Consistent Visual Quality Control in Video Coding

Long Xu, Songnan Li, King Ngi Ngan, *Fellow, IEEE,* and Lin Ma

***Abstract*—Visual quality consistency is one of the most important issues in video quality assessment. When people view a sequential video, they may have an unpleasant perceptual experience if the video has an inconsistent visual quality even though the average visual quality of the video is not compromised. Thus, consistent visual quality control is mostly expected in general video encoding with limited channel bandwidth and buffer resources. However, there still has not been enough study on such an issue. In this paper, a new objective visual quality metric (VQM) is proposed first, which can easily be incorporated into video coding for guiding video coding. Second, a VQM-based window model is proposed to handle the tradeoff between visual quality consistency and buffer constraint in video coding. Third, a window-level rate control algorithm is developed to accomplish visual quality control based on the above two proposals. Finally, experimental results prove that consistent visual quality, high rate-distortion efficiency, accurate bit control, and compliant buffer constraint can be achieved by the proposed rate control algorithm.**

***Index Terms*—Consistent visual quality, video coding, video quality assessment (VQA), window model, window-level rate control.**

I. Introduction

IN RECENT years, there has been increased interest in image quality assessment (IQA) and video quality assessment (VQA) that measure the perceptual visual qualities of images and videos. Since humans are the ultimate receivers of the visual signal, the most accurate way of assessing image/video quality is to ask humans for their opinions of the quality of an image or video, which is known as the subjective visual quality assessment. Researchers need to perform their subjective experiments to validate the proposed objective/automatic visual quality metrics (VQMs). The subjective scores are first gathered from subjective experiments. Then, the correlation between the subjective scores and objective values is analyzed to evaluate whether the proposed objective metrics are good at measuring the human perception of image/video quality. The methods of measuring such a correlation include Pearson's correlation coefficient, Spearman's rank order correlation coefficient (SROCC), and outlier ratio (OR). According to the recommendation of the International Telecommunication Union (ITU) [1], [2], a number of viewers are asked to rate the images or videos in the subjective experiments, and the scores of the viewers are processed as the mean opinion score (MOS) or the difference mean opinion score (DMOS). The subjective experiment is actually fundamental to rank a proposed objective metric. In [1] and [2], the comparative studies of objective video quality metrics are carried out by Video Quality Expert Group (VQEG). The subjective experiments have been performed on a large number of test video sequences by many laboratories and researchers. In [3] and [4], the quality degradation of video streaming was investigated to target at a good perceptual quality of multimedia services. The publicly available databases of subjective scores and test material were reported in [5]–[7] for quality degradation of compression and error-prone channels. Most researchers developed their objective metrics based on these available subjective quality databases.

The objective quality metric aims at automatically predicting human perceptual behavior in evaluating image or video quality. It is convenient and computationally efficient in real-world applications. Traditionally, mean square error (MSE)/peak signal-to-noise ratio (PSNR) were used to evaluate image or video quality. Almost all image/video compression standards use MSE/PSNR to measure the objective quality of the compressed signal. However, MSE/PSNR does not correlate well with the human visual system (HVS) [8]. Thus, a host of image/video quality metrics have been proposed in the last decade.

These quality metrics can be generally categorized into two classes [9]. One focused on psychologically modeling the human perception of the visual signal. The signal was decomposed into multiple channels to simulate the tuning properties of the HVS. Luminance masking, contrast sensitivity function, and contrast masking models are typically used to obtain visibility thresholds for each channel. The state-of-the-art HVS-model-based video metrics, such as moving pictures quality metric [10], perceptual distortion metric [11], and the Sarnoff just noticeable difference [12] vision model, filter the videos using one band-pass and one low pass filter along the temporal dimension. The digital video quality metric [13] and the scalable wavelet based video distortion index [14] utilize a single low pass filter along the temporal dimension. In [15], a VQM is developed based on the spatio-temporal distortions through a temporal analysis of spatial perceptual distortion maps.

The other class of quality metrics is based on image structural similarity. There exist strong relations between pixels in

Manuscript received April 11, 2012; revised August 5, 2012 and October 27, 2012; accepted December 12, 2012. Date of publication January 29, 2013; date of current version May 31, 2013. This work was supported in part by the Research Grants Council of Hong Kong, China, under Grant CUHK415712, and the National Natural Science Foundation of China under Grant 61202242. This paper was recommended by Associate Editor R. C. Lancini.

L. Xu is with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: lxu@jdl.ac.cn).

S. N. Li, K. N. Ngan, and L. Ma are with the Chinese University of Hong Kong, Shatin, Hong Kong (e-mail: snli@ee.cuhk.edu.hk; knngan@ee.cuhk.edu.hk; lma@ee.cuhk.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.2013.2243657

nature images that show the obvious structural information to the HVS. Since the HVS is highly adapted for extracting structural information from a video scene, a measure of the structural similarity can provide a good approximation to the perceived image quality. Wang *et al.* [16], [17] proposed an objective metric called structural similarity index (SSIM) to rank the image quality, which was further extended into multiple-scale SSIM (MS-SSIM) [18] to adapt to the variable viewing conditions. In [19], a statistical model of human motion perception [20], [21] was used to perceptually weigh the spatial and temporal pooling processes. Moorthy *et al.* [22] investigated video quality assessment for the scenario of wireless video communication. In [23], a computationally efficient motion compensated SSIM (MC-SSIM) along temporal trajectory was proposed for video quality assessment. In [9], the motion-based video integrity evaluation (MOVIE) was proposed based on the fact that the middle temporal (MT) visual area in the HVS is critical for human visual perception [24]. The National Telecommunications and Information Administration (NTIA) developed a general video quality metric [25] for quantifying perceptual quality degradation of video compression, which became the top performer in the VQEG phase II video quality study [2]. Most recently, a good guide to quality assessment was found in [26], where the state-of-the-art quality assessment algorithms were analyzed intensively and ranked in terms of correlation with subjective scores.

In IQA, the spatial distortions, e.g., block artifacts, blurring, mosaic, ringing, etc., have been investigated. The temporal distortions include motion compensation mismatches, mosquito effects, ghosting, smearing, temporal fluctuation of picture quality, etc. Hence, VQA needs to study them additionally for assessing video sequence quality. Yuen and Wu [27] summarized the spatial and temporal distortions in hybrid motion compensated (MC) and block-based discrete cosine transform (DCT) coding standards. In video coding, the consistent/smooth visual quality is essential to the integral quality of a video sequence. First, temporal quality fluctuation is an annoying experience to human visual perception. In addition, the inconsistent visual quality of video frames would result in many image distortions, including block artifacts, mosaic, and so on in general real-time video communications, since the channel bandwidth and buffer resources are limited. The overuse of the bits quote at some frames will result in the serious quality degradation of the other frames. The consistent visual quality control under conventional constant bit rate (CBR) encoding environment concerns the usual applications of video coding. It has been studied previously in [28] and [29]. He *et al.* [28] tried to provide the smooth quantization under CBR encoding by introducing a low filtering mechanism to smooth the quantization parameters (QPs) produced from the traditional rate control algorithm. In [29], a sequential rate control algorithm was proposed for real-time video coding. However, the MSE that measured visual quality in both [28] and [29] was widely criticized for not correlating well with the perceived quality. In this paper, a new objective VQM is first proposed. Both quantization and temporal motion information are included to ensure the friendliness of the metric to video coding application. Then, a window model and a window-level rate control algorithm are developed based on the proposed VQM.

The rest of this paper is organized as follows. In Section II, a new objective VQM is proposed to assess compressed video signals. In Section III, a window model is constructed based on the proposed VQM. Section IV proposes a VQM-based rate-distortion (R-D) model and a window-level rate control algorithm for consistent visual quality control in video coding. Section V shows the experimental results for evaluating the efficiencies of the proposed VQM and rate control algorithm. Finally, a brief conclusion is given in the last section.

II. Proposed VQM for Video Coding

VQA needs to further exploit temporal characteristic of HVS compared with IQA. There are three major aspects to explore temporal HVS characteristics in the literature. The first one decomposes the video signal into multiple spatial-temporal frequency channels and assigns a different weight to each of them [9], [33]. The second one simulates the visual masking that is another visual phenomenon critical for video quality assessment. Considering the masking effect, the visibility of video distortion not only concerns spatial activity but also the temporal activity. The spatial masking effect was investigated elaborately in the past research [9], [12]–[14]. However, the temporal masking effect has not been thoroughly studied. The last one involves the high-level characteristics of the HVS in the pooling process. The pooling process is believed to be capable of simulating the late stage of the visual pathway, where all visual information is spatially and temporally integrated for quality evaluation of the entire image or video [19], [33], [35].

To develop a VQM for evaluating visual quality of image/video signal, guiding video encoding and cooperating with other components of a video encoder, it needs to be easily optimized to obtain a close-form analytic solution for a specific video encoding task. In [31], to ensure the friendliness of the quality metric to quantization, MSE is simply weighed to simulate the HVS responses to the visual signals. In [32], the quantization artifact and frame rate were introduced into the visual quality assessment for scalable video coding. In [19], the motion information content and the perceptual uncertainty were analyzed for VQA. In [35], a VQM measuring temporal variations of spatial visual distortions was developed based on the spatio-temporal distortion evaluation, using both eye-fixation-level and long-term temporal pooling.

In this paper, a new VQM is proposed based on pioneering work in [19] and [31]. The motivations of the proposed metric are first to utilize the available information of video encoding without extra computations, second to be incorporated into rate control component to monitor and guide video encoding timely and dynamically. Referring to [31], the MSE was weighted to simulate HVS response as

$$VQM_m = 1 - k_m \times MSE_m \quad (m = 0, 1, ..., M-1) \qquad (1)$$

where k_m is a weighting factor for the mth MB, VQM_m is the visual quality level of the mth MB, and M is the number

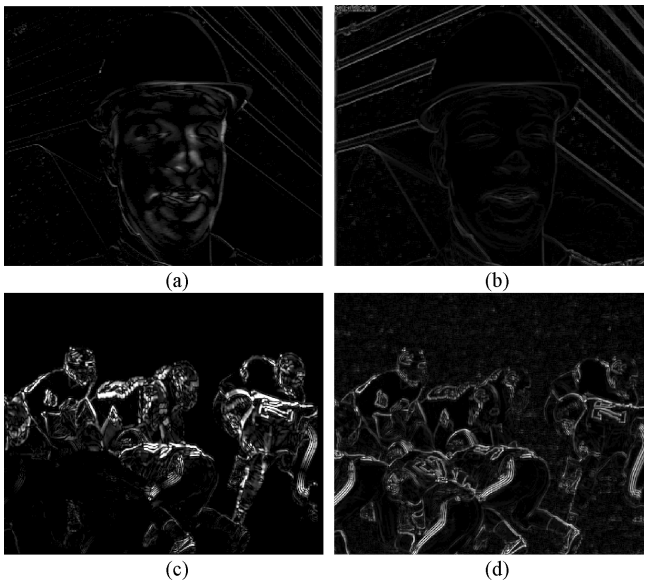

Fig. 1. Frame edges. (a), (c) Obtained by the temporal disparity between two consecutive frames. (b), (d) Obtained by the disparity of two pixels within a frame.

of MBs in a frame. Since the MSE change of low-detailed videos contributes more visual quality change than that of high-detailed videos, k_m serves as a weighting factor in (1). In [31], k was computed by the spatial edge strength, so edge detection was performed and the computational efficiency was compromised. In this paper, the temporal disparity between two consecutive frames is first computed. Second, two spatial disparities $I(x,y) - I(x - 2,y)$ and $I(x,y) - I(x, y - 2)$ are calculated along horizontal and vertical directions, respectively, within a frame $I(x,y)$. The sum of squares of the two spatial disparity images produces a spatial disparity image. The spatial and temporal disparity images are shown in Fig. 1 for *Foreman* and *Football*, respectively. Their combination produces a tridimensional disparity, which is used to compute the edge strength on MB basis as

$$k_m = \sum_{(x,y)\in \mathrm{MB}} |S(x, y)| + |T(x, y)| \tag{2}$$

where $S(x,y)$ and $T(x,y)$ represent the spatial and temporal disparity images, respectively. Equation (2) is computationally efficient in comparison to [31].

In [19], a motion information content and perception uncertainty were defined and used to assess video quality. A spatio-temporal weighting function was defined as

$$w = (\alpha \log v_r + \beta) - (\log v_g - \gamma \log c + \delta) \tag{3}$$

where v_r, v_g are the relative and global motion, c represents the stimulus contrast, and α, β, γ, and δ are the constant model parameters. The first component of the right part of (3) measures the motion information content, and the second component represents the perception uncertainty. The model indicates that an object with significant motion with respect to the background would be a strong surprise to the visual system. However, when the background motion is too large, the HVS cannot identify the objects with the same accuracy as in the statistic background. In this paper, a motion activity term is defined as

$$w_m = \log \sqrt{\frac{\sum_{i,j} v_x(i, j)^2 + v_y(i, j)^2}{d(i, j)}} \tag{4}$$

to measure the motion information content, where w_m is for the mth MB, (v_x, v_y) is the motion vector of a 4×4 block in an MB, and $d(i, j)$ is the distance from current frame to its reference. Compared to (3), only relative motion is considered in (4) for the purpose of little overhead computation. Considering both (1) and (4), the local MSE is adapted by both spatial and temporal factors to have a new VQM as

$$VQM_m = w_m \times (1 - k_m MSE_m). \tag{5}$$

The sum of VQMs over all MBs will come up with the frame VQM as

$$VQM_f = \sum_{m=0}^{M-1} w_m \times (1 - k_m MSE_m). \tag{6}$$

Similarly, the VQM of entire sequence is the sum of VQMs over all frames as

$$VQM = \sum_{f=0}^{L-1} \left(\sum_{m=0}^{M-1} w_m \times (1 - k_m MSE_m) \right). \tag{7}$$

In the practical usages of (5)–(7), w_m, k_m, and MSE_m are all normalized accordingly.

III. Proposed VQM-Based Window Model

Rate control consists of two sequential steps: bit allocation and QP decision. Considering buffer status in bit allocation, the coded bitstream conforms to both the target bit rate and buffer constraint. Rate control should guarantee a small gap between the target bit rate and the actual coded bit rate, and achieve good R-D performance. On the other hand, the picture quality smoothness and buffer compliance also play key roles in rate control. Generally, the consistent visual quality of encoded frames can give viewers a temporally consistent, and thus, comfortable visual experience. The consistent visual quality can be obtained by two-pass or multi-pass encoding [37], [38]. The compliant buffer constraint guarantees the successful transmission and decoding of bitstream under the given conditions of a video communication system. In video streaming applications, violating buffer constraint may cause annoying jitter.

In the traditional rate control algorithms, the anticipated complexity and mean absolute distortion (MAD) of future frames are used for bit allocation and QP calculation, respectively [39], [40]. These methods work well for the low motion object and similar scenes among the adjacent frames. However, they may result in the significant picture quality fluctuation or buffer violation when scene changes or high motion occurs. Given encoding conditions, an inverse relationship between picture quality change and buffer variation usually exists due to the nonstationary input signal of video coding. Thus, a tradeoff between them is worth studying in rate control.

To tackle the problem mentioned above, a theoretical window model was proposed in [41], in which two bounds of QP variation and frame bits variation were related to window size. The window size means the number of frames in a window. The picture quality smoothness was measured by QP variation, and the buffer fullness was computed from the frame bits output. The proposed window model tries to seek the relationship among window size, QP variation, and buffer size to get a good tradeoff between picture quality smoothness and buffer constraint. To derive the theoretical model, the frame bits variation is used instead of buffer constraint because there is an inherent relation between frame bits variation and buffer size given the channel bit rate. Actually, the buffer size is computed from the accumulated bits of the output frame bits minus the given channel bit rate per frame. Thus, the less the variation of output frame bits, the less the buffer size is. In this paper, the window model is updated by using buffer size directly. In addition, the proposed VQM is imported into the window model to measure visual quality smoothness instead of using the QP variation. The proposed VQM is expected to be better than QP for measuring visual quality smoothness under both single scene and multiple scenes. Assume window size is L, VQM bound is ΔV_Q and buffer constraint is B, the relation between L and ΔV_Q is investigated using the law of large numbers (LLN) [42], [43], where the number of samples and error bound correspond to L and ΔV_Q, respectively.

A. $L - \Delta V_Q$ Model

Suppose that the bound of picture quality variation is ΔV_Q represented by a Gaussian random variable $\xi(\omega)(\mu_\xi = 0, \sigma_\xi^2)$. According to central limit theory [43], the average of $\{\xi_k(\omega)\}(k = 1, 2, \cdots, n)$ converges at a Gaussian random variable, that is

$$P\left(\left|\frac{1}{n}\sum_{k=1}^{n}\xi_k - \mu_\xi\right| < \Delta V_Q\right) = \frac{1}{\sqrt{2\pi}}\int_{-\Delta V_Q(\sqrt{n}/\sigma_\xi)}^{\Delta V_Q(\sqrt{n}/\sigma_\xi)} e^{-\frac{t^2}{2}}dt. \tag{8}$$

Let the right-hand size of (8) be equal to p, and the integral limit of (8) be obtained by looking up the standardized normal table [43]. Assume $\Delta V_Q(\sqrt{n}/\sigma_\xi)=\sqrt{\alpha}$, replacing n by L, then the L–ΔV_Q model can be derived as

$$L = \alpha\frac{\sigma_\xi^2}{\Delta V_Q^2} \tag{9}$$

where L is the minimum window size that conforms to ΔV_Q. σ_ξ^2 is regarded as a constant, and merged into α in practical usage.

B. L-B model

He *et al.* [47] proved that the encoder buffer of the proposed rate control is subject to

$$|B_e(n)| \le \frac{1}{2}\Theta\sum_{j=1}^{M} a_j \cdot j \tag{10}$$

where Be (n) represents the buffer occupancy for the nth frame, Θ is such a constant that $\left|\log_2 \sigma^2(i)\right| \le \Theta$ holds true for all frames, where $\sigma^2(i)$ is the residue variance of the ith frame after intra or inter prediction. $\{\alpha_i \text{ s.t. } \alpha_i > 0, \sum_{i=1}^{M}\alpha_i = 1\}$ is an M-tap averaging filter for smoothing the distortion of frame encodings. Let $W_{\max} = \Theta\sum_{j=1}^{M} a_j \cdot j$; we have $|B_e(n)| \le \frac{1}{2}W_{\max}$. It can be regarded as the maximum encoder buffer that is needed for video streaming usage of the rate control of [28]. In particular, if an arithmetic averaging filter is employed in (10), we have

$$W_{\max} = \frac{L+1}{2}\Theta \tag{11}$$

which suggests that the larger the filter length (L) and variation of scene activity (Θ), the larger encoder buffer is needed for the video streaming. Obviously, the above filter length has the same functionality as the window size of our proposal. The parameter Θ is video content dependent, which is dynamically updated during encoding.

According to (11), the buffer constraint is directly related to the window size. There is no need to map buffer constraint to bits fluctuation, as proposed in [41]. In addition, we need to control the encoder buffer for the compliant buffer constraint, so more discussions about the L–B model are given as follows. Assume the encoder buffer occupancy of jth frame is $B_e(j)$, which is dependent on the initial buffer status of the start of the window, i.e., the $(j - L)$th frame. The frame encoding bits $\{R(i)\}$ and channel throughput $\{C(i)\}$ as

$$B_e(j) = B_e(j - L) + \sum_{i=j-L}^{j-1}(R(i) - C(i)). \tag{12}$$

$C(i) = min\{R_T, B_e(i)\}$ indicates that the channel throughput is not always at its payload capability, i.e., the given target bit rate R_T, which is the result of CAT constraint on an encoder buffer [44], [45]. $R(i)$ is calculated theoretically by the classical R-D model $R = \log_2\frac{\sigma}{D}$. Importing it into (12), the buffer variation can be finally expressed by

$$\begin{aligned} &B_e(j) - B_e(j - L)\\ &\ge L \times \left(\log_2\left(\prod_{i=j-L}^{j-1}\sigma(i)\right)^{\frac{1}{L}} - \log_2\left(\frac{1}{L}\sum_{i=j-L}^{j-1}\sigma^2(i)\right)^{\frac{1}{2}}\right). \end{aligned} \tag{13}$$

The left-hand side of (13) will approach 0 if $\sigma(k) = \sigma(l)$ $(k \neq l)$, which indicates that the buffer level will not increase.

From (12) and (13), we can conclude that the uniform frame bits output would result in a small buffer increase because the payload capability of the channel can be fully occupied at this situation. In addition, the situation of uniform frame bits exists if the characteristics of residual signals, such as variances, do not have large differences.

C. Window Model

Obviously, the smallest possible visual quality fluctuation under the given buffer constraint is preferred in video coding. However, small variations of visual quality and buffer usage cannot be achieved concurrently due to a contradiction between them. In this paper, the bounds of picture quality variation and buffer constraint (ΔV_Q and B) are assumed

to be two encoding requirements. They are contradictory to each other, i.e., the buffer occupancy will arise generally if a small ΔV_Q is required, and vice versa. According to (12), the buffer occupancy is determined by the frame bits and channel throughput. It can be adjusted by changing the encoding bits of each frame.

We first get the minimum window size L from the L–ΔV_Q model (9). Then, if $W_{\max} \leq B$, L from the L-ΔV_Q model will be the final solution; otherwise, L should be further regulated to guarantee that the derived buffer $W_{\max}$ from (11) is less than B. Usually, nature videos are non-stationary, so the video content varies significantly along time. Thus, the frame bits output would fluctuate a lot if the smooth QPs are assigned to a window. In such a situation, the larger L is, the larger B is, and vice versa. Therefore, L should be decreased if $W_{\max} > B$. Using B instead of $W_{\max}$ in (11), then we can obtain $B = \frac{L+1}{2}\Theta$. Integrating (9), a window model is proposed as

$$L = \begin{cases} \dfrac{\alpha}{\Delta V_Q^2}, & \text{if } W_{\max} \leq B \\ \dfrac{2 \times \beta \times B}{\Theta} - 1, & \text{else} \end{cases} \tag{14}$$

where α and β are two parameters of the window model, and they are adaptively updated for each window.

IV. Proposed Window-level Rate Control

The proposed window model tells us that the best tradeoff between visual quality smoothness and buffer constraint can be achieved in theory. To get such a tradeoff in practice, the window-level rate control algorithms should be provided additionally. A window-level rate control utilizes any existing R-D model such as linear R-D model to calculate QP. In addition, it employs window model as the top level control to decide window size. From [46], the R-D relation usually was modeled as a LOG function

$$R = \log_2 \frac{\sigma}{D} \tag{15}$$

and the theoretical distortion model was given by

$$D = \frac{Q^2}{12}. \tag{16}$$

They were both based on the assumption of the Laplacian distribution of DCT coefficients. σ represents the standard deviation of input signal, which is replaced by MAD in practice. Equation (15) was usually approximated by a linear or quadratic R-D model [39], [40]. The linear R-D model is written as

$$Q = \frac{a \times MAD}{R} \tag{17}$$

where a is a constant, Q represents QP, and R indicates bits quote of a coding basis unit, e.g., an MB, a frame or a window. MAD is derived from the pre-analysis as to be introduced below.

A. Proposed VQM-Based R-D Model

To achieve window-level rate control, (17) is extended to window level as

$$Q_f = \frac{a \times \sum_{j=f}^{L-1} MAD_j}{T - \sum_{j=0}^{f-1} R_j} \tag{18}$$

where Q_f is the QP of the fth frame, and MAD_f and R_f are the MAD and bits usage of the fth frame, respectively. T is the total bits allocated to a window. Equation (18) can also be used at MB level rate control. From (18), QP is computed for each frame of the window. In addition, the remaining bits quotes and MADs are updated after encoding a frame. Such a method could provide a frame-level smooth visual quality profile. In addition, the bits control is accurate because (18) is implemented progressively. Furthermore, there is no bit allocation operation in (18), which usually results in bad bits usage and bad visual quality as in the traditional rate control.

The distortion D in (16) is measured by MSE in practice. Theoretically, all coding basic units should have the same MSE if a constant QP is employed according to (16). From Section II, a new VQM is defined in (5) to measure the visual quality. Importing (16) into (5) will generate a new theoretical model as

$$VQM_m = w_m \times (1 - k_m Q_m^2). \tag{19}$$

Accompanied by (18), (19) can realize visual quality control. First, (18) calculates a QP for the mth coding unit. Second, (19) provides a visual quality map $\{p_m\}$ to all coding units (MBs or frames). Then, the QP of each coding unit is revised accordingly by

$$\begin{cases} Q_m = Q_m \times p_m \\ p_m = w_m \times (1 - k_m Q_f^2) \end{cases} \tag{20}$$

where Q_f is a frame-level QP if (20) is used at the MB level. p_m is actually adaptive to both video content and bit rate, so it is more suitable for video compression. If (20) is used at the frame level, all the parameters in (20) are transferred into the corresponding forms at frame level. Then, we can come up with a new form of (20) as

$$\begin{cases} Q_f = Q \times P_f \\ P_f = \sum_{m=0}^{M-1} p_m \end{cases} \tag{21}$$

which revises frame-level QPs in the sense of consistent visual quality at frame level. Thus, (21) cooperated with (18) will construct the proposed VQM-based R-D model. In addition, the proposed R-D model can be used in both frame-level and window-level rate controls.

B. Pre-Analysis

A pre-analysis process is introduced to provide MAD information in (18). In the pre-analysis, only 16×16 inter mode motion estimation (ME) is used to all frames of a window for

the computational efficiency. In addition, only one reference is used in the ME stage of pre-analysis. After pre-analysis, edge strength $\{k_m\}$ and motion activity $\{w_m\}$ of all MBs can be computed by using (2) and (4), respectively. In pre-analysis, a constant QP is used to all frames.

A frame VQM profile can be obtained approximately after pre-analysis according to (19). The frame QPs can be determined by employing (18) at the window level. Then, the QPs from (18) are regulated to get a smooth VQM profile. The pre-analysis has two functions. The one is to provide $\{MAD_j\}$ for QP calculation described in (18). The other is to produce visual quality maps $\{p_m\}$ and $\{P_j\}$, which are used for visual quality control as described in (20) and (21).

C. Remarks on the Calculation of Residual Signal Variance

In the proposed window model (14), Θ should be updated adaptively along with the video content change. As defined in [28], Θ is the maximum of all $\log_2 \sigma^2\ (j)$, where $\sigma^2(j)$ is the variances of the jth residual frame. In order to adapt to the bit rate or video quality of video compression, the variance of residual signal is computed on the quantized residual signal instead of the original residual signal. The original residual signal $s(j)$ is given by

$$s(j) = x(j) - r(j-1) \tag{22}$$

where $x(j)$ represents the original input signal, and $r(j-1)$ represents its reference. Assume the quantized $s(j)$ is $\hat{s}(j)$. The reconstructed jth frame can be written as

$$r(j) = r(j-1) + \hat{s}(j). \tag{23}$$

Thus, the quantized residual signal $\hat{s}(j)$ is actually the difference between $r(j)$ and $r(j-1)$, that is

$$\hat{s}(j) = r(j) - r(j-1). \tag{24}$$

Then, the variance of the quantized residual signal for each MB is calculated by

$$\sigma^2_{\mathrm{MB}}(m) = \frac{1}{256} \times \sum_{x=0}^{15} \sum_{y=0}^{15} (r(j,x,y) - r(j-1,x,y))^2 \tag{25}$$

where (x, y) represents the spatial coordination of a pixel in a MB.

D. Proposed Window-Level Rate Control Algorithm

For using the R-D model (18), the window size is first decided by (14) before encoding a window. The model parameter α of (14) is updated after encoding a window by using the actual VQM variation. For each time of calculating L, $W_{\max}$ in (11) is computed and compared to B. In the case of the condition $W_{max} > B$, L is regulated further according to the lower equation of (14). In real-time video communications, window size is dynamically updated along with the change of video content. And the change of window size would provide as smooth as possible visual quality for natural videos under the limited buffer and bit rate resources. It should be pointed out that the real-time concept is from the algorithm design instead of computer implementation here. Generally, both live video broadcasting and video streaming belong to real-time video communication because of the buffer/delay constraint.

After deciding window size, the pre-analysis is performed to provide the MADs $\{MAD_j\}$ and visual quality maps $\{P_f\}$ and $\{p_m\}$ to (18), (20) and (21), respectively. The proposed window-level rate control algorithm is summarized in Table II. The symbols used in the proposed algorithm are listed in Table I.

TABLE I
SYMBOLS USED IN THE FOLLOWING

Symbol	Description
r	Bit rate
f	Frame rate
L	Window size
B,$B(j)$	Buffer size, buffer fullness after encoding the jth frame
T	Target bits for a window
α, β, Θ	Window model parameters
a	R-D model parameter
$R\ (j)$	Encoded bits of the jth frame
$\sigma^2\ (j)$	Variance of the jth residual frame and the mth MB
σ^2 MB (m)	respectively
$\{P_f\}$, $\{p_m\}$	Visual quality map at frame level and MB level
$V_Q(j)$	Objective visual quality of the jth frame
Q_f, Q_m	QP of frame level and MB level for encoding
Q	Predicted QP for preanalysis

TABLE II
PROPOSED WINDOW-LEVEL RATE CONTROL ALGORITHM

Step 1:	If the window is the first one, initializing window size L to the GOP size, which is 16 in our setting; else, computing window size L from the latest α and β according to (14);
Step 2:	Getting the total bits of the current window $T = (r/f) \times L$;
Step 3:	Predicting Q from the last coded window for pre-analysis; for the first window, Q is initialized manually;
Step 4:	Performing pre-analysis on only 16×16 inter prediction for P and B frames and 16×16 intra prediction for I frames in the current window;
Step 5:	Computing MAD for each frame in the current window;
Step 6:	Computing edge strength and motion activity terms according to (2) and (4) for each MB, and computing visual quality map $\{P_f\}$ and $\{p_m\}$ according to (20) and (21);
Step 7:	Computing a QP for the jth frame by (18);
Step 8:	Computing a new QP Q_f by using (21) at frame level as $Q_f = Q_f \times P_f$ for the purpose of smooth visual quality;
Step 9:	Encoding the jth frame using Q_f from Step 8; meanwhile, Q_f is revised by using (20) at MB level as $Q_m = Q_f \times p_m$;
Step 10:	After encoding a frame, computing $V_Q\ (j)$ and $\sigma^2(j)$ and updating buffer $B(j)$ for the jth frame as followings: $V_Q(j)$ is calculated from (5)–(7); $\sigma^2\ (j)$ is the average of $\{\sigma^2_{\mathrm{MB}}(m)\}$ over all MBs: $\sigma^2(j) = \frac{1}{M}\sum_{m=0}^{M-1} \sigma^2_{\mathrm{MB}}(m)$; Buffer status is updated by $B(j) = B(j\text{-}1) - (r/f) + R(j)$, $B(0) = 0$ for the first frame of a window;
Step 11:	If the last frame of current window is reached, go to Step 12; else go to Step 7;
Step 12:	Updating window parameters α and β of (14) as: $\alpha = L \times \delta V_Q^2$ and $\beta = \frac{0.5\times(L+1)\times\Theta}{B(L-1)}$, where $\delta V_Q = \sum_{j=0}^{L-1}(V_Q(j) - \frac{1}{L}\sum_{j=0}^{L-1} V_Q(j))^2$ and $\Theta = max\{\sigma^2(j), j=0, \ldots, L\text{-}1\}$
Step 13:	If the sequence ends, terminates procedure; else go to Step 1. W5148

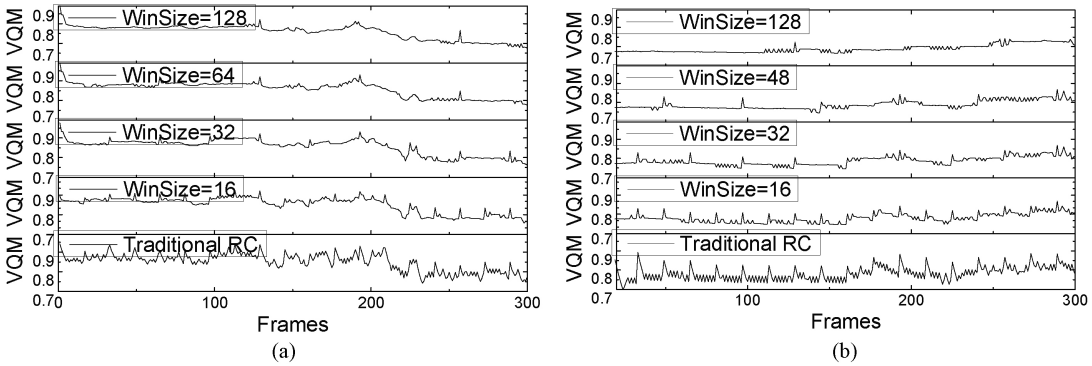

Fig. 2. Frame VQM profiles for the window size of {1, 16, 32, 64, 128}. (a) *Foreman*. (b) *Mobile*.

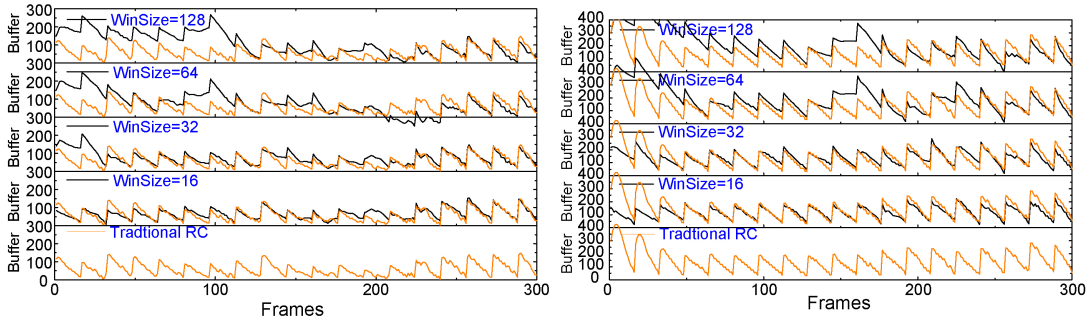

Fig. 3. Instant buffer fullness for the window size of {1, 16, 32, 64, 128}. (a) *Foreman*. (b) *Mobile*.

V. Experimental Results and Discussions

We implemented the proposed algorithm on JM14.0 of H.264/AVC reference software under the conditions: profile/level: 100/40, reference frames: 2, EPZS search, search range: 32 and 64 for CIF (352×288) and 720p (1280×720) resolutions, respectively, RDO on and CABAC, IPPP encoding structure. The following experiments will be arranged into three parts: the first part validates the proposed window model, the second part evaluates the proposed VQM, the third part exhibits both the objective and subjective coding improvement of the proposed window-level rate control algorithm.

A. Validation of Window Model

A large window size L would be computed from (14) if a small visual quality variation ΔV_Q is given for an encoding application. A larger window size needs more buffers for encoding according to (11). Thus, there is actually a tradeoff between visual quality smoothness and buffer constraint. The proposed window-level rate control algorithm is implemented with a constant window size of {1, 16, 32, 64, 128} respectively. It should be pointed out that the proposed window-level rate control would be the traditional rate control if the window size equals 1. The curves of VQMs and instant buffer fullness versus frame no. for *Foreman* and *Mobile* are shown in Figs. 2 and 3. From Fig. 2, it can be observed that larger window sizes can produce smoother frame VQM profile. Correspondingly, larger buffer is needed for larger window size as shown in Fig. 3. In Fig. 2, the VQM curve of WinSize = 16 is smoother than that of WinSize = 1. Meanwhile, the corresponding buffer curve of WinSize = 16 is below that of WinSize = 1.

For a CBR encoding given channel bandwidth and buffer delay, a good rate control tries to seek a tradeoff between visual quality smoothness and buffer constraint. Due to the nonstationary property of the input video signal, the buffer status varies along the video scene change. Thus, the buffer control should adapt to the video content to make full use of the given buffer resource. The experiments with adaptive window size are performed on a CIF sequence *Linker*, consisting of five standard CIF sequences *Foreman*, *Football*, *Tennis*, *News*, and *Silent* in order, 1300 frames in total. A new window size is computed from (14) after encoding a window. The experimental conditions are as follows. L ranges from 16 to 80 and is initialized to 16, VQM variation bound $\Delta V_Q = 0.1$. The experimental results are listed in Table III, where the column labeled "Buffer delay" lists the given buffer constraints. From Table III, we can see that the larger the buffer size is, the more the large windows are applied, and accordingly the less VQM variation is. The actual buffer of encoded bitstream conforms to the given buffer constraint, and the visual quality smoothness accords with the claim of window model. Given 1.0-s buffer delay, most of the windows select the largest window size of 80, and corresponding VQM variation is the least among all cases. While a large number of windows select $L = 16$ given 0.4-s buffer delay.

B. Performance Evaluation of the Proposed VQM

The performance of a VQM can be evaluated by depicting the relationship of the obtained VQM values and the provided subjective ratings, specifically the MOS/DMOS value of each distorted video. The MOS/DMOS value is obtained by

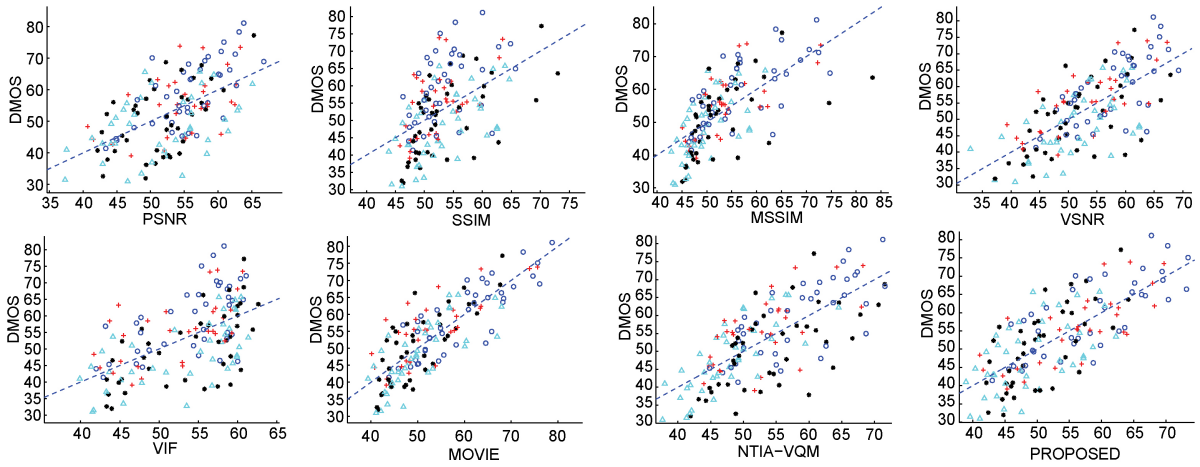

Fig. 4. Scatter plots of the DMOS values versus model predictions on the LIVE video quality database. Each sample point represents one test video. (The circle and "+" indicate the distorted video sequences under wireless network and IP network, respectively. The star indicates H.264 encoded video sequence, while the triangle indicates the MPEG-2 compressed one.) First row from left to right: PSNR, SSIM, MSSIM, and VSNR; second row from left to right: VIF, MOVIE, NTIA-VQM, and the proposed method.

TABLE III

VQM AND BUFFER VARIATIONS WITH THE ADAPTIVE WINDOW SIZE (ΔVQ=0.1, 1000 KB/S) [VQM VARIATION IS MEASURED BY THE STANDARD DEVIATION (STD) OF FRAME VQMS]

Buffer Delay (Seconds)	Percentage (%) of Each Window Size L					STD of VQM	Average VQM	Average Bit Rate
	=16	[16,32]	[32,48]	[48,64]	[64,80]			
0.40	30.712	17.893	14.840	23.952	12.689	1.364	0.9496	1000.68
0.50	2.031	3.539	4.677	10.977	78.797	1.212	0.9486	999.07
0.60	1.322	4.682	0.000	14.185	79.908	1.124	0.9565	998.29
0.80	0.445	1.712	0.000	7.084	90.766	1.023	0.9597	1000.28
1.00	0.000	0.000	1.558	3.332	95.090	0.785	0.9665	1000.34

TABLE IV

LINEAR CORRELATION COEFFICIENT

Algorithm	Wireless	IP	H.264	MPEG-2	All Data
PSNR	0.677	0.478	0.589	0.409	0.569
SSIM	0.473	0.537	0.611	0.582	0.503
MSSIM	0.684	0.684	0.692	0.632	0.676
VSNR	0.680	0.737	0.614	0.507	0.688
VIF	0.593	0.636	0.649	0.673	0.577
NTIA-VQM	0.742	0.655	0.666	0.801	0.716
MOVIE	0.836	0.756	0.790	0.797	**0.810**
Proposed	0.762	0.736	0.709	0.556	**0.741**

TABLE V

SPEARMAN RANK ORDER CORRELATION COEFFICIENT

Algorithm	Wireless	IP	H.264	MPEG-2	All Data
PSNR	0.671	0.430	0.477	0.394	0.553
SSIM	0.539	0.474	0.659	0.569	0.533
MSSIM	0.729	0.645	0.734	0.681	**0.735**
VSNR	0.694	0.693	0.641	0.587	0.672
VIF	0.538	0.553	0.638	0.635	0.558
NTIA-VQM	0.722	0.638	0.648	0.787	0.703
MOVIE	0.810	0.715	0.766	0.773	**0.789**
Proposed	0.753	0.724	0.664	0.564	**0.721**

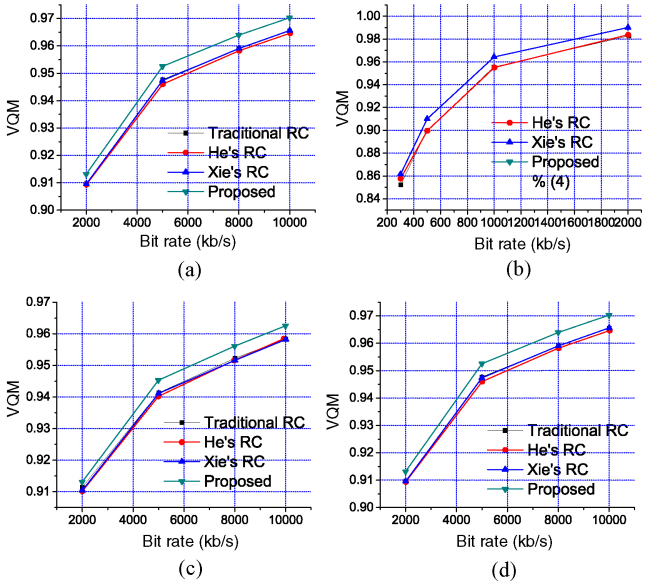

Fig. 5. VQM comparisons between the proposed algorithm and benchmarks. (a), (b) CIF sequences. (c), (d) 720p sequences. (a) *News*. (b) *Silent*. (c) *Crew*. (d) *Harbour*.

subjective viewing tests, where many observers participated and provided their opinions on the visual quality of each distorted video. Therefore, it can be regarded as the ground truth for evaluating the metric performances. As suggested by VQEG HDTV test [1] and that in [48], we follow their evaluation procedure to evaluate the performance of the proposed metric. Let x_j represent the visual quality index of the ith distorted image obtained from the corresponding

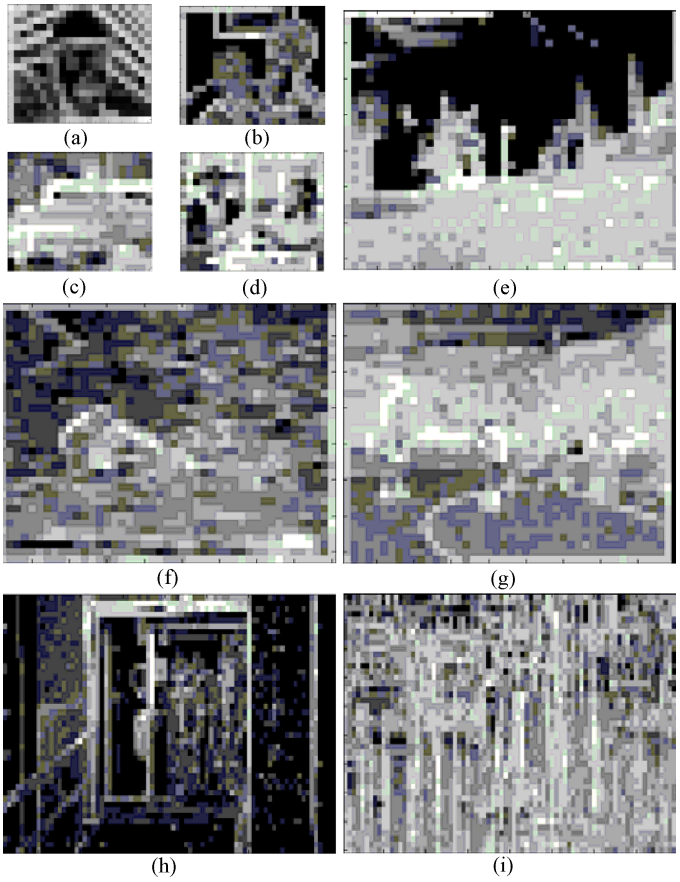

Fig. 6. Visual quality maps used in the proposed algorithm. (a), (b) CIF resolution. (e)–(g) SD (720×576) resolution. (h), (i) 720p resolution. (a) *Foreman*. (b) *Mother-daughter*. (c) *Bus*. (d) *Paris*. (e) *Flower*. (f) *Kayak*. (g) *Basketball*. (h) *Crew*. (i) *Harbour*.

VQA. The five parameter monotonic logistic function is employed to map χ_j and V_j

$$V_j = \beta_1 \times (0.5 - {}^{1}/_{1 + e^{\beta_2 \times (x_j - \beta_3)}}) + \beta_4 \times x_j + \beta_5. \quad (26)$$

The corresponding five parameters $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ are determined by minimizing the sum of squared differences between the mapped objecting score and the subjective DMOS value. Generally, three statistical measurements, LCC, SROCC, and RMSE, are employed to evaluate the performance of a VQM. LCC measures the prediction accuracy. SROCC provides an evaluation of the prediction monotonicity. The RMSE is introduced for evaluating the error during the fitting process. According to the definitions, larger values of LCC and SROCC mean that the objective and subjective scores correlate better, that is to say, a better performance of the VQM. And the smaller RMSE values indicate smaller errors between the two scores, therefore a better performance.

We evaluate the performance of the proposed metric on the LIVE Video Quality Database [49], which contains 150 distorted videos for ten uncompressed high-quality reference videos. There are 15 distorted videos for each reference video with four different distortion types, specifically, MPEG2 compression, H.264/AVC compression, simulated transmission of H.264 compressed bitstreams through error-prone IP networks, and through error-prone wireless networks. Each video in the

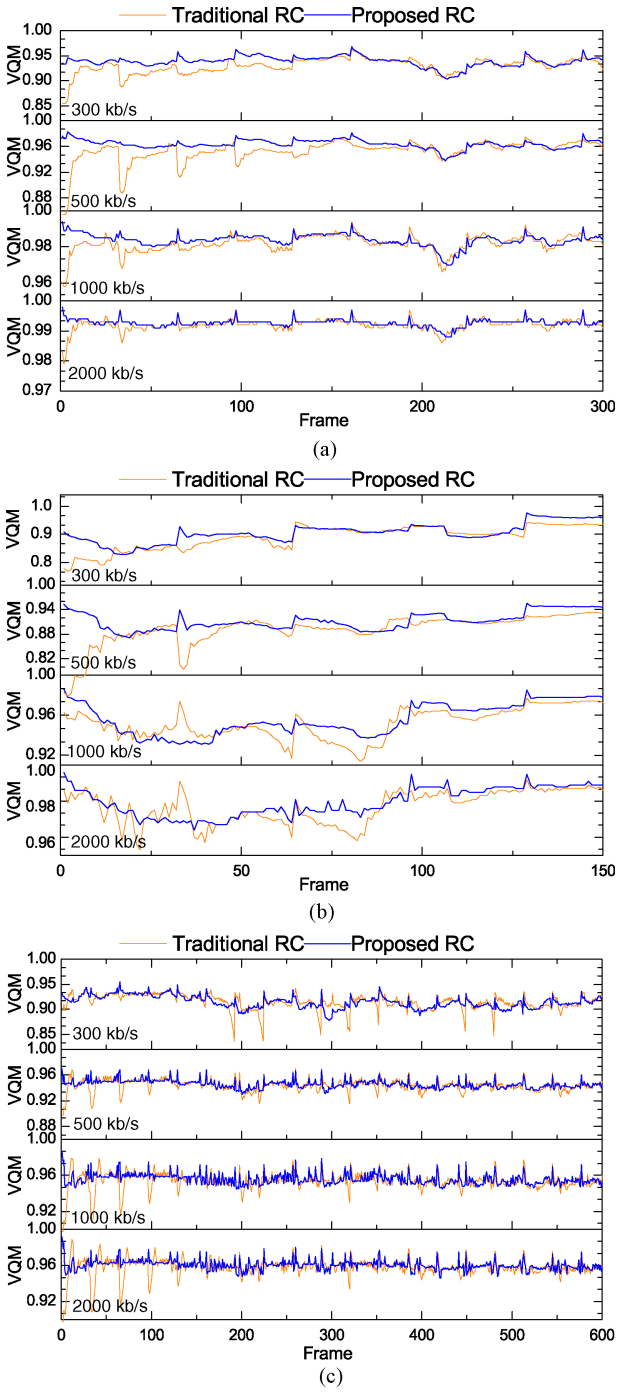

Fig. 7. Visual quality smoothness comparisons between traditional rate control and the proposed rate control. (a) Frame VQM of *Silent* (CIF). (b) Frame VQM of *Tennis* (CIF). (c) Frame VQM of *Crew* (720p).

LIVE Video Quality Database was assessed by 38 human subjects, who scored the video quality on a continuous quality scale. The subjective ratings obtained from the subjective experiments, along with the reference and distorted videos, are provided in the LIVE Video Quality Database.

The performance evaluation criteria, LCC, SROCC, and RMSE, are computed for measuring the correlation between the VQM values and subjective ratings. The computing results show that 0.74, 0.72, and 7.375 are for LCC, SROCC, and RMSE, respectively, which is better than PSRN, VSNR, VIF,

TABLE VI

CODING PERFORMANCE COMPARISONS AMONG ALL TESTING ALGORITHMS IN TERMS OF BIT CONTROL ERROR AND PSNR IMPROVEMENT

Resolution	Sequence	Target bit rate (kb/s)	Traditional Rate Control			He's Rate Control			Xie's Rate Control			Proposed Rate Control		
			Bit rate (kb/s)	PSNR	Error (%)	Bit rate (kbps)	PSNR	Error (%)	Bit rate (kbps)	PSNR	Error (%)	Bit rate (kbps)	PSNR	Error (%)
CIF	*Foreman*	*2000*	2003.59	42.35	0.18	2010.93	42.14	0.55	2000.29	42.22	0.01	1997.88	42.32	−0.11
		1000	1002.31	39.46	0.23	1003.52	39.25	0.35	1001.83	39.24	0.18	1000.74	39.45	0.07
		500	503.38	36.71	0.68	503.17	36.50	0.63	502.58	35.27	0.52	500.65	36.68	0.13
		300	303.76	34.54	1.25	305.86	34.35	1.95	300.96	33.10	0.32	301.38	34.71	0.46
	News	*2000*	2003.17	47.59	0.16	2005.47	46.80	0.27	2003.63	47.55	0.18	1998.70	48.20	−0.06
		1000	1002.53	44.96	0.25	997.46	44.30	−0.25	1001.11	44.75	0.11	1000.57	45.03	0.06
		500	502.37	41.90	0.47	502.04	41.60	0.41	501.18	41.65	0.24	500.26	42.01	0.05
		300	301.84	39.29	0.61	302.27	39.19	0.76	300.60	39.15	0.20	301.13	39.54	0.38
	Silent	*2000*	2004.06	45.33	0.20	2007.19	44.44	0.36	2001.77	45.26	0.09	2000.80	46.04	0.04
		1000	1004.05	41.89	0.40	997.71	41.33	−0.23	1001.38	41.77	0.14	999.88	42.30	−0.01
		500	502.13	38.59	0.43	498.29	38.20	−0.34	500.62	38.34	0.12	499.87	38.52	−0.03
		300	301.91	36.22	0.64	301.58	36.14	0.53	301.41	36.08	0.47	300.83	36.07	0.28
	Tennis	*2000*	2001.86	41.85	0.09	1960.98	41.35	-1.95	2003.99	41.61	0.20	2000.46	42.30	0.02
		1000	1002.89	38.45	0.29	996.50	38.33	−0.35	1002.74	38.15	0.27	1001.82	38.79	0.18
		500	502.50	35.19	0.50	501.06	35.10	0.21	500.10	34.88	0.02	501.59	35.48	0.32
		300	302.07	32.82	0.69	298.34	32.86	−0.55	308.56	32.44	2.85	300.02	33.36	0.01
Average				**39.82**	**+0.44**		**39.49**	**+0.15**		**39.47**	**+0.37**		**40.05**	**+0.11**
					0.44			**0.61**			**0.37**			**0.14**
720p	*Night*	*10 000*	10006.99	39.28	0.07	10025.51	39.19	0.26	10008.39	39.19	0.08	9992.78	39.39	−0.07
		8000	8006.98	38.55	0.09	7998.37	38.42	−0.02	7998.45	38.44	−0.02	7995.24	38.64	−0.06
		5000	5007.72	37.04	0.15	5011.75	36.97	0.23	5002.33	36.98	0.05	4996.27	37.33	−0.07
		2000	2005.02	33.52	0.25	2005.72	33.41	0.29	2002.43	33.49	0.12	2001.47	34.07	0.07
	Crew	*10 000*	10010.21	41.61	0.10	9927.34	41.53	−0.73	10007.05	41.52	0.07	9997.48	41.51	−0.03
		8000	8010.74	40.99	0.13	7970.62	41.07	−0.37	8009.60	40.96	0.12	7995.51	40.85	−0.06
		5000	5008.12	39.85	0.16	4995.38	39.77	−0.09	5006.36	39.72	0.13	4997.43	39.82	−0.05
		2000	2005.63	37.03	0.28	1999.71	36.88	−0.01	2006.25	36.75	0.31	1995.30	37.41	−0.24
	Harbour	*10 000*	10006.50	37.49	0.07	10020.98	37.52	0.21	10004.20	37.54	0.04	9997.64	37.52	−0.02
		8000	8005.90	36.52	0.07	8014.40	36.63	0.18	8005.49	36.65	0.07	7998.81	36.64	−0.01
		5000	5005.88	34.75	0.12	5007.71	34.77	0.15	4997.14	34.70	−0.06	5000.22	34.83	0.00
		2000	2004.24	31.27	0.21	2002.53	31.18	0.13	2001.35	31.19	0.07	1996.73	31.41	−0.16
Average				**37.32**	**+0.14**		**37.28**	**+0.02**		**37.26**	**+0.08**		**37.45**	**−0.06**
					0.14			**0.22**			**0.09**			**0.07**

MSSIM, and SSIM metrics. This indicates that the proposed VQM can better assess the visual quality than those metrics that do not take into account the temporal information. We also compare the proposed VQM with PSNR, SSIM [16], [17], MSSIM [18], VSNR [51], VIF [52], NTIA-VQM [25], and MOVIE [50]. As PSNR, SSIM, MSSIM, VSNR, and VIF only provide frame-level quality scores, the final quality index of the video sequence is generated by averaging their outputs for all frames. The experimental results of LCC and SROCC measurements are illustrated in Tables IV and V. The RMSEs of PSNR, SSIM, MSSIM, VSNR, VIF, and NTIA-VQM are 9.188, 8.267, 7.717, 9.777, 7.860, and 7.664, respectively, which are all larger than ours. Regarding computational complexity, for a 250-frame 768×432 sequence on a 3G-Hz quad-core CPU with 6G RAM, the computing times of PSNR, SSIM, MSSIM, VSNR, VIF, NTIA-VQM, MOVIE, and the proposed metric are 4, 24, 60, 26, 636, 57, 6320, and 43 s, respectively. The proposed metric is superior to MSSIM, VIF, NTIA-VQM, and MOVIE with respect to computing time. In addition, the less computing time would be for the proposed metric in video coding since the MSE of each MB is available.

From Table IV, it can be observed that PSNR performs poorly, because it is not related to the HVS perception. Also, the VSNR performs badly, which can be attributed to two reasons. The first reason is that VSNR analyzes the HVS perception of the distortion in the wavelet domain. But the MPEG-2 and H.264 compression schemes introduce the distortions during the quantization process in DCT domain. The second one is that VSNR is an image quality metric designed to capture the spatial distortions. For video quality assessment, the temporal information is very important and needs to be accounted for. This is also the reason why SSIM, MSSIM, and VIF perform successfully in image quality evaluation, but not so well on the video quality assessment. From Table IV, it can be observed that the performances of these metrics are not good enough, with SROCC values smaller than 0.6. The reason is that the temporal information is not included. Our proposed method outperforms PSNR, VSNR, SSIM, MSSIM, and VIF. This means that the proposed metric can effectively depict the perceptual quality of the distorted videos. The scatter plots of different VQMs over the LIVE Video Quality Database are illustrated in Fig. 4. It can be observed that for our proposed method the sample points scatter more closely around the

TABLE VII
CODING PERFORMANCE COMPARISONS AMONG ALL TESTING ALGORITHMS IN TERMS OF AVERAGE VQM (THE REAL BIT RATE OF EACH ITEM IS THE SAME AS THAT LISTED IN TABLE VI)

Sequence	Target Bit Rate	Traditional Rate Control	He's Rate Control	Xie's Rate Control	Proposed Control
Foreman	*2000*	0.9817	0.9817	0.9817	0.9855
	1000	0.9643	0.9651	0.9656	0.9696
	500	0.9386	0.9387	0.9392	0.9457
	300	0.9125	0.9140	0.9145	0.9218
News	*2000*	0.9939	0.9941	0.9949	0.9972
	1000	0.9856	0.9872	0.9869	0.9896
	500	0.9734	0.9756	0.9756	0.9793
	300	0.9645	0.9668	0.9671	0.9693
Silent	*2000*	0.9772	0.9827	0.9837	0.9901
	1000	0.9504	0.9555	0.9551	0.9643
	500	0.8889	0.8996	0.8996	0.9101
	300	0.8462	0.8522	0.8577	0.8615
Average		**0.9481**	**0.9511**	**0.9518**	**0.9570**
Night	*10 000*	0.9745	0.9743	0.9758	0.9807
	8000	0.9673	0.9667	0.9684	0.9736
	5000	0.9525	0.9520	0.9527	0.9587
	2000	0.9054	0.9042	0.9039	0.9094
Crew	*10 000*	0.9583	0.9584	0.9582	0.9625
	8000	0.9523	0.9516	0.9516	0.9561
	5000	0.9413	0.9402	0.9411	0.9453

fitted line. It means that the values predicted by the proposed method correlate better with the subjective ratings, specifically the DMOS values, demonstrating a better performance.

C. Objective and Subjective Coding Performance of the Proposed Algorithm

The competition occuring between the proposed window-level algorithm and three benchmark algorithms are the traditional one (JVT-H017r3) [40], Xie and Zeng's [29] algorithm, and He's [28] algorithm. The experimental results show that the proposed algorithm is better than the benchmarks in terms of both the objective and subjective measurements.

The objective coding performance is measured by both PSNR and the proposed VQM. The comparisons of PSNR, VQM, and bit control accuracy are listed in Tables VI and VII, where the proposed algorithm achieves a significant PSNR and VQM improvement than the benchmarks. The PSNR improvement of the proposed algorithm is up to 0.3 dB over the benchmarks on CIF sequences. In addition, the proposed algorithm is much better than the benchmarks in terms of bit rate control accuracy. The maximum bit rate mismatches of Xie's and He's algorithms are over 2% that is worse than that of the proposed algorithm with maximum bit rate mismatch of 0.32%, as shown in Table VI. The average bit control error and the average PSNR are listed in the "Average" row of Table VI, where the bottom number of bit control error is the average of absolute bit control errors. Table VII lists the VQM comparisons of the competitive algorithms, which are also illustrated in Fig. 5. From Fig. 5, the significant VQM improvement of the proposed algorithm over the benchmarks can be observed. The VQM of the proposed algorithm is about 1% more than that of the traditional one for all the

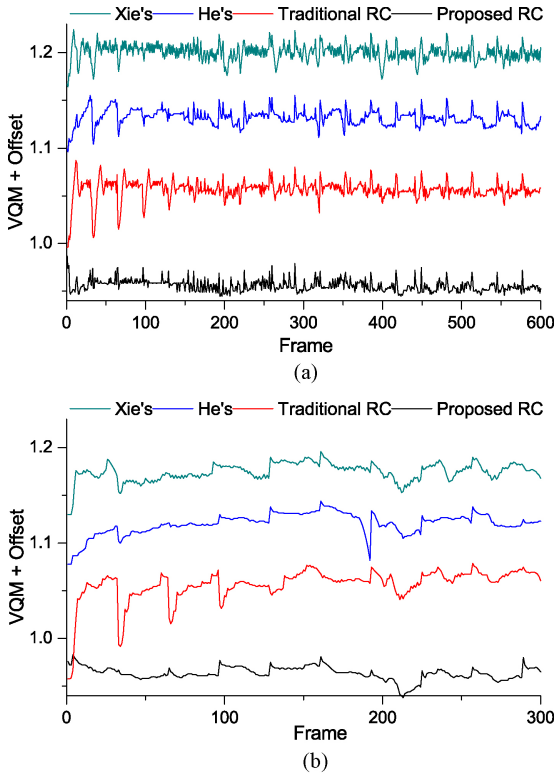

Fig. 8. Visual quality comparisons between the proposed algorithm and all benchmarks. (a) At 5000 kb/s, *Crew* (720p). (b) At 500 kb/s, *Silent* (CIF).

test sequences. Such a percentage is actually more significant than the percentage of 0.3-dB PSNR improvement over the 40-dB base. In the experiments, the proposed algorithm employs the visual quality map produced by (19) to improve the visual quality of encoding, while all the benchmarks employ a uniform visual quality map without distinguishing the contribution of each block area to the human eye's perception. As an example, the VQM maps of the first P frames of some test sequences are shown in Fig. 6. The VQM value is computed from (20) for each MB. In Fig. 6, the brightness of each pixel represents the VQM value of the corresponding MB.

Referring to [28] and [29], the significant smooth visual quality represented by PSNR or distortion profiles can be obtained by Xie's and He's algorithms than the traditional one. The frame VQM of the proposed algorithm is compared with those of the benchmarks in Fig. 7. It can be observed that the proposed algorithm achieves a much smoother frame VQM profile than the traditional algorithm. Comparing the proposed algorithm with [28] and [29], the proposed algorithm is much better than the other two from the aspect of picture quality smoothness, which can be observed from the frame VQM curves drawn in Fig. 8. In Fig. 9, some pictures are shown for comparing the subjective visual quality of proposed and traditional algorithms. In [28], the large filter length is good at smoothing visual quality, but may cause the significant bit control error. In [29], a more elaborate mechanism was provided to handle the situations when the conventional bit allocation and/or QP determination of rate control failed, which makes the algorithm more complex. Comparing the computational complexities of these algorithms, a slight overhead is needed

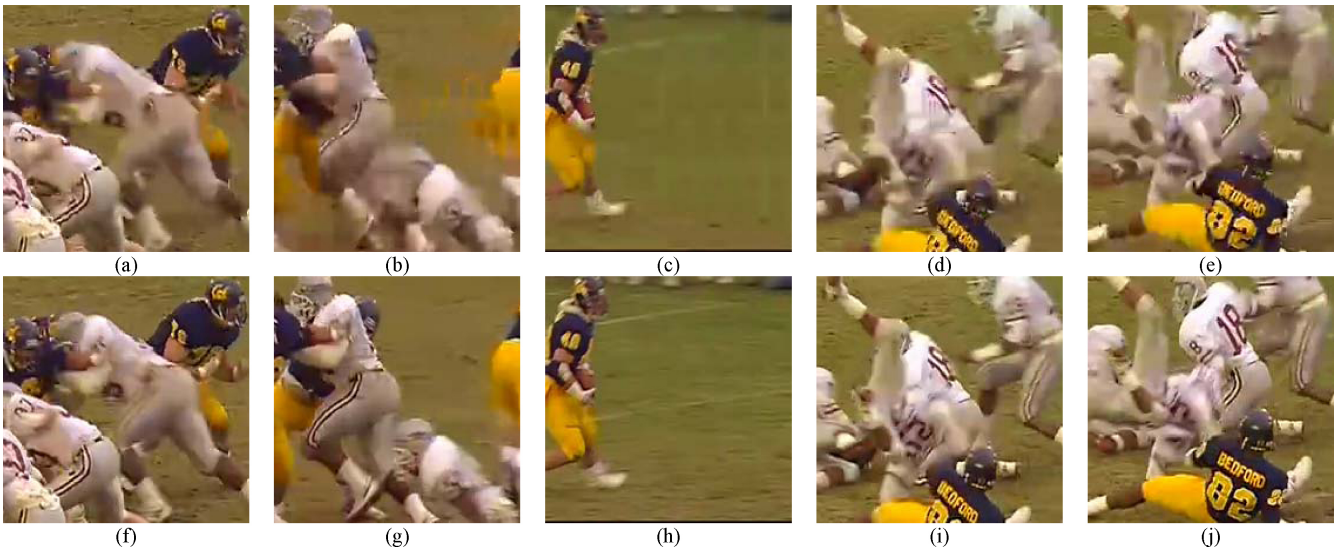

Fig. 9. Subjective visual quality comparison is between the traditional rate control and the proposed one on *Football* at 300 kb/s. (a)–(e) Decoded from the bitstreams of the traditional rate control algorithm. (f)–(j) Decoded from the bitstream of the constant QP encoding. (a), (f) 13th frame. (b), (g) 31th frame. (c), (h) 109th frame. (d), (i) 218th frame. (e), (j) 220th frame.

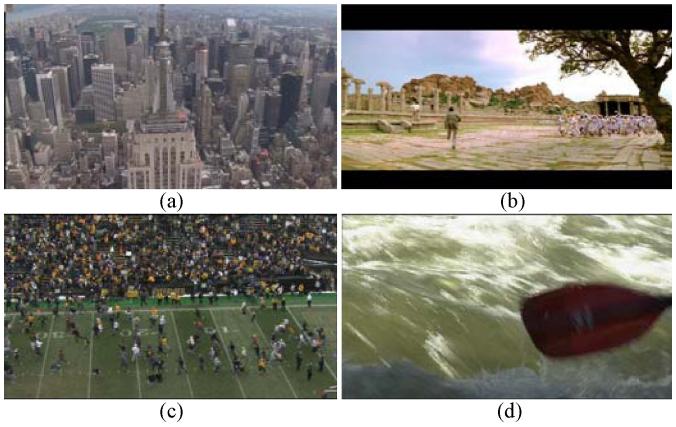

Fig. 10. One frame from each of the four source videos in the study (1080p represents 1920×1080 resolution, progressively scanned video). (a) *City* (720p). (b) *The Myth* (720p). (c) *RushField* (1080p). (d) *RedKayak* (1080p).

for the proposed algorithm because of a pre-analysis process in the proposed algorithm. However, the achievement of the proposed algorithm in visual quality control is significant. From the statistics of encoding time, the time of pre-analysis is about 4%–7% of the total encoding time as fast ME is enabled, and less than 4% as full ME is enabled. The time complexity of the proposed algorithm is only 1.8%–2.1% higher than that of the traditional one. In addition, the computational complexity could be further reduced by reusing the motion information of pre-analysis. In practical, the pre-analysis can be implemented separately from the other parts of a standard encoder, which is efficient in the hardware accelerator of an encoder.

We conducted the subjective evaluation in the IVP Laboratory, Chinese University of Hong Kong, Shatin, Hong Kong [54]. The evaluation was performed in a studio room with lighting condition satisfying the lab environment requirement of the ITU-R BT.500 standard [55]. The display monitor is a 65-inch Panasonic plasma display (TH-65PF9WK) and the viewing distance is three times (for 1920×1080 videos)/4.5 times (for 1280×720 videos) of the picture height. We used a high-performance workstation stored and displayed all videos in a format of raw TIF sequences. Eighteen observers 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 four source videos and 32 distorted videos. A single-stimulate method ACR [56] was used where each video (including the reference) occurred once in a random order, yet the two successive videos come from different source videos so as to remove contextual and memory effects in quality evaluation. Between the presentations of two videos, a mid-gray video in 5 s was displaying, and meanwhile, the evaluation was reported on the five-point scale: 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 [55] to derive the difference mean opinion score (DMOS) and the 95% confidence interval for each compressed video. The $\beta 2$ test suggested in [55] is used to identify the subjects whose quality judgments deviate from the distribution of the normal scores significantly. One out of the 18 subjects is rejected, which means that most of the subjective viewers achieved an agreement on the visual qualities of the encoded video sequences.

We used four uncompressed, high quality source videos of natural scenes, including a movie video "The Myth." We created four distorted videos from each source video, using four different quantizations for the traditional rate control algorithm and the proposed algorithm, respectively. The source videos are a high definition YUV 4:2:0 format. Fig. 10 shows one frame of each source video. All videos are 10 s long with a frame rate of 25 frames. The subjective experiments are performed to prove the better subjective visual quality

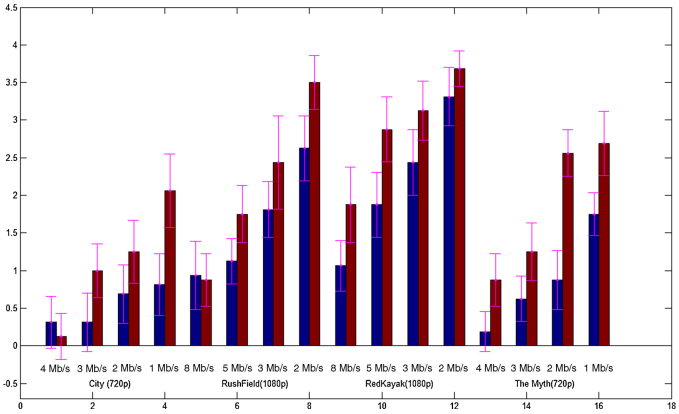

Fig. 11. Subjective quality comparison between the proposed algorithm and the traditional rate control algorithm.

of the proposed algorithm over the traditional one. The final result is shown in Fig. 11. The horizontal axis indicates the encoded video sequences in different bit-rates. The blue histograms are the DMOS values of video sequences generated by the proposed algorithm, while the red ones are the DMOS values from the traditional algorithm in [40]. The magenta error bar indicates the 95% confidence interval for each encoded sequence. It means that there is a 95% chance that the estimated mean value falls in the interval. The smaller DMOS value indicates the better visual quality of the encoded video sequence. It can be observed that the proposed method outperforms the other one [40] on most of the test sequences from Fig. 11, except for the 4 Mb/s of *City* and 8 Mb/s of *RushField*. For the 4 Mbps of *City*, both methods generate high quality videos. The DMOS values are around 0.3 that means that the quality is nearly as good as the reference. In this case, the subject viewers will have difficulties to judge the quality of each video sequence. For the 8 Mb/s of *RushField*, the difference of DMOS values is very small. For the other cases, the perceptual qualities of the proposed method are better than those of [40], which have clearly demonstrated the superiority of our proposed algorithm for consistent visual quality control.

VI. Conclusions

In this paper, we first proposed a practical VQM for video coding, taking into account both spatial and temporal activities to simulate the visual masking effect. Second, a theoretical window model to depict the relations among window size, quality variation, and bits variation was constructed based on the proposed VQM. Then, a VQM-based R-D model and a window-level rate control algorithm were developed and verified for video coding. In the proposed rate control algorithm, the tradeoff between visual quality consistency and buffer constraint was investigated. In addition, there was no bit allocation at the frame level in the proposed algorithm. Conversely, in the traditional algorithm, the visual quality fluctuation may be caused by the bit allocation that assumed stationary input video signal. Experimental results of video coding proved that the proposed algorithms can provide smooth visual quality of compressed videos with compliant buffer usages.

References

[1] Video Quality Expert Group. (2000, Mar.). *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment* [Online]. Available: http://www.vqeg.org

[2] P. Corriveau and A. Webster. (2003, Jul.). *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II* [Online]. Available: http://www.vqeg.org

[3] S. Winkler and R. Campos, "Video quality evaluation for Internet streaming applications," in *Proc. SPIE/IS&T Human Vision Electron. Imaging*, vol. 5007. Jan. 2003, pp. 104–115.

[4] G.-M. Muntean, P. Perry, and L. Murphy, "Subjective assessment of the quality-oriented adaptive scheme," *IEEE Trans. Broadcast*., vol. 51, no. 3, pp. 276–286, Sep. 2005.

[5] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. *Live Image Quality Assessment Database Release 2* [Online]. Available: http://live.ece.utexas.edu/research/quality, 2005.

[6] N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola, and F. Battisti, "Color image database for evaluation of image quality metrics," in *Proc. 10th IEEE Workshop Multimedia Signal Process.*, Oct. 2005, pp. 403–408.

[7] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective quality assessment of H.264/AVC video streaming with packet losses," *EURASIP J. Image Video Process.*, vol. 2011, no. 190431, pp. 1–12, Jan. 2011.

[8] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it—A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[9] 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.

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

[11] S. Winkler, "A perceptual distortion metric for digital color video," in *Proc. SPIE Conf. Human Vision Electron. Imaging*, vol. 3644. Jan. 1999, pp. 175–184.

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

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

[14] 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.

[15] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Sel. Topics Signal Process*., vol. 3, no. 2, pp. 253–265, Apr. 2009.

[16] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Proc.: Image Commun.*, vol. 19, no. 2, pp. 121–132, Feb. 2004.

[17] 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.

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

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

[20] K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1. Apr. 2007, pp. I869–I872.

[21] K. Seshadrinathan and A. C. Bovik, "An information theoretic video quality metric based on motion models," in *Proc. 3th Int. Workshop Video Proc. Quality Metrics Consumer Electron.*, Jan. 2007, pp. 25–26.

[22] A. K. Moorthy, K. Seshadrinathan, and R. Soundararajan, "Wireless video quality assessment: A study of subjective scores and objective algorithms," *IEEE Trans. Circuit Syst. Video Technol*., vol. 20, no. 4, pp. 587–597, Apr. 2010.

[23] 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.

[24] R. T. Born and D. C. Bradley, "Structure and function of visual area MT," *Annu. Rev. Neurosci*., vol. 28, no. 1, pp. 157–189, 2005.

[25] 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.
[26] 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.
[27] M. Yuen and H. R. Wu, "A survey of hybrid MC/DPC/DCT video coding distortions," *Elsevier, Signal Process.*, vol. 70, pp. 247–278, Nov. 1998.
[28] Z. H. He, W. Zeng, and C. W. Chen, "Low-pass filtering of rate-distortion functions for quality smoothing in real-time video communication," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 8, pp. 973–981, Aug. 2005.
[29] B. Xie and W. Zeng, "A sequence-based rate control framework for consistent quality real-time video," *IEEE Trans. Circuits Syst. Video Technol*., vol. 16, no. 1, pp. 56–71, Jan. 2006.
[30] T.-S. Ou, Y.-H. Huang, and H. H. Chen, "SIM-Based perceptual rate control for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 5, pp. 682–691, May 2011.
[31] 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.
[32] 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.
[33] C. Lee and O. Kwon, "Objective measurements of video quality using the wavelet transform," *SPIE Opt. Eng.*, vol. 42, no. 1, pp. 265–272, 2003.
[34] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 266–279, Apr. 2009.
[35] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 253–265, Apr. 2009.
[36] L. Su, Y. Lu, F. Wu, S. P. Li, and W. Gao, "Complexity-constrained H.264 video encoding," *IEEE Trans. Circuits Syst. Video Technol*., vol. 19, no. 4, pp. 1–15, Apr. 2009.
[37] P. H. Westerink, R. Rajagopalan, and C. A. Gonzales, "Two-pass MPEG-2 variable-bit-rate encoding," *IBM J. Res. Develop.*, vol. 43, no. 4, pp. 471–488, Jul. 1999.
[38] K. Wang and J. W. Woods, "MPEG motion picture coding with long-term constraint on distortion variation", *IEEE Trans. Circuits Syst. Video Technol*., vol. 18, no. 3, pp. 1–11, Mar. 2008.
[39] Z. G. Li, W. Gao, F. Pan, S. W. Ma, K. P. Lim, G. N. Feng, X. Lin, S. Rahardja, H. Q. Lu, and Y. Lu, "Adaptive rate control for H.264," *J. Visual Commun. Image Representation*, vol. 17, no. 2, pp. 376–406, Apr. 2006.
[40] S. W. Ma, Z. G. Li, and F. Wu, *Proposed Draft Adaptive Rate Control*, document JVT-H017r3, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, 8th Meeting, Geneva, May 2003.
[41] L. Xu, D. B. Zhao, X. Y. Ji, L. Deng, S. Kwong, and W. Gao, "Window- level rate control for smooth picture quality and smooth buffer occupancy," *IEEE Trans. Image Process*., vol. 20, no. 3, pp. 723–734, Mar. 2011.
[42] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*. Cambridge, MA, USA: Cambridge Univ. Press.
[43] Z. S. Wei, *Probability and Statistics*. Beijing, China: Higher Education Press, 1983.
[44] J. R. Corbera, P. A. Chou, and S. L. Regunathan, "A generalized hypothetical reference decoder for H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol*., vol. 13, no. 7, pp. 674–687, Jul. 2003.
[45] *Time-Shift Causality Constraint on the CAT-LB HRD*, document JVT-E133.doc, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, 5th Meeting, Geneva, Switzerland, Oct. 2002.
[46] L. Xu, X. Y. Ji, W. Gao, and D. B. Zhao, "Laplacian distortion model (LDM) for rate control in video coding," in *Proc. IEEE PCM*, Dec. 2007, pp. 638–646.
[47] Z. H. He, Y. K. Kim, and S. K. Mitra, "Low-delay rate control for DCT video coding via ρ-domain source modeling," *IEEE Trans. Circuits Syst. Video Technol*., vol. 11, no. 8, pp. 928–940, Aug. 2001.
[48] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process*., vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[49] *LIVE Video Quality Database*. (2010) [Online]. Available: http://live.ece.utexas.edu/research/quality/live_video.html
[50] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process*., vol. 19, no. 6, pp. 1427–1441, Jun. 2010.
[51] 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.
[52] 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.
[53] N. Kamaci, Y. Altunbasak, and R. M. Mersereau, "Frame bit allocation for the H.264/AVC video coder via Cauchy-density-based rate and distortion models" *IEEE Trans. Circuits Syst. Video Technol*., vol. 15, no. 8, pp. 994–1006, Aug. 2005.
[54] F. Zhang *et al*. (2011). *IVP Video Quality Database* [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/index.shtml
[55] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R Rec. BT.500-11, ITU, Geneva, Switzerland, 2002.
[56] *Subjective Video Quality Assessment Methods for Multimedia Applications*, ITU-T Rec. P.910, ITU, Geneva, Switzerland, 2008.

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

He was a Senior Research Associate with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, from August 2009 to July 2011. He was a Post-Doctoral Fellow with the Department of Electronic Engineering, Chinese University of Hong Kong, Shatin, Hong Kong, from July 2011 to July 2012. He is now with the Department of Automatic Engineering, University of Science and Technology Beijing, Beijing, China. His current research interests include image/video coding, wavelet-based image/video coding, and computer vision.

Songnan Li (S'08) received the B.S. and M.S. degrees from the Harbin Institute of Technology, Harbin, China, in 2004 and 2006, respectively, and the Ph.D. degree from the Chinese University of Hong Kong, Shatin, Hong Kong, in 2012.

His professional skills and current research interests include free-viewpoint videos, visual quality assessment, anaglyph image generation, video deinterlacing, video compression, and code optimization.

King Ngi Ngan (F'00) received the Ph.D. degree in electrical engineering from Loughborough University, U.K.

He is currently a Chair Professor at the Department of Electronic Engineering, Chinese University of Hong Kong, Shatin, Hong Kong. He was previously a Full Professor with Nanyang Technological University, Singapore, and the University of Western Australia, Australia. He holds honorary and visiting professorships of numerous universities in China, Australia and South East Asia. He has published extensively, including three authored books, six edited volumes, over 300 refereed technical papers, and edited nine special issues in journals. He also holds ten patents in the areas of image/video coding and communications.

Prof. Ngan has served as 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 has chaired a number of prestigious international conferences on video signal processing and communications, and served on advisory and technical committees of numerous professional organizations. He co-chaired the IEEE International Conference on Image Processing, Hong Kong, in September 2010. He is a fellow of IET (U.K.), and IEAust (Australia) and was an IEEE Distinguished Lecturer for 2006 to 2007.

Lin Ma (M'13) received the B.E. and M.E. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively. He is currently pursuing the Ph.D. degree at the Department of Electronic Engineering, Chinese University of Hong Kong (CUHK), Shatin, Hong Kong.

He was a Research Intern with Microsoft Research Asia, Beijing, China, from October 2007 to March 2008. He was a Research Assistant with the Department of Electronic Engineering, CUHK, from November 2008 to July 2009. He was a Visiting Student with the School of Computer Engineering, Nanyang Technological University, Singapore, from July to September 2011. His current research interests include image/video quality assessment, super-resolution, restoration, and compression.

Mr. Ma was a recipient of the Best Paper Award in the Pacific-Rim Conference on Multimedia in 2008. He was awarded the Microsoft Research Asia Fellowship in 2011. He was a finalist for the HKIS Young Scientist Award in engineering science in 2012.