

Moreover, A is expected to capture the motion information and thus predict future frames. The optimal A can be found by minimizing the squared prediction error as

$$\hat{A}(l) = \underset{A}{\text{argmin}} \left|\left| X_{k+1}^l - AX_k^{l-1} \right|\right|. \tag{17}$$

Therefore, the optimal solution can be obtained as

$$\hat{A}(l) = X_{k+1}^l X_k^{l-1+} \tag{18}$$

where X_k^{l-1+} is the pseudoinverse of X_k^{l-1}. We can predict the future frame $y(l+1)$ based on the obtained model parameters, i.e., $A(l)$, $C(l)$ that characterize the temporal activities of sequence Y_k^l. The prediction error can be calculated as

$$R_T(l+1) = |y(l+1) - C(l)A(l)x(l)| \tag{19}$$

where $R_T(l+1) \in \mathbb{R}^m$ is the noise that can not be predicted with regular information. This value reveals the predictability

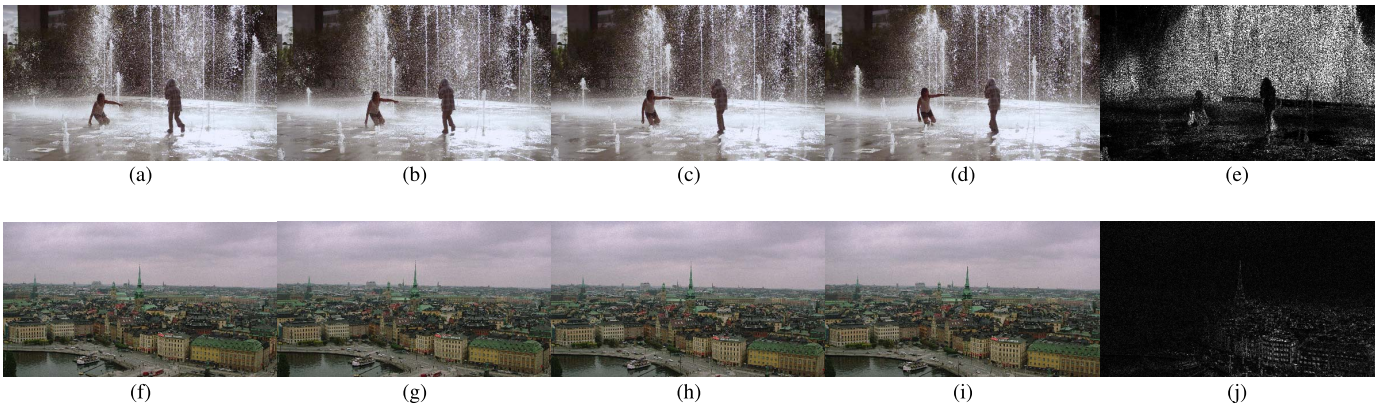


Fig. 3. Visual illustration of temporal randomness on two different video sequences. (a)-(d) Consecutive frames of the sequence ElFuente2. (e) Temporal randomness for the sequence ElFuente2. (f)-(i) Consecutive frames of the sequence OldTownCross. (j) Temporal randomness for the sequence OldTownCross.

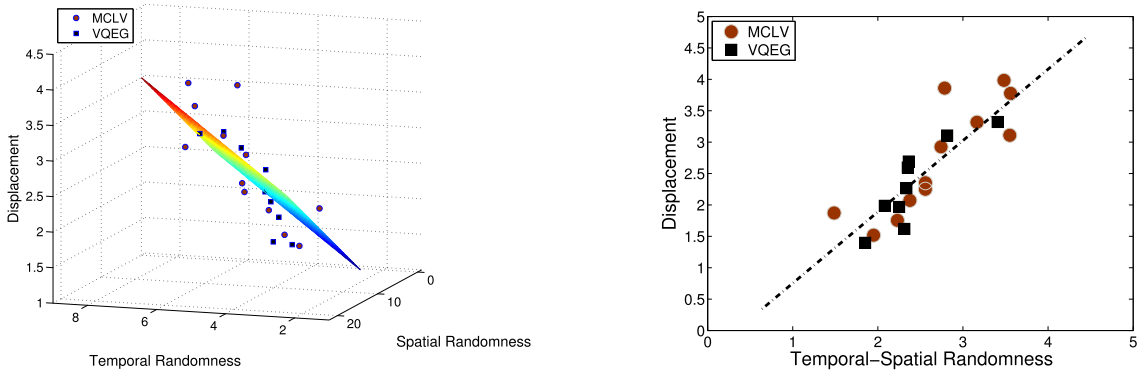


Fig. 4. (a) Relation between horizontal displacement P and temporal randomness and spatial complexity. (b) Combined temporal and spatial randomness.

of the next frame according to the trajectory of moving objects in the past frames and thus reflect its temporal randomness. Usually smooth and regular motions in videos will make future frames more predictable than massive and random motions. Fig. 3 shows the temporal randomness for two sequences. Fig. 3(a)–(d) and (f)–(i) shows the frames of the sequence ElFuente2 and OldTownCross, respectively, and Fig. 3(e) and (j) shows the corresponding temporal randomness calculated from (19). In the background of the sequence ElFuente2, the motion of water drops is unpredictable, and thus its temporal randomness is large. While in the sequence OldTownCross, the motion is smooth and regular. Consequently, its temporal randomness is much smaller than that of the sequence ElFuente2. Finally, the average temporal randomness is used to represent the overall temporal randomness of the whole video as

$$\bar{R}_T = \frac{1}{m \cdot L} \sum_{l=1}^{L} \sum_{i=1}^{m} R_T^i(l) \tag{20}$$

where $R_T^i(l) \in \mathbb{R}^m$ is the ith component of $R_T(l)$, and L is the total number of frames. The number of previous frames $(k-l)$ will affect the prediction error. For smooth motion, usually larger number of previous frames will result in a smaller prediction error, and for nonlinear motion (in higher order), a smaller number will give a lower prediction error. In this approach, we assume that humans use a fixed duration of their past experience to predict future movement. In our experiment, we set t to 1 s, and in case the frame rate is 30 frames/s, the number of frames to predict the future frame is 30.

Besides the temporal domain, the spatial activities of the frame also affect the masking effect. The pixel variance of $N \times N$ block is computed to indicate the local spatial randomness, and the logarithm of the mean of the local spatial randomness is utilized as spatial randomness of the whole video as

$$\bar{R}_S = \ln\left(\frac{1}{M \cdot L} \sum_{t=1}^{L} \sum_{i=1}^{B} \sigma^2(i,t)\right) \tag{21}$$

where $\sigma^2(i)$ is the variance of the ith $N \times N$ block in the tth frame; B and L are the total number of blocks within a frame and total number of frame within a sequence.

C. Modulation

As discussed, the displacement of metric curves in (14) reflects the masking effect, and it relates to the temporal and spatial activities of the video sequences. To investigate its relation to temporal randomness $\bar{R}_T$ and spatial randomness $\bar{R}_S$, we have to measure the actual horizontal displacement first. The displacement can be determined by measuring the horizontal position of the crossing points of the metric curves with any horizontal lines such as MOS $= 3.0$. The relation of the actual displacement P with the temporal randomness $\bar{R}_T$ and the spatial randomness $\bar{R}_S$ is shown in Fig. 4. In Fig. 4(a), each point represents a video sequence from either

the database MCLV or the database VQEG, we can see that the displacement has a linear relation with $\bar{R}_T$ and $\bar{R}_S$, respectively. Thus, it could be approximated with a linear surface, and the displacement can be predicted as

$$\hat{P}_i = \alpha \bar{R}_T + \beta \bar{R}_S \tag{22}$$

where α and β are model parameters and fixed at 0.315 and 0.372, respectively. Fig. 4(b) shows the relation between the actual and the predicted displacement. Combining the (13), (14), and (22), we have

$$\begin{aligned} \widehat{\text{MOS}} &= F(\ln(\text{MSE}_f) - \alpha \bar{R}_T - \beta \bar{R}_S) \\ &= G(\text{MSE}_f \cdot e^{-(\alpha \bar{R}_T + \beta \bar{R}_S)}) \end{aligned} \tag{23}$$

where $G(\cdot) = F(\ln(\cdot))$. It is acceptable for a quality metric to predict MOS through a nonlinear mapping, because the mapping is easy to be found, and it depends on various environmental factors, such as the range of MOS and evaluation methodology. Therefore, in [43] and [44], a nonlinear mapping is not considered as part of VQM, rather it is left to the final stage of the performance evaluation. $G(\cdot)$ can be obtained by fitting the objective prediction scores to the subjective quality scores, as described in [43] and [44]. We use the perceptually weighted distortion

$$\text{MD} = \text{MSE}_\text{f} \cdot e^{-(\alpha \bar{R}_T + \beta \bar{R}_S)} \tag{24}$$

as the MOS predictor. In this way, the MSE is modified according to the HVS characteristics and thus become more correlated with the perceptual quality.

D. Context Effect

The MOS of a video is not only determined mainly by its perceptual quality but it is also affected by the perceptual quality of other videos during subjective tests. For example, when a video with medium quality is evaluated in a pool of severely impaired videos, it will get a higher MOS than when it is evaluated in a pool of high quality videos. Such phenomenon is called context effect. Although various subjective tests are designed carefully to reduce such an effect, it cannot be removed completely in subjective tests [45], [46]. Usually the quality of former displayed videos will affect MOS of latter videos, but since the display order of the videos are random for each subject, it is reasonable to assume that each video has an equal chance to be affected by other videos in subjective tests. Assuming that the MOS of a video would be equally affected by other videos, a slight shift in MOS might be caused with the general perceptual quality of the context, which is expressed as

$$\text{MOS} = Q - \eta \cdot \bar{Q} \tag{25}$$

where $\bar{Q}$ is the average perceptual quality of all videos displayed in subjective tests, and η is a penalty coefficient reflecting how much other videos would affect the quality of the current video. For example, $\eta = 0$ means that MOS is not affected by the quality of other videos. So far, such a shift in MOS does not affect the performance evaluation of the quality assessment.

However, in the actual subjective test, the MOS of a particular video may receive a different impact from different videos. The MOS of a video is more likely to be affected by videos with similar contents and distortion types. In other words, when the subjects provide quality scores, they intend to compare the quality of the current video with previous similar videos with similar distortion types. The resultant quality score will be affected by these videos more than others. In this paper, we focus on the same distortion types, i.e., compression distortion, and thus only the content is considered. To measure the similarity of videos, besides the temporal randomness in (19) and spatial randomness measured in (21), the color information is also extracted because color plays an important role in quality assessment, as described in [47] and [48]. Therefore, the color feature for each frame is extracted as

$$cv = \det \begin{pmatrix} \sigma_Y^2 & \sigma_{YU}^2 & \sigma_{YV}^2 \\ \sigma_{YU}^2 & \sigma_U^2 & \sigma_{UV}^2 \\ \sigma_{YV}^2 & \sigma_{UV}^2 & \sigma_V^2 \end{pmatrix} \tag{26}$$

where σ_Y^2, σ_U^2, and σ_V^2 are the variance of Y, U, and V components in the YCbCr color space, respectively; σ_{YU}^2, σ_{YV}^2, and σ_{UV}^2 are the covariances of three components, respectively. The mean value $\bar{cv}$ along the temporal domain is used for each sequence. Therefore, we measure the distance between the ith and the jth videos in the feature space as

$$d(i,j) = \frac{\kappa_1 |\bar{cv}_i - \bar{cv}_j|}{\bar{cv}_i + \bar{cv}_j} + \frac{\kappa_2 |\bar{R}_{Ti} - \bar{R}_{Tj}|}{\bar{R}_{Ti} + \bar{R}_{Tj}} + \frac{\kappa_3 |\bar{R}_{Si} - \bar{R}_{Sj}|}{\bar{R}_{Si} + \bar{R}_{Sj}} \tag{27}$$

where κ_1–κ_3 are constant model parameters indicating the importance of the features, and they are set to 1 in our experiments. The videos with smaller distance $d(i,j)$ will affect the MOS of each other more than the videos with larger distance.

To simulate the impact of other video quality on the MOS while taking the content distance into consideration, we modify the quality metric in (24) and propose the perceptually weighted MSE as

$$\text{PW-MSE}(i) = \text{MD}(i) - \eta \left(\frac{1}{\Delta_i} \sum_{j \in V, j \neq i} e^{-d(i,j)} \cdot \text{MD}(j) \right) \tag{28}$$

where $e^{-d(i,j)}$ is the weighting factor, and $\Delta_i = \sum_{j \in V, j \neq i} e^{-d(i,j)}$ is used for normalization; V is the set of videos in context, and $\eta = 1$. If the content similarity among videos is identical, (28) becomes (25), and the context effect vanishes in terms of quality prediction, because a constant added to the metric will not affect the final performance.

IV. Experimental Results

A. Subjective Databases and Performance Metrics

The performance of the proposed video quality matric was evaluated in the three databases, including the MCLV [49], the VQEG [50], and the IRCCyN databases [51]. In the MCLV, there are 12 original video sequences with the resolution of 1920 × 1080. Two types of compression distortion are involved in the MCLV database. In the first type of distortion,

TABLE II
INTERMEDIATE PERFORMANCE AT EACH STAGE

	PCC			SROCC			RMSE		
	MSE	**CSF**	**PW-MSE**	**MSE**	**CSF**	**PW-MSE**	**MSE**	**CSF**	**PW-MSE**
MCLV	0.4526	0.6029	0.9723	0.4442	0.5881	0.9667	2.7975	1.7704	0.5191
VQEG	0.6907	0.7051	0.9348	0.6816	0.6983	0.9151	0.6309	0.6186	0.3098
IRCCyN	0.7960	0.8893	0.9245	0.8050	0.8807	0.9204	0.6369	0.4971	0.4144

the original sequences are compressed with H.264/AVC codec, generating four different quality levels. In the second type of distortion, the original sequences are first downscaled and compressed with H.264/AVC codec at four quality levels. Then, the compressed sequences are upscaled to the original resolution. A total of 96 distorted video sequences is in the MCLV database. In the VQEG database, the original sequences are from the VQEGHD 3 of the VQEG project, and there are nine original sequences with the resolution of 1920×1080. In the database VQEGHD 3, besides the compression distortion types, there are several other distortion types such as transmission error. Since we are interested in only compression distortion, only six distorted sequences with compression distortion were selected for each original sequence. There is a total of 54 distorted video sequences. In the IRCCyN database, there are sixty original sequences with the resolution of 640×480. The videos are encoded with H.264/AVC and the codec of scalable video coding (H.264/SVC). Each original video is encoded at four different quality levels. Thus, there is a total of 240 distorted videos.

Since some performance metrics, such as the linear correlation coefficient, require to compare linear correlation, for a fair comparison, the nonlinear mapping is carried out between the objective score and MOS. The following nonlinear function is employed before the performance evaluation for all video quality metrics:

$$q(x) = \alpha_1 \left(0.5 - \frac{1}{1 + \exp(\alpha_2(x - \alpha_3))}\right) + \alpha_4 x + \alpha_5 \quad (29)$$

where α_1 to α_5 are the parameters obtained by regression between the input and output data. As for metrics of performance evaluation, the Pearson correlation coefficient (PCC), the Spearman rank order correlation coefficient (SROCC) and root MSE (RMSE) are employed as described in [43] and [44]. PCC generally indicates the goodness of linear relation. The SROCC is computed on ranks and thus depicts the monotonic relationships, whereas the RMSE computes the prediction errors and thus depicts the prediction accuracy.

B. Performance at Two Stages

The proposed algorithm consists of two main stages to simulate the visual signal processing in the HVS. In the first stage, the foveated low-pass filtering is implemented to simulate the initial processing of the HVS. Then, the masking effect is considered to simulate high-level processing in the HVS. To verify the effectiveness of each step in the proposed algorithm, the intermediate results of each step were investigated, including the performance of the model with foveated low-pass filtering only (denoted as CSF), and complete model. The results are summarized in Table II.

As we can see in Table II, the performance under each performance evaluation method is improved at each stage under all databases. In the MCLV database, MSE does not perform well compared with other databases, achieving only around 0.45 and 0.44 in PCC and SROCC, respectively. Even after processing with the foveated low-pass filtering, the performance has not improved significantly because in the MCLV database, the video contents are quite diverse. That makes the masking effect vary dramatically among different sequences and, as a consequence, MSE becomes inconsistent over different video content. When only the masking effect is considered, performance is improved significantly compared with the foveated low-pass filtering. When both models are taken into consideration in the final stage, we can see that the performance has improved to 0.972, 0.967, and 0.519 in PCC, SROCC, and RMSE, respectively. As far as the VQEG and IRCCyN databases are concerned, MSE achieves better performances than in the MCLV database, and the performance is further improved at each step.

C. Overall Performance

In this section, we compare the performance of the proposed method with other benchmarks, including: MS-SSIM [52], VIF [53], ST-MAD [54], VQM [55], MOVIE [56]. Default settings were used for all the benchmarks, except for MOVIE.[1] Only the luminance component is used for analysis. Table III summarizes the performance of all the video quality metrics in the MCLV, the VQEG, and the IRCCyN databases, where the best performance is highlighted in boldface.

From Table III, we can see that the proposed PW-MSE achieves the best performance among all the video quality metrics and performs consistently well that it obtains PCC and SROCC above 0.9 on all the three databases.

The scatter plots of the subjective quality scores against objective quality scores are shown in Fig. 5 for the three databases. In order to plot in the same scale, the MOS was normalized and the objective scores were obtained after applying the nonlinear fitting to MOS. We can see the width of the PW-MSE's scatter plot is the narrowest among the quality metrics, which implies that it has a more direct correlation between the objective and subjective quality scores than other metrics.

[1]Due to the limited computational capability, the frame interval of MOVIE is set to 32 for the MCLV and VQEG databases, instead of the default value 8.

TABLE III
OVERALL PERFORMANCE ON VARIOUS DATABASES

	MCLV			VQEG			IRCCyN		
	PCC	**SROCC**	**RMSE**	**PCC**	**SROCC**	**RMSE**	**PCC**	**SROCC**	**RMSE**
MS-SSIM	0.681	0.663	1.625	0.854	0.855	0.454	0.917	0.911	0.433
VIF	0.518	0.511	1.898	0.704	0.663	0.619	0.837	0.832	0.595
ST-MAD	0.579	0.623	1.810	0.670	0.558	0.648	0.912	0.909	0.446
MOVIE	0.625	0.627	1.733	0.768	0.877	0.559	0.753	0.900	0.716
VQM	0.763	0.783	1.433	0.880	0.876	0.415	0.918	0.910	0.431
PW-MSE	**0.972**	**0.967**	**0.519**	**0.935**	**0.915**	**0.310**	**0.925**	**0.920**	**0.414**

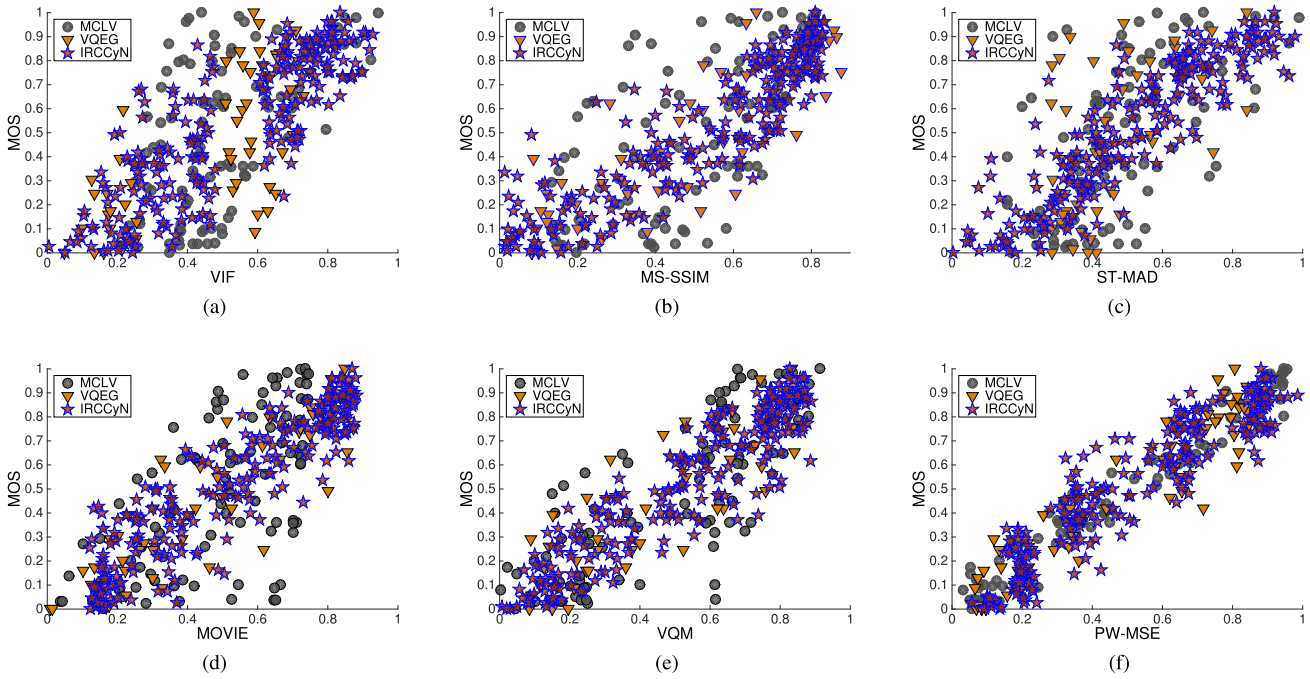


Fig. 5. Scatter plot of MOS versus predicted MOS by various quality metrics. (a) VIF. (b) MS-SSIM. (c) ST-MAD. (d) MOVIE. (e) VQM. (f) PW-MSE.

V. CONCLUSION

In this paper, PW-MSE is proposed for compressed videos. The masking effect as well as the low-passing filter characteristics of the initial process of HVS is explored. To mathematically model and simulate the initial process in HVS, the foveated CSF is adopted as the transfer function in the frequency domain. The error signal from the compression distortion is filtered with the proposed transfer function in the spatial domain, which removes most errors in high frequency that cannot be perceived by humans. Furthermore, after processing the initial part of HVS, the error signal is highly affected by various masking effects from different image contents. To study the masking effect quantitatively, the randomness is proposed to measure it by modeling the video with a dynamic system. Moreover, a modulation relation among the randomness and the distortion before masking and after masking is investigated across various video contents. By observing the relation of MOS and the distortion before masking effect, a masking modulation model is proposed based on the randomness measurement. PW-MSE is tested on databases with various compression distortions. By validating at every step, each step of the proposed PW-MSE contributes to the overall performance improvement. The performance comparison with other benchmark image quality metrics and video quality metrics demonstrates the effectiveness of PW-MSE.

APPENDIX
INVERSE FOURIER TRANSFORM OF CSF

To simplify the notation, CSF in (4) is expressed as

$$\mathrm{CSF}(\omega) = b \cdot \exp(-a \cdot \omega) \quad \omega \geq 0 \tag{30}$$

where $b = c_0(k_1+k_2|\log(\varepsilon \cdot v_r/3|^3 \cdot v_r$ and $a = c_1 \cdot (e+e_2)/e_2 \cdot (\varepsilon \cdot v_r + 2)/k_3$. Equation (30) defines only $\omega \geq 0$, and the negative axis is not defined. If we assume the filter in spatial domain is a real even function, CSF should be symmetric along the y-axis in the frequency domain as

$$\mathrm{CSF}(\omega) = b \cdot \exp(-a \cdot |\omega|). \tag{31}$$

By applying the inverse Fourier transform to (31), we can have the filter in the spatial domain

$$\begin{aligned} f(d_F) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathrm{CSF}(\omega) e^{j\omega d_F} d\omega \\ &= \frac{b}{2\pi} \int_{-\infty}^{+\infty} (e^{-a\omega} \cdot u(\omega) + e^{a\omega} \cdot u(-\omega)) e^{j\omega d_F} d\omega \\ &= \frac{1}{\pi} \frac{ab}{a^2 + d_F^2}. \end{aligned} \tag{32}$$

REFERENCES

[1] S. J. Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," *Proc. SPIE*, vol. 3299, pp. 180–191, Jul. 1998.

[2] J. M. Foley and G. M. Boynton, "New model of human luminance pattern vision mechanisms: Analysis of the effects of pattern orientation, spatial phase and temporal frequency," *Proc. SPIE*, vol. 2054, pp. 32–42, Mar. 1994.

[3] D. H. Kelly, "Motion and vision. II. Stabilized spatio-temporal threshold surface," *J. Opt. Soc. Amer. A*, vol. 69, no. 10, pp. 1340–1349, Oct. 1979.

[4] S. J. Daly, "Visible differences predictor: An algorithm for the assessment of image fidelity," *Proc. SPIE*, vol. 1666, pp. 2–15, Aug. 1992.

[5] C. A. Burbeck and D. H. Kelly, "Spatiotemporal characteristics of visual mechanisms: Excitatory-inhibitory model," *J. Opt. Soc. Amer.*, vol. 70, no. 9, pp. 1121–1126, 1980.

[6] Y. Jia, W. Lin, and A. A. Kassim, "Estimating just-noticeable distortion for video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 7, pp. 820–829, Jul. 2006.

[7] Z. Wei and K. N. Ngan, "Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 3, pp. 337–346, Mar. 2009.

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

[9] S. Li, L. Ma, and K. N. Ngan, "Full-reference video quality assessment by decoupling detail losses and additive impairments," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 7, pp. 1100–1112, Jul. 2012.

[10] F. L. Van Nes, J. J. Koenderink, H. Nas, and M. A. Bouman, "Spatiotemporal modulation transfer in the human eye," *J. Opt. Soc. Amer.*, vol. 57, no. 9, pp. 1082–1088, Sep. 1967. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=josa-57-9-1082

[11] Z. Wang, A. C. Bovik, L. Lu, and J. L. Kouloheris, "Foveated wavelet image quality index," *Proc. SPIE*, vol. 4472, pp. 42–52, Dec. 2001.

[12] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video quality assessment," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 129–132, Mar. 2002.

[13] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: Based on eye-tracking data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 7, pp. 971–982, Jul. 2011.

[14] U. Engelke, H. Kaprykowsky, H. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 50–59, Nov. 2011.

[15] J. You, T. Ebrahimi, and A. Perkis, "Attention driven foveated video quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 200–213, Jan. 2014.

[16] V. Laparra, J. Muñoz-Marí, and J. Malo, "Divisive normalization image quality metric revisited," *J. Opt. Soc. Amer. A*, vol. 27, no. 4, pp. 852–864, 2010.

[17] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," *J. Opt. Soc. Amer. A*, vol. 14, no. 9, pp. 2379–2391, Sep. 1997.

[18] D. M. Chandler and S. S. Hemami, "Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions," *J. Opt. Soc. Amer. A*, vol. 20, no. 7, pp. 1164–1180, Jul. 2003.

[19] A. B. Watson, R. Borthwick, and M. Taylor, "Image quality and entropy masking," *Proc. SPIE*, vol. 3016, pp. 2–12, Jun. 1997.

[20] S. He, P. Cavanagh, and J. Intriligator, "Attentional resolution and the locus of visual awareness," *Nature*, vol. 383, pp. 334–337, Sep. 1996.

[21] L. Xu, S. Li, K. N. Ngan, and L. Ma, "Consistent visual quality control in video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 6, pp. 975–989, Jun. 2013.

[22] S. T. L. Chung, G. E. Legge, and B. S. Tjan, "Spatial-frequency characteristics of letter identification in central and peripheral vision," *Vis. Res.*, vol. 42, no. 18, pp. 2137–2152, Aug. 2002.

[23] P. G. J. Barten, "Contrast sensitivity of the human eye and its effects on image quality," *Proc. SPIE*, vol. 72, p. 232, Dec. 1999.

[24] A. B. Watson and A. J. Ahumada, Jr., "A standard model for foveal detection of spatial contrast," *J. Vis.*, vol. 5, no. 9, pp. 717–740, 2005.

[25] J. Laird, M. Rosen, J. Pelz, E. Montag, and S. Daly, "Spatio-velocity CSF as a function of retinal velocity using unstabilized stimuli," *Proc. SPIE*, vol. 6057, p. 605705, Feb. 2006.

[26] J. Intriligator and P. Cavanagh, "The spatial resolution of visual attention," *Cognit. Psychol.*, vol. 43, no. 3, pp. 171–216, 2001.

[27] P. Cavanagh and G. A. Alvarez, "Tracking multiple targets with multifocal attention," *Trends Cognit. Sci.*, vol. 9, no. 7, pp. 349–354, 2005.

[28] W. S. Geisler and J. S. Perry, "Real-time foveated multiresolution system for low-bandwidth video communication," *Proc. SPIE*, vol. 3299, pp. 294–305, Jul. 1998.

[29] F. Rieke and M. E. Rudd, "The challenges natural images pose for visual adaptation," *Neuron*, vol. 64, no. 5, pp. 605–616, 2009.

[30] P. J. Bex, S. G. Solomon, and S. C. Dakin, "Contrast sensitivity in natural scenes depends on edge as well as spatial frequency structure," *J. Vis.*, vol. 9, no. 10, p. 1, 2009.

[31] J. A. Ferwerda, S. N. Pattanaik, P. Shirley, and D. P. Greenberg, "A model of visual adaptation for realistic image synthesis," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, 1996, pp. 249–258.

[32] R. M. Boynton and G. Kandel, "On responses in the human visual system as a function of adaptation level," *J. Opt. Soc. Amer.*, vol. 47, no. 4, pp. 275–286, 1957.

[33] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[34] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[35] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1597–1604.

[36] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.

[37] A. Bugeau and P. Pérez, "Detection and segmentation of moving objects in highly dynamic scenes," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–8.

[38] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Conf. CVPR*, Sep. 2009, pp. 2106–2113.

[39] W. Einhäuser, U. Rutishauser, and C. Koch, "Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli," *J. Vis.*, vol. 8, no. 2, p. 2, 2008.

[40] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 564–577, May 2003.

[41] B. Boots, G. J. Gordon, and S. M. Siddiqi, "A constraint generation approach to learning stable linear dynamical systems," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 20. 2008, pp. 1329–1336.

[42] R. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[43] "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase I," Video Quality Experts Group (VQEG), Tech. Rep., Mar. 2000.

[44] "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II," Video Quality Experts Group (VQEG), Tech. Rep., Aug. 2003.

[45] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," *Proc. SPIE*, vol. 5150, p. 573, Jun. 2003.

[46] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, International Telecommunication Union, BT-500 Series, Rec. ITU-R BT.500-11, 2002.

[47] A. Bhat, S. Kannangara, Y. Zhao, and I. Richardson, "A full reference quality metric for compressed video based on mean squared error and video content," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 2, pp. 165–173, Feb. 2012.

[48] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.

[49] J. Y. Lin, R. Song, T. Liu, H. Wang, and C.-C. J. Kuo, "MCL-V: A streaming video quality assessment database," *J. Vis. Commun. Image Represent.*, vol. 30, pp. 1–9, Jul. 2015. [Online]. Available: http://mcl.usc.edu/mcl-v-database

[50] "Report on the validation of video quality models for high definition video content," Video Quality Experts Group, Tech. Rep., 2010.

[51] Y. Pitrey, M. Barkowsky, R. Pépion, P. Le Callet, and H. Hlavacs, "Influence of the source content and encoding configuration on the perceived quality for scalable video coding," *Proc. SPIE*, vol. 8291, p. 82911K, Feb. 2012.

[52] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, vol. 2, pp. 1398–1402.
[53] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[54] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *Proc. 18th IEEE ICIP*, Sep. 2011, pp. 2505–2508.
[55] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.
[56] 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.

Sudeng Hu received the B.Eng. degree from Zhejiang University, Hangzhou, China, in 2007; the M.Phil. degree from the Department of Computer Science, City University of Hong Kong, Hong Kong, in 2010; and the Ph.D. degree from the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA, in 2015.

Since 2016, he has been with Apple, Cupertino, CA, USA, as a Software Development Engineer. His research interests include image and video compression, scalable video coding, 3-D video coding, and image and video quality assessment.

Dr. Hu received the 2014 Chinese Government Award for Outstanding Self-Financed Students Abroad.

Lina Jin received the B.S. degree from Jilin University, Changchun, China, in 2005, and the M.Sc. and Ph.D. degrees from Tampere University of Technology (TUT), Tampere, Finland, in 2010 and 2015, respectively.

From 2009 to 2014, she was a Researcher with TUT. In 2013, she joined the Multimedia Communication Laboratory, University of Southern California, Los Angeles, CA, USA, as a Research Assistant. Her research interests include image and video quality metrics, quality of experience for multimedia, image and video compression, and image enhancement.

Hanli Wang (M'08–SM'12) received the B.S. and M.S. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from City University of Hong Kong, Hong Kong, in 2007.

From 2007 to 2008, he was a Research Fellow with the Department of Computer Science, City University of Hong Kong, and a Visiting Scholar with Stanford University, Stanford, CA, USA, invited by Prof. C. K. Chui. From 2008 to 2009, he was a Research Engineer with Precoad, Inc., Menlo Park, CA, USA. From 2009 to 2010, he was an Alexander von Humboldt Research Fellow with the University of Hagen, Hagen, Germany. In 2010, he joined the Department of Computer Science and Technology, Tongji University, Shanghai, China, as a Professor. He has authored more than 80 papers in his research fields. His research interests include digital video coding, image processing, computer vision, and machine learning.

Yun Zhang (M'12–SM'16) received the B.S. and M.S. degrees in electrical engineering from Ningbo University, Ningbo, China, in 2004 and 2007, respectively, and the Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China, in 2010.

From 2009 to 2014, he was a Post-Doctoral Researcher with the Department of Computer Science, City University of Hong Kong, Hong Kong. In 2010, he became an Assistant Professor with the Shenzhen Institute of Advanced Technology, CAS, Shenzhen, China, where he has been an Associate Professor since 2012. His current research interests include video compression, 3-D video processing, and visual perception.

Sam Kwong (F'13) received the B.S. degree in electrical engineering from State University of New York at Buffalo, Buffalo, NY, USA, in 1983; the M.S. degree in electrical engineering from University of Waterloo, Waterloo, ON, Canada, in 1985; and the Ph.D. degree from University of Hagen, Hagen, Germany, in 1996.

From 1985 to 1987, he was a Diagnostic Engineer with Control Data Canada, Mississauga, ON, Canada. He joined Bell Northern Research Canada, Ottawa, ON, Canada, as a member of the Scientific Staff. In 1990, he became a Lecturer with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, where he is currently a Professor with the Department of Computer Science. His research interests include video and image coding and evolutionary algorithms.

C.-C. Jay Kuo (F'99) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1980, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1985 and 1987, respectively.

He is the Director of the Multimedia Communications Laboratory and a Professor of Electrical Engineering, Computer Science, and Mathematics with the Ming-Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA. He has co-authored about 200 journal papers, 850 conference papers, and ten books. His research interests include digital image/video analysis and modeling, multimedia data compression, communication and networking, and biological signal/image processing.

Dr. Kuo is a fellow of the American Association for the Advancement of Science and the International Society for Optical Engineers.