Free-Energy Principle Inspired Video Quality Metric and Its Use in Video Coding

Long Xu, *Member, IEEE*, Weisi Lin, *Fellow, IEEE*, Lin Ma, *Member, IEEE*, Yongbing Zhang, *Member, IEEE*, Yuming Fang, *Member, IEEE*, King Ngi Ngan, *Fellow, IEEE*, Songnan Li, *Member, IEEE*, and Yihua Yan

***Abstract*—In this paper, we extend the free-energy principle to video quality assessment (VQA) by incorporating with the recent psychophysical study on human visual speed perception (HVSP). A novel video quality metric, namely the free-energy principle inspired video quality metric (FePVQ), is therefore developed and applied to perceptual video coding optimization. The free-energy principle suggests that the human visual system (HVS) can actively predict "orderly" information and avoid "disorderly" information for image perception. Basically, "orderly" is associated with the skeletons and edges of objects, and "disorderly" mostly concerns textures in images. Based on this principle, an image is separated into orderly and disorderly regions, and processed differently in image quality assessment. For videos, visual attention, or fixation, is associated with the objects with significant motion according to HVSP, resulting in a motion strength factor in the FePVQ so that the free-energy principle is extended into spatio-temporal domain for VQA. In addition, we investigate the application of the FePVQ in perceptual rate distortion optimization (RDO). For this purpose, the FePVQ is realized with low computational cost by using the relative total variation model and the block-wise motion vectors of video coding to simulate the free-energy principle and the HVSP, respectively. The experimental results indicate that the proposed FePVQ is highly consistent with the HVS perception. The linear correlation coefficient and Spearman's rank-order correlation coefficient are up to 0.8324 and 0.8281 on the LIVE video database. Better perceptual quality of encoded video sequences is achieved by FePVQ-motivated RDO in video coding.**

***Index Terms*—Free-energy principle, perceptual video coding optimization, video coding, visual quality assessment (VQA).**

I. Introduction

During recent years, there has been an increasing interest in visual quality assessment of image/video. Since humans are the ultimate receivers of visual signal being processed in most situations, the most accurate way of assessing image/video quality (VQ) is to ask humans for their opinions on the quality of the image/video, which is known as subjective visual quality assessment. However, subjective visual quality assessment is very time consuming, laborious and expensive. Moreover, it is infeasible to have subjects' intervention with in-loop and on-service processes. Therefore, objective quality assessment targeting at automatically predicting visual quality by computer model is in demand.

Objective quality assessment metrics can be classified into two categories: signal fidelity measures and perceptual visual quality metrics (PVQMs). The signal fidelity measures, like mean square error (MSE), signal-to-noise ratio (SNR) and peak SNR (PSNR), generally are poor predictor of perceived visual quality. PVQMs, which are designed to measure the perceived visual quality, have been intensively developed during the last decades. Some well-known PVQMs include moving pictures quality metric [1], perceptual distortion metric [2], Sarnoff just noticeable difference (JND) [3] vision model, digital video quality [4], and scalable wavelet based video distortion index [5]. In [6], You *et al.* proposed a visual perception model based on foveated vision for video quality assessment (VQA). In [7], an advanced foveal imaging model was proposed by incorporating visual attention into [6]. In [8], Lu *et al.* proposed a set of heuristic fuzzy rules that use both relative and absolute motion information to account for motion suppression and visual attention. It was shown that these rules are effective in improving VQA performance of the standard MSE/PSNR measures as well as the SSIM [9]–[11] approach. In [12], the contribution of spatial and temporal factors and their interaction were explored by machine learning [13], [14] and a low-complexity VQA using temporal quality variation is proposed.

With recent studies on brain theory and neuroscience, the free-energy principle has been found to be existing in action, perception and learning of human brain [15], [16]. It can be incorporated into existing image quality assessment (IQA) metrics with the forms of JND [17], importance weight for pooling local quality measurement, and so on. The free-energy principle claims that any self-organizing system that is at equilibrium with its environment must minimize its free energy. The essence of this principle is a sort of mathematical formulation of how adaptive systems resist a natural tendency to disorder. The free-energy principle indicates that the human visual system (HVS) can actively predict the sensory information and avoid the residual uncertainty/disorder for

Manuscript received July 6, 2015; accepted January 25, 2016. Date of publication February 3, 2016; date of current version March 15, 2016. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61232016, Grant U1405254, Grant 61572461, Grant 61202242, Grant 61571212, and Grant 11433006, in part by CAS "100-Talents" (Dr. X. Long), in part by the NSFC Guangdong Joint Fund under Grant U1201255 and Grant U1301257, and in part by the Research Grants Council of the Hong Kong SAR under Project CUHK 415712. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Klara Nahrstedt. *(Corresponding author: Lin Ma.)*

L. Xu is with the Key Laboratory of Solar Activity, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China, and also with the School of Computer and Software, Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: lxu@nao.cas.cn).

W. Lin is with the Department of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: WSlin@ntu.edu.sg).

L. Ma is with Huawei Noah's Ark Lab, Hong Kong 999077, China (e-mail: forest.linma@gmail.com).

Y. Zhang is with the Graduate School at Shenzhen, Tsinghua University, Shenzhen 518005, China (e-mail: zhang.yongbing@sz.tsinghua.edu.cn).

Y. Fang is with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang 330032, China (e-mail: fa0001ng@e.ntu.edu.sg).

K. N. Ngan and S. Li are with the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong 999077, China (e-mail: knngan@ee.cuhk.edu.hk; snli@ee.cuhk.edu.hk).

Y. Yan is with the Key Laboratory of Solar Activity, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China (e-mail: yyh@nao.cas.cn).

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/TMM.2016.2525004

image perception and understanding [17]. In other words, the HVS tries to extract as much information as possible to minimize the uncertainty of an input scene. However, it is impossible to fully process all of the sensation information and thus it has to discard some uncertainties in a visual scene. There have been a few efforts to apply this principle to visual quality assessment [17], [18]. In [17], an IQA metric was proposed inspired by the Bayesian brain theory [19] and the free-energy principle which indicates that the human brain works with an internal generative mechanism (IGM) for visual information perception and understanding. Regarding IGM, the HVS acts as an inference system that actively predicts the visual sensation and aviods the residual uncertainty/disorder. And, an auto-regression model was adopted to decompose a visual scene into two portions: predicted sensory content and residual uncertain content. Then, they were processed differently by existing IQA metrics. In [18], a new psychovisual image quality metric is developed based on free-energy principle. However, there have not yet been considerations of free-energy principle in application of VQA. Compared to image signal, video signal is depicted as a signal in a three-dimensional (3-D) field (x, y, z). To investigate the free-energy principle in video signal processing, we have to extend spatial orderly and disorderly terms into spatio-temporal domain beyond [17]. Specifically, for inter-frame predicted video coding, the motion estimation (ME) [20] and motion compensation (MC) revise the spatial signal extensively, where the highly textured regions would be disorderly and noisy after ME/MC with the repetitive pattern of textures removed by ME/MC process. As a result, the free-energy principle needs to be further explored in depth in video coding.

Our study targets at a VQA metric which is computationally efficient, and can be easily integrated into video coding process to have a close-loop solution of video coding optimization. Consequently, we adopt the pooling of blocked MSE, which is the only quality measurement in video coding, in our proposed scheme. Our study is inspired by the recent revealed free-energy principle [15], [16] and the psychophysical study by Stocker and Simoncelli on human visual speed perception (HVSP) [21], [22]. In HVSP, the HVS was first modeled as an efficient information extractor [23]. To achieve such efficiency, the HVS should pay more attention on the areas/regions that contain more information in a visual scene. For video signal, it is believed that object motion is associated with visual attention. Second, the visual perception was modeled as an information communication process and the HVS was an error-prone communication channel. The amount of information that can be perceived at the receiver end depends on the noise level of the communication channel (the HVS), so the internal noise in the HVS results in perception uncertainty. These two models suggest the spatial visual attention and the temporal visual attention, respectively in video signal perception. In addition, they separate the video signal into orderly and disorderly parts (sensory information and perception uncertainty) in spatio-temporal domain so that they could further finetune visual perception by applying different policies to the two separated video signal parts.

In this paper, a spatio-temporal weight function is defined and incorporated as weighting factors for pooling block based MSE for VQA. The block size is 4×4 which corresponds to the smallest processing unit of hybrid video coding (e.g., H.264/AVC and HEVC). Thus, a novel VQA metric, namely free-energy principle inspired video quality metric (FePVQ) is proposed for VQA. As a variant of MSE, FePVQ can be easily incorporated into hybrid video coding to have a close-loop solution of video coding optimization. To reach a low computational cost, the available information of video coding is reused in FePVQ as much as possible. We reuse the motion vectors to approximate HVSP for motion perception. In addition, we employ a computationally efficient model, namely relative total variation (RTV) [24] to distinguish orderly and disorderly information instead of the one [17] which was applied to image and was unallowable to video processing due to extremely high computational cost. The contributions of this paper lie in the following respects:

1) extending the free-energy principle accounting for IQA into spatio-temporal video signal quality assessment;
2) incorporating the recent psychophysical study on HVSP into VQA for measuring motion speed perception;
3) developing a low-complexity VQ metric for perceptual video coding optimization; and
4) simplifying the free-energy principle and HVSP models by employing the RTV and reusing motion vectors of video coding for saving computations.

The rest of this paper is organized as follows. Section II presents the proposed FePVQ algorithm in detail. Section III investigates the perceptual RDO of video coding based on FePVQ. Section IV evaluates the performance of FePVQ for predicting perceptual VQ, and the R-D efficiency of FePVQ-based video coding optimization by extensive experimental results. Finally, a brief conclusion is given in the last section.

II. Free-Energy Principle Inspired Video Quality Metric

According to free-energy principle [15], [16], orderly and disorderly regions are loosely associated with structural edges and textures, respectively in visual scene. The structural edges could provide more semantics information of a visual scene. For example, the structural edges of an image could clearly reveal the visual objects contained, such as peoples, animals, flowers, cars and buildings. Thus, they are most important to the HVS and should be preserved as much as possible in digital image processing. The textures of a scene are usually the surface of objects, such as texture patterns of wallpaper, grass, tree and road surface, which contain details of objects, and could further augment the objects with more appealing properties, such as fine texture, smooth grayscale transition and rich color, and make them vivid to human perception. Thus, structure and texture jointly render the users impressive perspective of visual scenes. However, for bit rate constraint usages, such as image/video coding, the quality of these two visual forms cannot be guaranteed at the same time. Generally, they are contradictory due to the constant bit rate in the communication systems. In these scenarios, the structure regions should be given the priority of bit quota compared with texture regions. Specifically, concerning free-energy principle in motion-compensated inter-frame video coding (named hybrid video coding, such as H.264/AVC), the highly textured regions may become disorderly with noise property after inter-frame ME/MC, as the repetitive pattern of textures have been removed by ME process. As a result, only structural edges after ME/MC still keep orderly and reveal clear profiles of visual objects. Thus, it is essential to process these two visual forms differently.

In this paper, 3-D gradient of video sequences is investigated as a 3-D cube consisting of tiles which are individual frames. In the 3-D cube, the moving objects would stretch out a plane along temporal trajectory. As a result, the video clip would vary along a certain spatio-temporal direction. Each pixel in the cube is associated with a 3-D gradient represented by a triple parameter (∂x, ∂y, ∂z), where ∂x, ∂y and ∂z are the partial derivatives along x, y and z axes, respectively. ∂x and ∂y are computed from the spatial

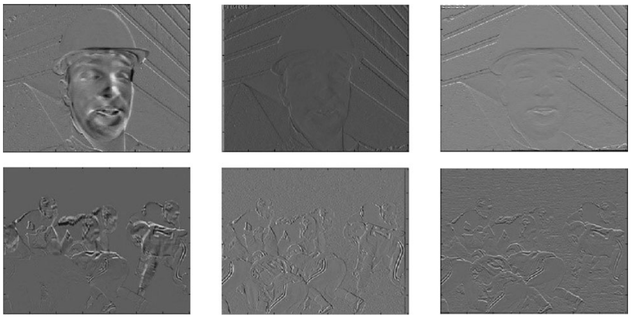


Fig. 1. Temporal and spatial (horizontal and vertical) gradients of video sequences (left: temporal gradient for video sequence; center: spatial vertical gradient; right: spatial horizontal gradient).

displacement of one pixel along horizontal and vertical directions, respectively. ∂z is computed from the temporal displacement of two neighboring frames. Fig. 1 shows the examples of $(\partial x, \partial y, \partial z)$ on two video sequences.

A. Textural Strength in Spatio-Temporal Gradient Domain

Generally, the gradient is directly used to compute the mutual correlation/similarity as in SSIM. In addition, SSIM is usually computed on pixel-basis, which requires more computations than blockwise metrics. For better integration with video coding, the proposed FePVQ uses the gradients as weights for MSE [25], [26] to simulate how the HVS responses to visual signals. Nowadays, PSNR/MSE is still widely used, especially for video compression. Thus, the quality metric in the variant form of MSE can be easier to be integrated into current visual application systems. In this work, we propose the texture strength at macroblock (MB) basis to depict the local texture properties for quality assessment, which is defined as

$$TS_m = \sum_{q \in MB} \left(\left|\partial_x F\right|_q + \left|\partial_y F\right|_q + \left|\partial_z F\right|_q \right) \qquad (1)$$

where F is the input signal, ∂x, ∂y and ∂z represent the partial derivatives, q denotes one pixel. Usually, the regions with detailed visual content, such as fine texture, moving objects, have the large gradient, so the texture strength is large. Besides spatial gradient, temporal gradient defined by the difference between two consecutive frames is also included in (1). From Fig. 1, we had the same observation that the texture regions and moving objects had large gradient. Thus, these three gradients, ∂x, ∂y and ∂z are combined to measure texture strength in (1).

Based on the knowledge that detailed texture regions can conceal more distortions than smooth regions [25], [26], a VQ metric is defined as the MSE weighted by texture strength [25] as

$$VQ_m^{TS} = \frac{1}{TS_m} \times MSE_m;\ m = 0, 1, 2, \ldots, M-1 \qquad (2)$$

where TS_m is a weight for the mth block of a frame, VQ_m^{TS} represents visual quality of the mth block and M is the total number of blocks in a frame.

B. Structure Strength by Considering Free-Energy Principle

In hybrid video coding, such as H.264/AVC, ME is widely used to eliminate temporal redundancy between successive frames for inter frame coding. The signal after ME/MC is of much less energy than the original one. That's why video can be compressed dramatically. However, there are still large residual coefficients for structural edges and fast moving objects, which consume a large part of the bit budget. Considering this aspect of the free-energy principle in hybrid video coding, texture after ME/MC is largely disorderly. The disorderly signal is similar in characteristic as noise, where there is little correlation between pixels. The structural signal mostly concerns the skeleton of objects in image. From (2), it can be deduced that the larger the texture strength, the less the visual quality loss under the same MSE. Referring to (1), it can be observed that both structure and texture regions are of large texture strength. Since the HVS is highly adapted to extract structural information, it is necessary to distinguish the structure information from the texture information so that they can be processed differently in VQA and video coding. The algorithm of [17] is of significant computational complexity although it can be used to separate structure and texture ideally. In [24], two models, *pixel-wise windowed total variation* and *windowed inherent variation* were defined as

$$\begin{cases} D_x(p) = & \sum_{q \in R(p)} g_{p,q} \left| (\partial_x F \right| \\ L_x(p) = & \left| \sum_{q \in R(p)} g_{p,q} (\partial_x F \right| \end{cases} \qquad (3)$$

where ∂x represents the partial differential operator on the input signal F, $R(p)$ is a neighboring region around the pixel p, $D_x(p)$ and $L_x(p)$ indicate the *pixel-wise windowed total variation* and *windowed inherent variation* are computed along x direction, and $g_{p,q}$ is a weighting function of Gaussian. Here, $D_x(p)$ counts the absolute spatial difference within the window $R(p)$, while $L_x(p)$ can measure the gradient consistency of $R(p)$. $D_y(p)$ and $L_y(p)$ can be computed in the same way as (3), but along y direction. Then, a RTV model was defined as $L(p)$ divided by $D(p)$ for structure-texture decomposition. A window only including texture generally has a smaller inherent variation $L_(p)$ than that including both texture and structural edges, since the gradient directions of the structural region are consistent while the gradient directions of the texture region are random. Fig. 2 shows the extracted structure from texture by using the RTV, where the original images (see Fig. 2(a)–(d)) have plentiful textures, grass, water and uneven ground, and the structural edges of these images are accurately extracted from texture background as shown in Fig. 2(e)–(h). In this study, the

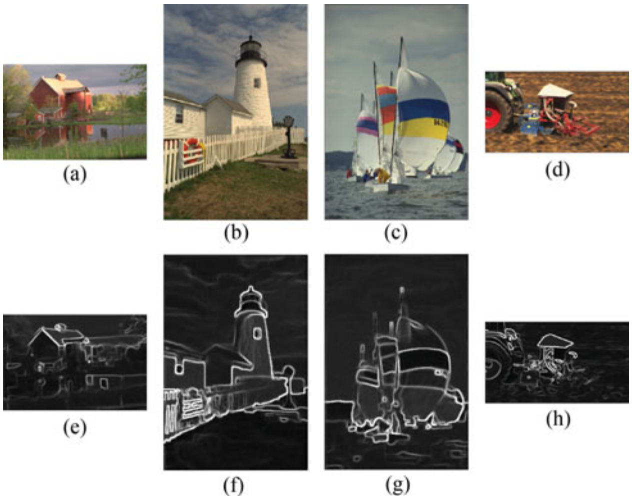


Fig. 2. Extraction of structures from textures. (a)-(d) are original images, "house," "linghthouse," "sailing2" (LIVE image database), and "tractor" (extracted from tractor YUV sequence), and (e)-(h) are extracted structure from (a)-(d).

RTV is further extended to temporal domain to classify structure and texture of video signals. Denote the three-dimension gradients of video as (∂x, ∂y, ∂z), the structure strength (SS) of each MB is defined as

$$SS_m = \left|\sum_{q\in MB}(\partial_x F)_q\right| + \left|\sum_{q\in MB}(\partial_y F)_q\right| + \left|\sum_{q\in MB}(\partial_z F)_q\right| \tag{4}$$

where SS_m represents SS of a block. Different from TS_m defined in (1), SS_m depends on the factor whether the gradient directions ("$+$" or "$-$") in a block are coincident or not. The blocks containing only texture generally lead to a smaller SS_m while those blocks containing structural edges have a larger SS_m.

Observing Fig. 1, which show temporal gradients for two video sequences, we can notice the obvious skeleton of each video frame. In addition, the skeleton is closely related to the boundaries of significant objects, which indicates that the structural edges can be separated from the background in an image by using temporal gradient information. To further enhance the fidelity of SS, the temporal gradient ∂z is included in (4) additionally. Furthermore, the temporal gradient can represent temporal saliency very well. Intuitively, (4) is a supplement to (1) since (1) cannot handle the difference between texture and structure regions.

Regarding disorderly and orderly regions [17], they are basically corresponding to texture and structure, respectively in ME/MC based hybrid video coding. In [17], the HVS responses to disorderly and orderly signals differently, which is integrated into a JND model. It is believed that texture regions usually become disorderly after ME/MC in hybrid video coding, so they have a larger JND threshold, and contribute less than structure ones to overall perceptual visual quality. From (4), SS_m of the texture regions is less than that of structure regions since the latter has inconsistent gradients.

C. Motion Strength Considering Human Visual Speed Perception to the HVS

In video, a moving object should be associated with visual attention and can be used for predicting visual fixations [29]. This is because an object with significant motion with respect to the background would be a strong surprisal. However, as the background motion is too large, the HVS cannot identify the objects with same accuracy as in static background, which would result in human perception uncertainty [21], [22]. For a natural video, a scene usually lasts for several seconds, where an almost same background remains for all the frames within the scene. For simplicity, only the relative motion represented by motion vectors is considered in this work for the sake of computational efficiency. In video coding, motion vectors are available, so they are utilised to approximate the perceptual motion model without extra computations.

Assume motion vector ($\vec{v}_r = (v_x, v_y)$), which is given at $N \times N$ block basis, e.g., a 4×4 block in H.264/AVC. When the multiple references are used, we first project $\vec{v}_r$ onto its nearest forward reference by $\vec{v}_r = \vec{v_r}/d$, where d represents distance from current frame to its reference regarding to v_r. Then, the motion strength is defined as

$$\begin{aligned} MS_m &= \alpha \times \sum_{i,j\in MB} \log|\vec{v}_r(i,j)| + \beta \\ |\vec{v}_r(i,j)| &= \sqrt{\frac{v_x(i,j)^2 + v_y(i,j)^2}{d(i,j)^2}} \end{aligned} \tag{5}$$

where MS_m is defined at MB basis, (v_x, v_y) is the motion vector of a 4×4 block and $d(i,j)$ is the distance from current frame to its reference; α and β are two constant parameters for fine-tuning the profile of MS_m. The logarithm in (5) accounts for "Weber-Fechner law" which states that the resolution of perception diminishes for the stimuli of greater magnitude, specifically, the subjective sensation is proportional to the logarithm of the stimuli intensity.

Referring to [22] and [21], (5) can be explained as the self-information of motion vector (relative motion) whose prior distribution was assumed as a power-law function

$$p(v) = Cv^{-\alpha} \tag{6}$$

where $\alpha > 0$ and $C > 0$ are two constant parameters; we use v to replace $|v|$ (motion vector length as defined in (5)) for simplicity. The most common approach for evaluating empirical data against a hypothesised power-law distribution is to observe that the power-law implies the linear form

$$-\log p(v) = \alpha \log v + \beta \tag{7}$$

where $\beta = -\log C$, "$-$" before $\log p(x)$ indicates self-information of $p(x)$. Although (7) is in a linear form with respect to α and β, the linear regression cannot be relied on to estimate α and β since it would result in significant systematic biases [31], [32]. Thus, the maximum likelihood estimation is employed to estimate α and β from empirical data [31], [32]. By collecting a set of motion vectors $\{\vec{v}_r(i)\}$ from practical video coding, the maximum likelihood estimation is performed as

$$\begin{cases} \alpha = 1 + n\Big[\sum_{i=1}^{n} \log \dfrac{v_i}{v_{\min}}\Big]^{-1} \\ \beta = -\log C = -\log\left((\alpha-1)v_{\min}^{\alpha-1}\right) \end{cases} \tag{8}$$

where $v_i, i = 1, \ldots, n$ is the observed value of v, $v_{\min}$ is the minimum value of v. It should be noticed that $v_{\min}$ corresponds not to the smallest value of v measured but to the smallest for which the power-law behaviour holds. We collect motion vectors by coding the first 16 frames (about half a second for 30 f/s) of the standard video sequences: "Foreman," "Football," "Mobile," "Bus," "Tennis" and "Flower" (CIF). The frames are coded in "P" frame by using quarter pixel ME and only one forward reference. A pair of α and β can be computed from (8) for each sequence, and then the averages of them ($\alpha = 4.55$, $\beta = 4.20$) are used to initialize α and β in (5). We found that α and β are related to motion vector scale and specific video coding system closely. That is why the values of α and β are much different from those in [22]

By combining (1), (4), (5) and local MSE, a new VQA metric, i.e., FePVQ is designed at MB basis as

$$VQ_m = \frac{MS_m^a \times SS_m^b}{TS_m^c} \times MSE_m \tag{9}$$

where MS_m, SS_m and TS_m representing motion strength, SS and texture strength, respectively, are weighting factors of local MSE; a, b and c are constants to finetune the profiles of MS_m, SS_m and TS_m, respectively. They are handpicked and set to 1.25, 1.25 and 1.2 in our experiments. We found that this setting is competitive among all the testings. However, more systematic way to choose these parameters could recur to optimization methods. Equation (9) gives visual quality score of each MB, and the visual quality score of a frame is computed by summing up the scores of all the MBs.

III. Perceptual Video Coding Optimization

The proposed FePVQ is naturally associated with quantization parameter (QP) since MSE in (9) can be theoretically formulated by QP function [33]. Thus, it can be directly integrated into video coding systems to obtain a closed-form analytic solution for perceptual video coding optimization. Since FePVQ is designed based on blocks, it is suitable for optimizing any block based process in video coding, such as block-based ME, intra prediction, inter prediction, and RDO. In video coding, multiple modes are provided to adapt to video content variation. Different modes represent different block sizes, prediction types, transforms and partitions of an MB. For MBs with simple textured video content, a larger size mode is preferred. On the contrary, the MBs with fine textured video content may benefit from the small partitions of an MB. The best partition mode is decided by RDO, which seeks the best tradeoff between distortion and bit cost as [34], [35]

$$J = D + \lambda \times R \tag{10}$$

where D is measured in MSE for conventional RDO, R is the number of bits for coding an MB, and λ is known as the Lagrange multiplier which controls the tradeoff between D and R. With the same distortion, the mode with the least bit cost would be the best mode. Equivalently, the mode with the least MSE would be the best mode under the same bit cost. For the perceptual optimization, a perceptual measurement of distortion instead of MSE is desirable for RDO. In [36], the authors proposed to use SSIM instead of MSE to make mode decision and RDO. However, the existing RDO scheme [34] which was theoretically optimized for MSE measurement may not be optimal to SSIM. The proposed FePVQ is extended from MSE, and consequently it is completely compatible with the existing RDO scheme. In our previous work [37], the optimal perceptual VQ control was investigated for bit rate constrained video coding task. In this paper, we go deep into the bit rate constrained optimization problem of video coding. Firstly, perceptual RDO is studied by replacing D measured by MSE in (10) with FePVQ as

$$\begin{aligned} J &= VQ_m + \lambda \times R_m \\ &= \frac{MS_m \times SS_m}{TS_m} \times MSE_m + \lambda \times R_m \\ &= p_m \times MSE_m + \lambda \times R_m \end{aligned} \tag{11}$$

which seeks the best tradeoff between perceptual visual quality measured by FePVQ and bit rate cost. Secondly, according to masking property of the HVS, which describes why similar artifacts are more visible in certain region of a frame while they are hardly noticeable in other regions, the QP is adapted accordingly to different video contents to achieve consistent visual quality in a whole visual scene.

With the assumptions of uniform quantization scheme and uniform DCT coefficients distribution, MSE is usually modeled by $Q_2/12$ [33]. Thus, the proposed FePVQ can be modeled theoretically by

$$VQ_m = \frac{MS_m \times SS_m}{TS_m} \times \frac{Q_m^2}{12} = p_m \times Q_m^2 \tag{12}$$

which indicates that the VQs of MBs are different from each other under the same quantization. The underlying reason is that different kinds of input visual signals are perceived differently by the HVS. So $\{p_m\}$ actually provides a visual perception map of a frame consisting of various kinds of visual signals. In practice, p_m in both (11) and (12) acts as a scaling factor, which is normalized by a median value of p_m and clipped into [0.75, 1.25]. The distortions of regions with simple texture background, strong structural edge and high motion strength are more sensitive to the HVS. For consistent visual quality, after obtaining QP of each frame, denoted by Q_f, we could compute a new QP for each MB by referring to Q_f

$$Q_m = Q_f / \sqrt{p_m}. \tag{13}$$

Referring to (11), p_m changes the weight of MSE in RDO, which would revise the final mode selection, and therefore obtain better perceptual quality for same bit cost. In addition, the QP also changes in a certain range so as to have more choices for perceptual optimization. Specifically, for p_m larger than 1.1, the RDO is performed as

$$\begin{aligned} &\min\{J(s,T)\} \\ &J(s,t) = VQ_m(s,t) + \lambda(t) \times R_m(s,t) \\ &s \in [Q-2, Q-1, Q]; \\ &t \in \text{skip/direct}, 16\times16, 16\times8, \ldots, 4\times8,\ 4\times4 \end{aligned} \tag{14}$$

where t and s represents mode and QP, respectively. For p_m less than 0.9, the RDO is performed as

$$\begin{aligned} &\min\{J(s,T)\} \\ &J(s,t) = VQ_m(s,t) + \lambda(t) \times R_m(s,t) \\ &s \in [Q+2, Q+1, Q]; \\ &t \in \text{skip/direct}, 16\times16, 16\times8, \ldots, 4\times8,\ 4\times4. \end{aligned} \tag{15}$$

Such a process can achieve a better tradeoff between perceptual visual quality and bit cost.

In FePVQ, significant structural edges would have larger SS_m and thus increase the weight of visual quality contribution in a whole image. This situation is consistent to the perception process of the HVS to an input image. This is why FePVQ is much better than the previous scheme [37]. By integrating FePVQ into video encoders, the structural edges of an image can be largely preserved while the quality of disorderly texture regions is compromised to save more bits for coding structural edges. Thus, better perceptual visual quality of video encoding can be achieved under the same target bits quota.

IV. Experimental Results and Discussions

In Part A below, we evaluate the performance of FePVQ to predict visual quality of images and videos. Then, it is integrated into HM14.0 reference software of HEVC to optimize video encoding from the perspective of perceptual visual quality. In Part B, we show how FePVQ benefits video coding optimization. Part C discusses the associated computational complexity analysis.

A. Performance of FePVQ Metric

The performance of a VQA metric can be evaluated by depicting the relationship of the predicted objective quality scores and the provided subjective quality ratings [difference mean opinion score (DMOSs)]. The DMOS is provided by subjective viewing tests where the viewers are requested to give their opinions on visual quality of distorted visual signals. The DMOSs and associated images/videos consist of subjective quality databases to be the benchmarks for evaluating user-defined objective quality metrics. The used databases in our experiments are introduced briefly in Table I. We follow the procedure [38] recommended by video quality expert group to evaluate the performance of VQA approaches.

TABLE I
IMAGE SUBJECTIVE QUALITY DATABASES

Database	Description
LIVE Image [39]	There are 29 reference images, and totally 779 distorted images with five distortions types: JPEG, JPEG2000, white noise, Gaussian blur, and simulated Rayleigh fading channel.
TID2008 [40]	The TID2008 contains 25 reference images and 1700 distorted images (25 reference images × 17 types of distortions × 4 levels of distortions). Distortion types mostly concern all kinds of noises, blur and JPEG/JPEG2000 compression and transmission error and block-wise distortions.
CSIQ [41]	There are 30 reference images, and totally 866 distorted images generated from six different types of distortion: JPEG, JPEG2000, Global contrast decrements, Gaussian blur, addictive Gaussian white noise, and addictive Gaussian pink noise at four to five different levels of distortion.
LIVE Video [45]	Ten videos: 250 frames, 768×432, 25/50f/s, YUV 4:2:0 and progressive; 150 distorted videos, the distortion types include: transmission over wireless networks, transmission over error-prone IP networks, MPEG-2 compression and H.264 compression; camera patterns: still, camera pan, circular camera motion and zoom in; ownership: Image & Video Engineering (LIVE) Laboratory, University of Texas at Austin.
IVP Video [46]	10 high-resolution reference videos, 1920×1088, YUV 4:2:0 and progressive; 128 distorted ones generated using 4 types of distortions: MPEG-2 compression, Dirac wavelet compression, H.264 compression and packet loss on the H.264 streaming through IP networks.
CSIQ Video [47]	12 high-quality reference videos, 832×480, YUV 4:2:0 and progressive; 216 distorted videos with six different types of distortion, i.e., H.264 compression, HEVC compression, MJPEG compression, Wavelet-based compression, H.264 bitstream over wireless channel and AWN. Each distortion type has three different distortion levels.

TABLE II
PERFORMANCE COMPARISON ON MULTIPLE DATABASES

Database		PSNR	SSIM	MSSIM	VIF	ESSIM	FSIM	FePVQ (18)
TID2008	PLCC	0.5706	0.7912	0.8425	0.8090	**0.8808**	0.8497	0.8493
	SROCC	0.5800	0.7965	0.8528	0.7496	**0.8856**	0.8574	0.8607
	RMSE	1.1020	0.3740	0.7299	0.7888	0.2890	0.3220	0.3230
CSIQ	PLCC	0.8001	0.8613	0.8998	0.9277	0.8763	**0.9120**	0.8133
	SROCC	0.8057	0.8756	0.9138	0.9193	0.8879	**0.9242**	0.8267
	RMSE	0.1575	0.1334	0.1145	0.0980	0.0968	0.1077	0.1078
LIVE Image	PLCC	0.8744	0.8242	0.9430	0.9598	0.9547	**0.9604**	0.9318
	SROCC	0.8813	0.9112	0.9445	0.9631	0.9611	**0.9652**	0.9375
	RMSE	13.2300	15.4430	9.0956	7.6734	8.1180	7.5970	9.8990

The monotonic logistic function with five parameters β_1, β_2, β_3, β_4, β_5 is used to map visual quality index computed from VQA metrics into final objective score V_j as

$$V_j = \beta_1 \times \left(0.5 - \frac{1}{1 + e^{\beta_2 \times (x_j - \beta_3))}}\right) + \beta_4 \times x_j + \beta_5. \tag{16}$$

The parameters are determined by some parametric minimization approaches, such as least squares. Three statistical measurements, linear correlation coefficient (LCC), Spearman's rank order correlation coefficient (SROCC) and RMSE are used to measure the degree of correlation between DMOS and predicted values of VQA metric. Given two sets A and B, LCC is defined as the correlation coefficient of A and B as

$$LCC(A, B) = \frac{\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{n}(A_i - \bar{A})^2}\sqrt{\sum_{i=1}^{n}(B_i - \bar{B})^2}} \tag{17}$$

where $\bar{A}$ and $\bar{B}$ represent the means of A and B, respectively. Intrinsically, LCC measures the accuracy of a VQA metric to predict DMOS. SROCC evaluates the prediction monotonicity, and RMSE measures the error during the fitting process. Larger LCC and SROCC values indicate the better correlation between objective and subjective scores, while smaller RMSE means smaller error of VQA prediction, therefore a better performance.

1) Performance of FePVQ on Images: Although (9) is designed for VQA, it can be directly adapted as

$$VQ_m = \frac{SS_m}{TS_m} \times MSE_m \tag{18}$$

for IQA, where TS_m and SS_m are associated with image instead of video. We compare (18) with several state-of-the-art IQA metrics: PSNR, SSIM [9], MSSIM [10], VIF [41], FSIM [42] and ESSIM [43] over LIVE image [39], TID2008 [40], and CSIQ [41] databases. The statistics of LCC and SROCC are tabulated in Table II. It shows that (18) can improve LCC and SROCC up to 0.9318 and 0.9375, respectively on LIVE image database [40], which is significantly beyond PSNR/MSE whose LCC and SROCC are 0.8744 and 0.8813, respectively. The reason is the use of texture strength and SS inspired by the free-energy principle of HVS.

2) Performance of FePVQ on Videos: To evaluate FePVQ on videos, the LIVE video [45], Image Video Processing (IVP) video [46] and CSIQ video [44] databases are used. In this experiment, the proposed FePVQ in (9) for videos is compared with state-of-the-art metrics including PSNR, SSIM [9], MSSIM [10], FSIM [42], ESSIM [43], MOVIE [45], VSNR [46], NTIA-VQM [48], VIF [41] and our previously proposed one [37]. The LCC and SROCC statistics on LIVE video database are tabulated in Table III. As PSNR, SSIM, MSSIM, VSNR and VIF only provide frame-level quality scores, the final quality index of video is generated by averaging frame-level quality scores. From Table III, it can be observed that PSNR performs poorly since it is loosely related to the HVS perception. The VSNR is also not satisfactory since it measures distortion as to the HVS perception in wavelet domain, while MPEG-2 and H.264/AVC are associated with quantization distortion in DCT domain. In addition, VSNR is an image quality metric accounting for spatial distortion. For VQA, the temporal information is very important and needs to be highlighted. This is

TABLE III
PERFORMANCE COMPARISON OF FEPVQ AND BENCHMARKS ON LIVE VIDEO DATABASE

Algorithm		Wireless	IP	H.264	MPEG-2	All Data
PSNR	LCC	0.6768	0.4779	0.5883	0.4088	0.5690
	SROCC	0.6705	0.4296	0.4773	0.3939	0.5527
SSIM	LCC	0.4732	0.5371	0.6109	0.5815	0.5025
	SROCC	0.5386	0.4741	0.6585	0.5690	0.5326
MSSIM	LCC	0.6841	0.6838	0.6924	0.6315	0.6762
	SROCC	0.7291	0.6449	0.7343	0.6810	0.7351
VSNR	LCC	0.6950	0.7364	0.6187	0.6782	0.6884
	SROCC	0.6940	0.6930	0.6405	0.5874	0.6715
VIF	LCC	0.5929	0.6364	0.6488	0.6727	0.5760
	SROCC	0.5375	0.5533	0.6377	0.6346	0.5579
WSNR	LCC	0.6797	0.7218	0.5824	0.5853	0.6776
	SROCC	0.6458	0.6932	0.5733	0.6154	0.6296
ESSIM	LCC	0.4420	0.5346	0.6675	0.5671	0.5017
	SROCC	0.4056	0.5158	0.6396	0.5682	0.4958
FSIM	LCC	0.7289	0.7416	0.6964	0.6641	0.7013
	SROCC	0.7435	0.7090	0.6947	0.6941	0.7322
NTIA-	LCC	0.7777	0.7087	0.7295	0.8631	0.7741
VQM	SROCC	0.7685	0.6761	0.7388	0.8414	0.7527
MOVIE	LCC	**0.8360**	**0.7566**	0.7905	**0.7969**	0.8101
	SROCC	**0.8109**	**0.7157**	0.7664	**0.7733**	0.7890
L. Xu	LCC	0.7617	0.7357	0.7085	0.5550	0.7410
[37]	SROCC	0.7533	0.7237	0.6642	0.5637	0.7211
FePVQ	LCC	0.8333	0.6787	**0.8983**	0.7348	**0.8326**
(9)	SROCC	0.8073	0.7417	**0.8725**	0.7513	**0.8279**

also the reason why SSIM, MSSIM and VIF perform successfully in image quality evaluation, but not so well for VQA.

FePVQ outperforms all the competitors, which means that it can effectively predict the perceptual quality of the distorted videos. Both motion strength and SS in (9) account for the success of FePVQ. As compared to our previous work [37], FePVQ is improved due to incorporating the additional SS. From Table III, FePVQ outperforms MOVIE with LCC and SROCC being 0.8324 and 0.8281, respectively. Especially, the LCC and SROCC can be up to 0.8979 and 0.8728 on H.264/AVC compression, which are much better than all the competitors. Since the video coding, specifically H.264/AVC, is highly concerned in this work, so the good performance of FePVQ for H.264/AVC compression is indeed desirable. It should be pointed out that although FePVQ is more competitive on H.264/AVC compression, it is a general approach for VQA. In Fig. 3, the scatter-plots after mapping are compared against the different VQA approaches over the LIVE VQ database. It can be observed that for FePVQ, the sample points scatter more closely around the fitted line. It means that the values predicted by FePVQ correlate better with the subjective ratings, specifically the DMOS values, demonstrating a better performance than other metrics.

From Table III, the performance of FePVQ is the best on "H.264" videos (high correlations), and performs less well (not as competitive with benchmarks on the other distortions ["IP," "Wireless" and "MPEG-2")]. From Table III, we can see that the performance of the proposed FePVQ is not as good as MOVIE on "IP," "Wireless" and "MPEG-2." To explain this observation in depth, we checked the videos encoded by "Wireless," "IP," "H.264" and "MPEG-2," and extracted some frames shown in Fig. 4. From Fig. 4, the "Wireless" and "IP" have the unnatural blockiness, structural edges due to packet loss and some error concealment techniques, such as copy the neighor block. These two distortions are usually charactered by unnatural blockiness, structural edges. The free-energy priciple is more likely applicable to natural texture, structures and object edges, however it fails to those kinds of unnatural artifacts. Thus, FePVQ on "Wireless" and "IP" performs not as good as it on "H.264" and "MPEG-2." Comparing "MPEG-2" with "H.264," we can notice that the severe blockiness exists in "MPEG-2," as shown in Fig. 4(c), while it is not obvious in "H.264" (blur artifact is more obvious) as shown in Fig. 4(d). This phenomenon is caused by several reasons: 1) ME with variable block size and multiple references result in better prediction, and thus smaller residual in "H.264," which finally reduces blockiness caused by block-based transform followed by quantization; 2) the smaller size DCT transform in "H.264" (4 × 4 and 8 × 8), "MPEG-2" uses 8 × 8 DCT; 3) improved deblocking filter (loop-filtering) in "H.264." From Fig. 4(c), blockiness containing a special textural and structural pattern which however is more likely to be "artificial" but not "natural" that the free-energy principle accounts for. Thus, FePVQ performs better on "H.264" than "MPEG-2."

IVP video database is established by the IVP Laboratory of the Chinese University of Hong Kong. The distorted videos are generated by four types of distortions: MPEG-2 compression (MPEG-2), Dirac wavelet compression (Dirac), H.264 compression (H.264) and packet loss on the H.264 streaming through IP networks (IP). We tested the VQ metrics on this database and reported LCC and SROCC statistics in Table IV. From Table IV, it is noticed that PSNR performs better than SSIM, which is because IVP database was optimized on PSNR [46]. FePVQ is the best among all the compared metrics on H.264 and MPEG-2 distortions since it is specifically optimized for block based video codings, such as H.264, MPEG-2 and HEVC. MOVIE performs better than FePVQ on Dirac distortion, which may be because MOVIE also employs a sub-bands signal decomposition as the Dirac does. Across all distortion types, it can be observed FePVQ is the best among all the compared methods.

The CSIQ video database developed in last year is the newest VQ database. It includes the newest HEVC compression distortion. Besides HEVC, H.264, Motion JPEG compression (MJPEG), Wavelet-based compression using SNOW codec [47] (SNOW), H.264 bitstream over wireless channel (Wireless) and additive white noise (AWN) distortions are included. We also tested FePVQ and the benchmarks on this database and reported LCC and SROCC statistics in Table V. From Table V, FePVQ is significantly better than all the benchmarks on H.264 and HEVC distortions, which is because FePVQ is specifically designed for video coding optimization. FePVQ is also the best on MJPEG distortion. For Wireless and Snow distortions, MOVIE performs better than FePVQ. MOVIE is defined on the Gabor sub-bands of signals, which make it more efficient on Snow because Snow also concerns the wavelet sub-bands decomposition. For AWN distortion, it can be observed that PSNR and SSIM perform better than others. The statistics across all the distortion types in Table V shows that FePVQ is the best among all the compared metrics.

B. Perceptual Video Encoding Optimization

Replacing MSE measurement of conventional encoder by FePVQ in RDO, perceptual optimization of video encoding can be achieved. We use HM14.0 reference software of HEVC [49] under the default configuration (*IntraPeriod=−1, MaxCUWidth/Height=64, MaxPartitionDepth=4, QuadtreeTU Log2MaxSize=5, QuadtreeTULog2MinSize =2, QuadtreeTU MaxDepthInter/Intra=3, DecodingRefreshType=0, GOPSize= 4, FastSearch=1, SearchRange=64, RDOQ=1, RDOQTS=1*) to conduct the experiment. Six standard video sequences (see Table VI) of HEVC are tested. They are with plenty of scenes, includ-

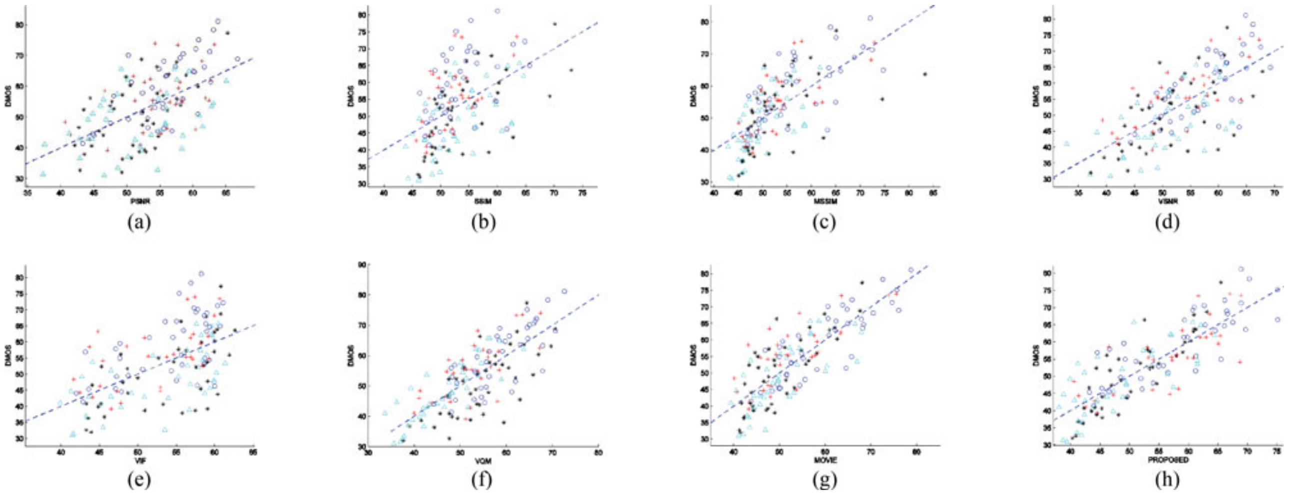


Fig. 3. Scatter plots of the DMOS values versus model predictions on the LIVE VQ database. Each sample point represents one test video. (The circle and "+" indicate the distorted video sequences under wireless network and IP network, respectively. The star indicates H.264/AVC encoded video sequence, while the triangle indicates the MPEG-2 compressed one.) First row from left to right: (a) PSNR, (b) SSIM, (c) MSSIM, and (d) VSNR; second row from left to right: (e) VIF, (f) NTIA-VQM, (g) MOVIE, and (h) the proposed FePVQ (9).

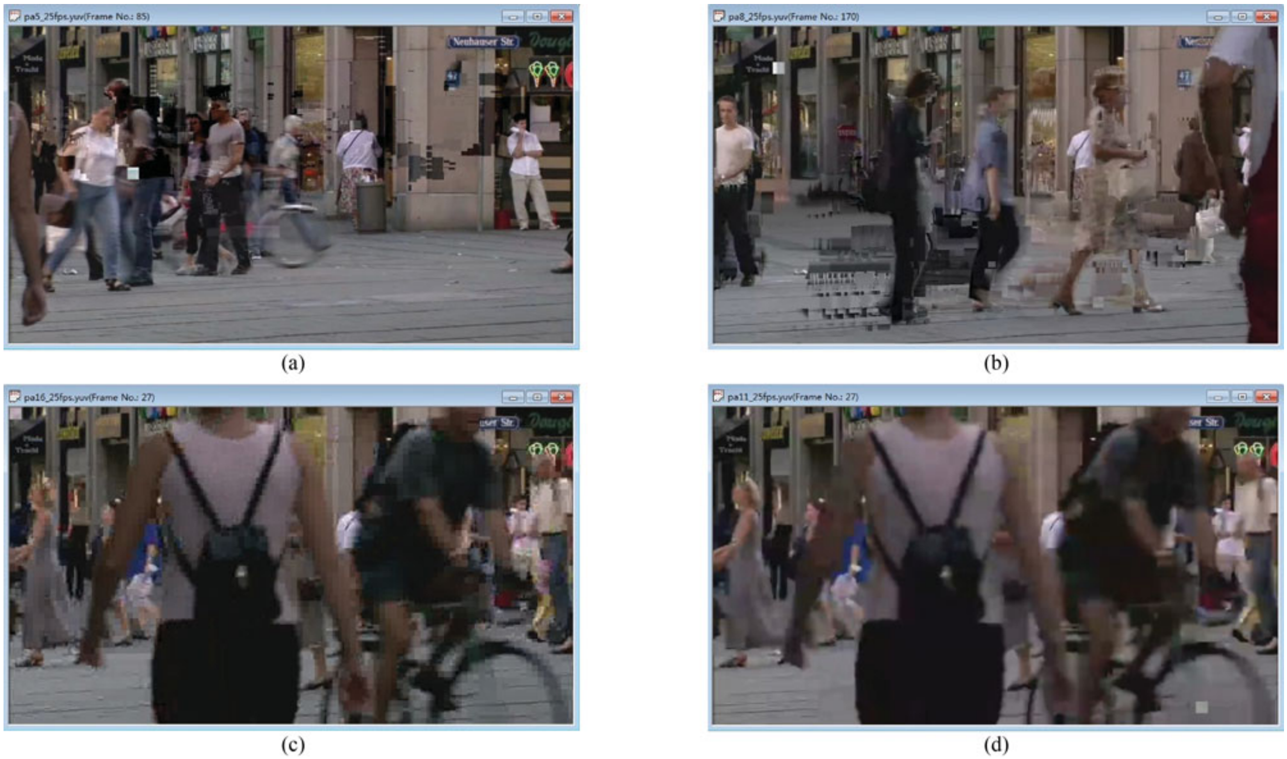

pa5_25fps.yuv(Frame No.: 85)
Neuhauser Str.
(a)
pa8_25fps.yuv(Frame No.: 170)
(b)
pa16_25fps.yuv(Frame No.: 27)
(c)
pa11_25fps.yuv(Frame No.: 27)
(d)

Fig. 4. Different characteristics of each kind of distortions (all frames are from video sequence "Pedestrian Area"; (a) is the 85th frame from "Wireless" distortion, (b) is the 170th frame from "IP" distortion, and (c) and (d) are the 27th frame from "MPEG-2" and "H.264" distortions, respectively).

ing indoors, outdoors, natural scenery, surveillance and busy street. Each video sequence has 300 frames, and resolution ranges widely. The constant QP coding (QP $=$ 24, 28, 32 and 36) is performed in this experiment.

The R-D comparison is conducted between the original RDO [34] and the proposed one. The original RDO means the traditional MSE based optimization, and the proposed one employs FePVQ for RDO. Significant R-D improvement can be achieved by the proposed RDO over the original one. On average, an improvement about 0.25 dB VQ and 3.4% bit rate saving can be achieved (see Appendix). In this case, the R-D improvement is measured by FePVQ. To avoid the bias that the R-D improvement is just because of the different measurements instead of the RDOs, NTIA-VQM [48] is also used to measure R-D performance. The R-D statistics are listed in Table VII, where the smaller the NTIA-VQM value, the better the visual quality. From Table VII, the proposed RDO is

TABLE IV
PERFORMANCE COMPARISON OF FEPVQ AND BENCHMARKS ON IVP VIDEO DATABASE

Distortion		PSNR	SSIM	NTIA-VQM	MOVIE	FePVQ
Dirac	LCC	0.826	0.739	0.865	**0.888**	0.881
	SROCC	0.846	0.789	0.884	**0.870**	0.867
H.264	LCC	0.771	0.664	0.868	0.823	**0.861**
	SROCC	0.799	0.629	0.864	0.845	**0.856**
MPEG-2	LCC	0.613	0.608	0.786	0.858	**0.873**
	SROCC	0.700	0.572	0.787	0.842	**0.856**
IP	LCC	0.668	0.536	0.535	0.823	**0.837**
	SROCC	0.639	0.511	0.466	**0.824**	0.819
All	LCC	0.682	0.546	0.611	0.880	**0.911**
	SROCC	0.709	0.565	0.643	0.879	**0.884**

TABLE V
PERFORMANCE COMPARISON OF FEPVQ AND BENCHMARKS ON CSIQ VIDEO DATABASE

Distortion		PSNR	SSIM	NTIA-VQM	MOVIE	FePVQ
H.264	LCC	0.802	0.755	0.919	0.897	**0.920**
	SROCC	0.835	0.717	0.916	0.904	**0.912**
Wireless	LCC	0.851	0.716	0.801	**0.886**	0.881
	SROCC	0.802	0.722	0.806	**0.882**	0.873
MJPEG	LCC	0.509	0.734	0.647	0.887	**0.896**
	SROCC	0.460	0.747	0.641	0.882	**0.887**
Snow	LCC	0.759	0.716	0.874	**0.900**	0.889
	SROCC	0.769	0.733	0.840	**0.898**	0.875
AWN	LCC	**0.906**	0.894	0.884	0.843	0.869
	SROCC	**0.949**	0.897	0.918	0.855	0.848
HEVC	LCC	0.785	0.523	0.906	0.933	**0.943**
	SROCC	0.805	0.580	0.915	0.937	**0.941**
All	LCC	0.579	0.516	0.789	0.806	**0.821**
	SROCC	0.565	0.558	0.769	0.788	**0.810**

TABLE VI
SEQUENCE INFORMATION

Sequence	Resolution	Frame rate
Flowervase	832 × 480	30.00
FourPeople	1920 × 1080	60.00
Mobisode	416 × 240	30.00
ParkScene	1920 × 1080	24.00
PeopleOnStreat	2560 × 1600	30.00
Vidyo	1280 × 720	60.00
Keiba	832 × 480	30.00

TABLE VII
PERFORMANCE COMPARISON BETWEEN THE PROPOSED PERCEPTUAL RDO AND THE ORIGINAL RDO ("IPPP" ENCODING STRUCTURE, FOR A FAIR COMPARISON, BOTH OF THE TWO RDOS USE NTIA-VQM FOR EVALUATING VISUAL QUALITY)

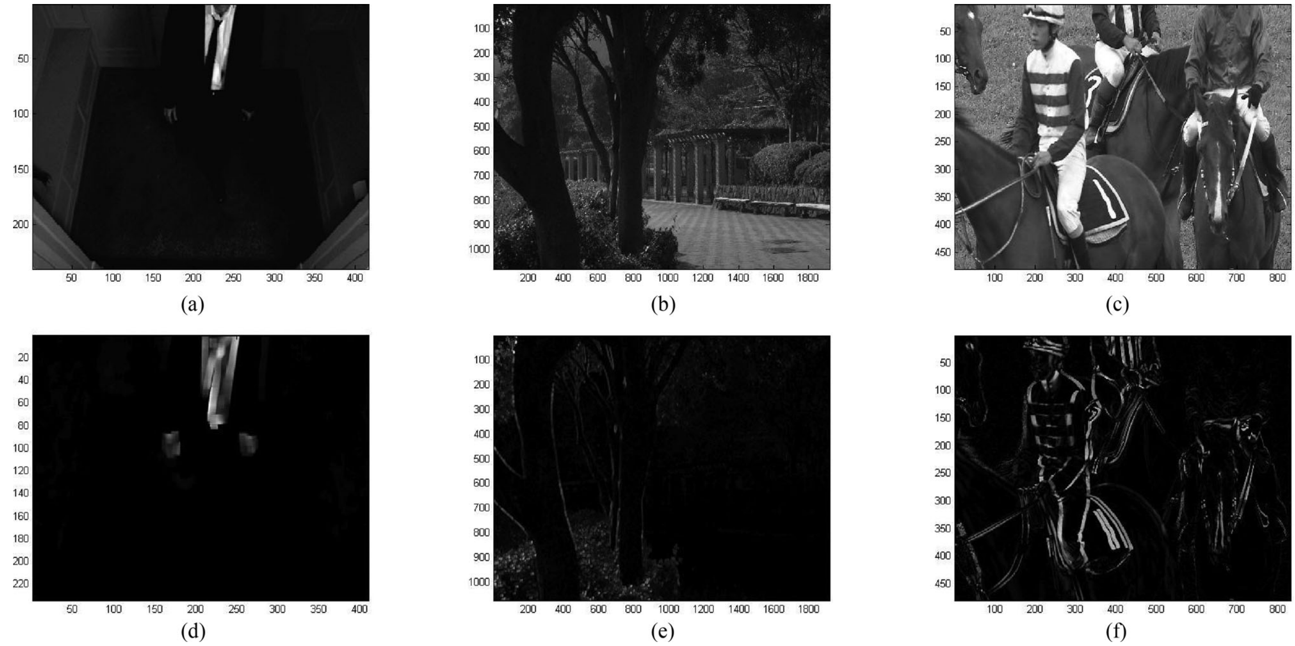
	Original		Proposed		R-D gain	
Sequence	Bit rate	NTIA -VQM	Bit rate	NTIA -VQM	ΔBit rate	ΔNTIA
	(kb/s)	(dB)	(kb/s)	(dB)	(%)	-VQM
Flowervase	447.22	0.1197	433.07	0.1161	−3.16	−0.0036
	221.82	0.2202	214.36	0.2202	−3.36	0.0000
	119.43	0.3579	114.89	0.3527	−3.80	−0.0052
	67.79	0.4955	64.57	0.4927	−4.75	−0.0028
FourPeople	2062.93	0.0911	1999.71	0.0917	−3.06	0.0007
	1148.40	0.1781	1112.30	0.1748	−3.14	−0.0033
	706.05	0.2897	684.27	0.2893	−3.08	−0.0004
	452.67	0.4178	438.18	0.4246	−3.20	0.0068
Mobisode	70.60	0.1087	67.71	0.1066	−4.09	−0.0021
	43.06	0.2056	40.68	0.1995	−5.54	−0.0060
	28.18	0.3189	26.93	0.3138	−4.46	−0.0051
	20.06	0.4497	19.33	0.4443	−3.66	−0.0054
ParkScene	6598.66	0.1279	6399.74	0.1232	−3.01	−0.0047
	3312.21	0.2303	3211.71	0.2261	−3.03	−0.0042
	1709.36	0.3573	1657.72	0.3514	−3.02	−0.0060
	894.92	0.4936	867.44	0.4884	−3.07	−0.0052
People -OnStreat	28653.02	0.0751	27792.09	0.0716	−3.00	−0.0035
	15797.31	0.1678	15322.92	0.1665	−3.00	−0.0013
	9399.33	0.2801	9115.77	0.2809	−3.02	0.0008
	5974.74	0.3964	5795.30	0.3978	−3.00	0.0013
Vidyo	1838.74	0.1585	1783.33	0.1501	−3.01	−0.0084
	936.03	0.2658	906.81	0.2573	−3.12	−0.0085
	542.64	0.3935	525.34	0.3849	−3.19	−0.0086
	339.09	0.5287	328.91	0.5124	−3.00	−0.0163
Average					**−3.41**	**−0.0038**

superior to the original one with 3.41% bit rate saving and 0.0038 VQ improvement measured by NTIA-VQM.

We draw the visual perception map $\{p_m\}$ [referring to (12)] in Fig. 5 for the test sequences, which shows a good agreement between high perception and significant structural edges, boundaries of motion objects. For mapping $\{p_m\}$ onto original images, we compute the visual perception map for each pixel by overlapping blocks size of 8 × 8, so the same size of visual perception map as original image is given in Fig. 5. It can be observed that:

1) the clear profiles of objects can be successively outlined by the visual perception maps;
2) for static objects, the visual perception map does not give the high visual perception value as the same as that of moving objects. For example, Fig. 5(b) shows a surveillance video in which a man steps out of elevator. The elevator is static relative to surveillance camera, so it is with low visual perception; and
3) for the natural scenes, the significantly structural edges can be successfully distinguished from the high texture background, e.g., Fig. 5(c) and (h), the trunk of trees, human bodies, horses are obvious against the background.

For evaluating the subjective visual quality improvement, we conducted the subjective evaluation in IVP Lab of CUHK [46]. The evaluation was performed in a studio room with lighting condition conforming to the BT.500 standard [50]. The display monitor is a 65" Panasonic plasma display (TH-65PF9WK). 15 subjects participated in the subjective test. All of them are non-experts. Their eyesight was either normal or had been corrected to be normal with spectacles. Each observer assessed 6 source videos and 48 distorted videos. A single-stimulate method [51] with absolute category rating (ACR) scale was used where each video (including the reference) occurred once in a random order, and yet the two successive videos come from different source videos so as to remove contextual and memory effects in quality evaluation. The ACR scale employs a five-category discrete quality judgment, specifically 5-excellent, 4-good, 3-fair, 2-poor, and 1-bad. At the beginning of the test, three videos were arranged as the training videos to stabilize the observers' opinion. Subjective rating of the compressed video was subtracted from that of the reference video. The difference values were processed using the method described in the BT.500 standard [50] to derive the DMOS.



(a) (b) (c)

(d) (e) (f)

Fig. 5. Visual perception maps [(a), (d) Mobisode; (b), (e) ParkScene; (c), (f) Keiba; all frames are the 11th frame coded by "P"].

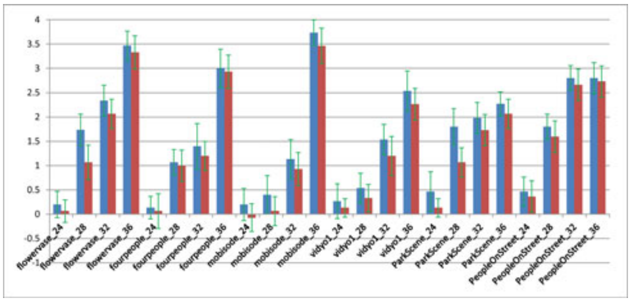


Fig. 6. Subjective quality comparison between the proposed perceptual RDO and the original one.

Five uncompressed, high quality source videos of natural scenes are used to create the distorted videos using four different QPs (QP = 24, 28, 32 and 36) for the standard RDO of HEVC and the proposed perceptual RDO, respectively. All videos contain 300 frames. The subjective experiments are performed to prove the better subjective visual quality of the proposed RDO over the traditional one. The subjective evaluation result is shown in Fig. 6. The horizontal axis indicates the encoded video sequences in different bit-rates. The red histograms are the DMOS values of video sequences generated by the proposed RDO by employing FePVQ, while the blue ones are the DMOS values from the original RDO of HEVC. The green error bar indicates the variance of the subjective rating scores for each encoded sequence. The smaller DMOS value indicates the better visual quality of the encoded video sequence. It can be observed that the proposed RDO can outperform the other one on most of the test sequences.

C. Computational Complexity Analysis

Regarding the computational complexity of FePVQ, for a 250-frame 768 × 432 sequence (LIVE video Database) on a 3G Hz quad-core CPU with 6G RAM, processing time of PSNR, SSIM, MS-SSIM, VSNR, VIF, NTIA-VQM, ESSIM, FSIM, MOVIE and FePVQ is given in Table VI. All codes are implemented using MATLAB (without specific optimization) except for MOVIE which uses C++. In Table VIII, the time complexity is measured in second and the value represent the average time per each video sequence.

TABLE VIII
COMPUTING TIME OF FEPVQ AND THE BENCHMARKS

Metric	PSNR	SSIM	MS-SSIM	VSNR	NTIA-VQM	ESSIM	FSIM	VIF	MOVIE	FePVQ (9)
Average processing time (in second)	4	24	60	26	57	64	84	636	6320	45

From Table VIII, FePVQ [(9)] is superior to MOVIE, VIF, MS-SSIM, ESSIM, FSIM and NTIA-VQM in terms of computational complexity. In addition, it would be better when applied to video coding, since the MSEs of MBs have been computed in RDO stage of video encoding. Furthermore, FePVQ builds a relation between perceptual visual quality and quantization of video compression in order to guide the perceptual video encoding besides assessing visual quality. The complexity of VIF and MOVIE majorly comes from their subbands decomposition. VIF uses overcomplete steerable pyramid decomposition. MOVIE uses 3-D Gabor filters to decompose the video locally into more than 100 spatiotemporal channels. The complexity makes their uses in video coding impractical.

V. CONCLUSION

Although much research has been done in IQA and VQA, it is important to derive a well-grounded perceptual measurement suitable for video coding. In this paper, a novel VQA metric, FePVQ is proposed for VQA. The free energy principle in neuroscience is first formulated into IQA for better representing structural edges of video frames. Second, the RTV accounting for separating orderly and disorderly regions is employed to implement the free-energy principle due to its low computational complexity. Thirdly, the HVSP accounts for the visual speed perception of video signal. Since FePVQ is closely related with MSE, it can be easily integrated into a RDO process of hybrid video coding for the purpose of perceptual video coding optimization. The remarkable merit of this work over the previous algorithms is to introduce free-energy principle and HVSP into VQA, contributing a better representation of orderly and disorderly signals in visual scenes, and therefore a better performance.

References

[1] C. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatio-temporal model of the human visual system," in *Proc. SPIE*, vol. 2668, 1996, pp. 450–461.

[2] S. Winkler, "A perceptual distortion metric for digital color video," in *Proc. SPIE*, Jan. 1999, pp. 175–184.

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

[4] A. Watson, J. Hu, and J. McGowan, "Digital video quality metric based on human vision," *J. Electron. Imag.*, vol. 10, no. 1, pp. 20–29, Jan. 2001.

[5] 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.

[6] J. You, T. Ebrahimi, and A. Perkis, "Modeling motion visual perception for video quality assessment," in *Proc. ACM Multimedia*, 2011, pp. 1293–1296.

[7] 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.

[8] Z. K. Lu, W. Lin, X. K. Yang, E. P. Ong, and S. S. Yao, "Modeling visual attention's modulatory afterreffects on visual sensitivity and quality evaluation," *IEEE Trans. Image Process.*, vol. 14, pp. 1928–1942, Nov. 2005.

[9] Z. Wang, H. R. Sheikh, A. C. Bovik, 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.

[10] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. 37th IEEE Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, pp. 1398–1402.

[11] A. K. Moorthy and A. C. Bovik, "Efficient video quality assessment along temporal trajectories," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 20, no. 11, pp. 1653–1753, Nov. 2010.

[12] M. Narwaria, W. S. Lin, and A. M. Liu, "Low-complexity video quality assessment using temporal quality variations," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 525–535, Jun. 2012.

[13] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, and S. Li, "Incremental support vector learning for ordinal regression," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 26, no. 7, pp. 1403–1416, Jul. 2015.

[14] B. Gu, V. S. Sheng, Z. Wang, D. Ho, S. Osman, and S. Li, "Incremental learning for γ-support vector regression," *Neural Netw.*, vol. 67, pp. 140–150, 2015.

[15] K. Friston, J. Kilner, and L. Harrison, "A free energy principle for the brain," *J. Physiol.*, vol. 100, nos. 1–3, pp. 70–87, 2006.

[16] K. Friston, "The free-energy principle: A unified brain theory?," *Nature Rev. Neurosci.*, vol. 11, no. 2, pp. 127–138, 2010.

[17] J. Wu, G. Shi, W. Lin, A. Liu, and F. Qi, "Just noticeable difference estimation for images with free-energy principle," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1705–1710, Nov. 2013.

[18] G. T. Zhai, X. L. Wu, X. K. Yang, W. S. Lin, and W J Zhang, "A psychovisual quality metric in free-energy principle," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 41–52, Jan. 2012.

[19] D. C. Knill and R. Pouget, "The Bayesian brain: The role of uncertainty in neural coding and computation," *Trends Neurosci.*, vol. 27, pp. 712–719, 2004.

[20] Z. Q. Pan, Y. Zhang, and S. Kwong, "Efficient motion and disparity estimation optimization for low complexity multiview video coding," *IEEE Trans. Broadcast.*, vol. 61, no. 2, pp. 166–176, Jun. 2015.

[21] A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception," *Nature Neurosci.*, vol. 9, pp. 578–585, 2006.

[22] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Amer. A, Opt. Image Sci. Vis.*, vol. 24, no. 12, pp. B61–B69, Dec. 2007.

[23] E. P. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Ann. Rev. Neurosci.*, vol. 24, pp. 1193–1216, 2001.

[24] L. Xu, Q. Yan, Y. Xia, and J. Y. Jia, "Structure extraction from texture via relative total variation," *ACM Trans. Graphics*, vol. 32, no. 6, Nov. 2012, Art. no. 139.

[25] A. Bhat, S. Kannangara, Y. F. 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.

[26] Y.-F. Ou, Z. Ma, and Y. Wang, "Perceptual quality assessment of video considering both frame rate and quantization artifacts," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 3, pp. 286–298, Mar. 2010.

[27] Z. Wang, L. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Process.*, vol. 12, no. 12, pp. 243–254, Feb. 2003.

[28] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Phys.*, vol. 46, no. 5, pp. 323–352, Sep. 2005.

[29] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, 2009.

[30] L. Xu, X. Y. Ji, W. Gao, and D. B. Zhao, "Laplacian distortion model (LDM) for rate control in video coding," in *Proc. Multimedia 8th Pacific Rim Conf. Adv. Multimedia Inf. Process.*, Dec. 2007, pp. 638–646.

[31] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 74–90, Nov. 1998.

[32] L. Xu, D. Zhao, X. Ji, L. Deng, S. Kwong, and W. Gao, "Window level rate control for smooth visual quality and smooth buffer occupancy," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 723–734, Mar. 2011.

[33] S. Q. Wang, A. Rehman, Z. Wang, S. W. Ma, and W. Gao, "SSIM-motivated rate-distortion optimization for video coding," *IEEE Trans. Circuits Syst. Video Techn*. vol. 22, no. 4, pp. 516–529, Apr. 2012.

[34] L. Xu, S. N. 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.

[35] The Video Quality Experts Group, "Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment," 2000. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv phaseI

[36] H. R. Sheikh, Z. Wang, L. Cormack and A. C. Bovik, "LIVE image quality assessment database release 2," 2005. [Online]. Available: http://live.ece.utexas.edu/research/quality

[37] N. Ponomarenko *et al.*, "TID2008-A database for evaluation of full-reference visual quality assessment metrics," *Adv. Modern Radioelectron.*, vol. 10, pp. 30–45, 2009.

[38] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, Mar. 2010, Art. no. 011006.

[39] 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.

[40] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[41] X. D. Zhang, X. C. Feng, W. W. Wang, and W. F. Xue, "Edge strength similarity for image quality assessment," *IEEE Signal Process. Lett*. vol. 20, no. 4, pp. 319–322, Apr. 2013.

[42] K. Seshadrinathan, R. Soundararajan, A. C. Bovik and L. K. Cormack, "LIVE video quality database," 2009. [Online]. Available: http://live.ece.utexas.edu/research/quality/live_video.html

[43] F. Zhang, S. Li, L. Ma, and K. N. Ngan, "IVP video quality database," 2011. [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/index.shtml.

[44] P. V. Vu and D. M. Chandler, "ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imag.*, vol. 23, no. 1, pp. 13–16, Feb. 2014.

[45] 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.

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

[47] F. Bellard *et al.*, “FFMPEG tool,” Nov. 15, 2012. [Online]. Available: http://www.ffmpeg.org

[48] M. 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.

[49] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[50] “Methodology for the subjective assessment of the quality of television pictures,” ITU-R Rec. BT.500-11, ITU, Geneva, Switzerland, 2002.

[51] “Subjective video quality assessment methods for multimedia applications,” ITU-T Rec. P.910, ITU, Geneva, Switzerland, 2008.

Long Xu (M’13) received the M.S. degree in applied mathematics from Xidian University, Xi’an, China, in 2002, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

He was a Postdoc with the Department of Computer Science, City University of Hong Kong, Hong Kong, China, and the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, China, and from July 2009 to December 2012. From January 2013 to March 2014, he was a Postdoc with the School of Computer Engineering, Nanyang Technological University, Singapore. He is currently with the Key Laboratory of Solar Activity, National Astronomical Observatories, Chinese Academy of Sciences. His research interests include image/video processing, solar radio astronomy, wavelet, machine learning, and computer vision.

Dr. Xu selected into the 100-Talents Plan, Chinese Academy of Sciences, 2014.

Weisi Lin (S’91–M’92–SM’00–F’16) received the Ph.D. degree from King’s College, London University, London, U.K.

He was previously the Lab Head of Visual Processing, Institute for Infocomm Research, Singapore. He is currently an Associate Professor with the School of Computer Engineering, Institute for Infocomm Research. His research interests include image processing, perceptual signal modeling, video compression, and multimedia communication, in which he has authored or coauthored more than 130 journal papers and 200 conference papers, filed seven patents, and authored two books.

Prof. Lin is a Fellow of the IET. He is an Associate Editor for the IEEE Transaction Image Processing, the IEEE Signal Processing Letters, and the *Journal of Visual Communication and Image Representation*, and a past Associate Editor for the IEEE Transaction Multimedia. He has also been a Guest Editor for eight special issues in international journals. He has been a Technical Program Chair for IEEE ICME 2013, PCM 2012, and QoMEX 2014. He chaired the IEEE MMTC Special Interest Group on QoE (2012–2014). He has been an invited/panelist/keynote/tutorial speaker in more than ten international conferences, as well as a Distinguished Lecturer of Asia-Pacific Signal and Information Processing Association (2012–2013).

Lin Ma (M’13) received the B.E. and M.E. degrees from the Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively, and the Ph.D. degree from the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, China, in 2013.

He is currently a Researcher with Huawei Noah’s Ark Lab, Hong Kong, China. He was a Research Intern with Microsoft Research Asia, Beijing, China, from October 2007 to March 2008. He was a Visiting Student with the School of Computer Engineering, Nanyang Technological University, Singapore, from July 2011 to September 2011. His research interests include the areas of deep learning and multimodal learning, specifically for image and language, image/video processing, and quality assessment.

Yongbing Zhang (M’13) received the B.A. degree in english and the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2004, 2006, and 2010, respectively.

He is currently an Associate Professor with the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China. His current research interests include video processing, image and video coding, video streaming, and transmission.

Prof. Zhang was the recipient of the Best Student Paper Award at IEEE International Conference on Visual Communication and Image Processing in 2015.

Yuming Fang (M’13) received the B.E. degree from Sichuan University, Chengdu, China, the M.S. degree from Beijing University of Technology, Beijing, China, and the Ph.D. degree in computer engineering from Nanyang Technological University, Singapore.

He is currently an Associate Professor with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. He was also a Visiting Postdoc Research Fellow with the IRCCyN laboratory, PolyTech’ Nantes & University Nantes, Nantes, France, the University of Waterloo, Waterloo, ON, Canada, and Nanyang Technological University, Singapore. His research interests include visual attention modeling, visual quality assessment, image retargeting, computer vision, and 3-D image/video processing.

Prof. Fang was a Secretary of the Ninth Joint Conference on Harmonious Human Machine Environment (HHME2013). He was also a Special Session Organizer at VCIP 2013 and QoMEX 2014.

King Ngi Ngan (M’79–SM’91–F’00) received the Ph.D. degree in electrical engineering from the Loughborough University, Loughborough, U.K.

He is currently a Chair Professor with the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, China. He was previously a Full Professor with the Nanyang Technological University, Singapore, and the University of Western Australia, Crawly, WA, Australia. Since 2012, he has been a Chair Professor at the University of Electronic Science and Technology, Chengdu, China, under the National Thousand Talents Program. He holds honorary and visiting professorships at numerous universities in China, Australia, and South East Asia. He has published extensively, including three authored books, seven edited volumes, and more than 370 refereed technical papers, and he has edited nine special issues in journals. He holds 15 patents in the areas of image/video coding and communications.

Prof. Ngan is a Fellow of the IET (U.K.), and IEAust (Australia), and an IEEE Distinguished Lecturer for 2006–2007. He was an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology, the *Journal on Visual Communications and Image Representation*, the *EURASIP Journal of Signal Processing: Image Communication*, and the *Journal of Applied Signal Processing*. He chaired and co-chaired a number of prestigious international conferences on image and video processing including the 2010 IEEE International Conference on Image Processing, and served on the advisory and technical committees of numerous professional organizations.

Songnan Li (S'10–M'12) received the B.S. and M.P. degrees in computer science and technology from the Harbin Institute of Technology, Beijing, China, in 2004 and 2006, respectively, and the Ph.D. degree from the Department of Electronic Engineering, Chinese University of Hong Kong (CUHK), Hong Kong, China, in 2012.

He joined the CUHK as a Research Assistant in 2007. From 2012 to 2014, he was a Postdoc with the Department of Electronic Engineering, CUHK. He is currently a Research Assistant Professor in the same department. His research interests include image and video processing, RGBD computer vision, and visual quality assessment.

Yihua Yan received the B.E and M.E. degrees from Northwestern Polytechnical University, Xi'an, China, in 1982 and 1985, respectively, and the Ph.D. degree from the Dalian University of Technology, Dalian, China, in 1990.

He is currently a Professor and Chief Scientist of Solar Radio Research and the Director of the Key Laboratory of Solar Activity and the Solar Physics Division, National Astronomical Observatories, Science Academy of Sciences, Beijing, China. He was a Foreign Research Fellow. He is currently the President of the International Astronomical Union Division E: Sun & Heliosphere, Science Academy of Sciences, from 2015 to 2018 with the National Astronomical Observatory of Japan, Tokyo, Japan, from 1995 to 1996, and an Alexander von Humboldt Fellow with the Astronomical Institute, Wurzburg University, Wurzburg, Germany, from 1996 to 1997. His research interests include solar magnetic fields, solar radio astronomy, space solar physics, and radio astronomical methods.