SSIM-Motivated Two-Pass VBR Coding for HEVC

Shiqi Wang, *Member, IEEE,* Abdul Rehman, Kai Zeng, Jiheng Wang, and Zhou Wang, *Fellow, IEEE*

***Abstract*— We propose a structural similarity (SSIM)-motivated two-pass variable bit rate control algorithm for High Efficiency Video Coding. Given a bit rate budget, the available bits are optimally allocated at group of pictures (GoP), frame, and coding unit (CU) levels by hierarchically constructing a perceptually uniform space with an SSIM-inspired divisive normalization mechanism. The Lagrange multiplier λ, which controls the tradeoff between perceptual distortion and bit rate, is adopted as the GoP level complexity measure. To derive λ, Laplacian distribution-based rate and perceptual distortion models are established after the first pass encoding, and the target bits are dynamically allocated by maintaining a uniform Lagrange multiplier level for each GoP through λ equalization. Within each GoP, rate control is further performed at frame and CU levels based on SSIM-inspired divisive normalization, aiming to transform the prediction residuals into a perceptually uniform space. Experiments show that the proposed scheme achieves high accuracy rate control and superior rate-SSIM performance, which is further verified by subjective visual testing.**

***Index Terms*— Divisive normalization, High Efficiency Video Coding (HEVC), structural similarity (SSIM) index, two-pass rate control, variable bit rate (VBR) coding.**

I. Introduction

THE exponentially increasing demand for high-definition (HD) and beyond-HD videos has been creating an ever-stronger demand for high-performance video compression technologies. The High Efficiency Video Coding (HEVC) standard [1], jointly developed by ITU-T Video Coding Experts Group and ISO/IEC Moving Picture Experts Group (MPEG), was claimed to achieve potentially more than 50% coding gain compared with H.264/AVC [2], thanks to many novel techniques being adopted. At the block level, an adaptive quadtree structure based on the coding tree unit (CTU) is employed, and three new concepts, namely, coding unit (CU), prediction unit (PU), and transform unit (TU), were developed to specify the basic processing unit of coding, prediction, and transform [3]. In contrast to the 16×16 macroblock (MB) in H.264/AVC, the CTU size can be $L \times L$, where L can be chosen from $16, 32, 64$, and a larger size usually enables higher compression performance, especially for HD and beyond-HD video sequences. At the frame level, the flexible reference management scheme based on the concept of reference picture set is adopted [4], which efficiently manages reference pictures under the constraint of limited decoded picture buffer. To further improve the coding efficiency, the quality of each picture, according to the reference structure, is optimized by adjusting the quantization parameter (QP) and Lagrangian multiplier. These unique features bring new challenges to the design of optimal HEVC encoders.

In practice, many digital video applications are constrained by limited storage space or bandwidth. Therefore, rate control schemes have been widely employed in the encoder implementation. When delivering a compressed bitstream under a bandwidth constraint, the goal of rate control is to avoid overflow and underflow, and meanwhile optimize the overall coding efficiency. To perform efficient and accurate rate control, appropriate rate and distortion models should be established [5], [6]. Previous rate control algorithms were proposed with specific considerations of the corresponding video coding standards (for example, TM5 for MPEG-2 [7], TMN8 for H.263 [8], and VM8 for MPEG-4 [9]). In view of this, several rate control algorithms were proposed for HEVC, targeting at achieving constant bit rate (CBR) coding. The first rate control algorithm for HEVC was described in [10], which was previously adopted into HM software. In [11], considering the new reference frame selection mechanism, rate-group of pictures (GoP)-based distortion and rate models were established and ρ domain rate control was proposed, where ρ denotes the percentage of zero coefficients in a frame after quantization [12]. In [13], an adaptive rate control scheme was presented by modeling the rate-quantization relationship with frame complexity, and Laplacian distribution-based CTU level bit allocation was further developed to improve the coding performance. In [14] and [15], Lagrange parameter (λ) domain rate control was proposed and adopted in HM, where the QP value for each frame is obtained from the corresponding λ value.

Although these CBR rate control algorithms have significantly improved the control accuracy and achieved desirable coding performance, little investigation on perceptual relevant rate control of variable bit rate (VBR) coding of HEVC has been done. In the literature, there have been existing studies on VBR for H.264/AVC coding [16]–[19]. Kamran *et al.* [20] proposed a novel frame-level fuzzy VBR rate control scheme for HEVC, which satisfies the buffer constraint and reduces the fluctuations of QP and peak signal-to-noise ratio (PSNR) simultaneously. In contrast to CBR coding, the advantage of VBR is that it allows for a varying amount of output data per time segment. Regarding the video compression scenario, the video content is usually nonstationary, such that the compression performance can be optimized if the output size

Manuscript received October 19, 2015; revised January 25, 2016 and April 27, 2016; accepted June 8, 2016. Date of publication June 13, 2016; date of current version October 3, 2017. This paper was recommended by Associate Editor M. Hannuksela.

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

A. Rehman, K. Zeng, J. Wang, and Z. Wang are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sqwang1986@gmail.com).

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

Digital Object Identifier 10.1109/TCSVT.2016.2580398

of the video file is allowed to change over time. For example, we can distribute fewer bits to the easier-to-code content and reserve bandwidth and buffer capacities for the content that is more challenging and requires more bitrates. Regarding the two-pass VBR coding, the basic principle is that the first pass is performed with constant QPs or CBR to collect the information and infer the scene complexity. This information is subsequently employed to guide bit allocation and adjust the coding parameters in the second pass compression so that higher coding performance and/or more consistent video quality can be achieved. In this scenario, the fundamental issue of bit allocation is to obtain the best quality under bit rate constraint by optimally distributing the coding bits. Central to such problems is rate–distortion optimization (RDO), which attempts to optimize the perceptual quality of the whole sequence D subject to the constraint R_c

$$\min\{D\} \text{ subject to } R \leq R_c. \tag{1}$$

Such an RDO process can be converted into an unconstrained optimization problem [21] by

$$\min\{J\} \text{ where } J = D + \lambda \cdot R \tag{2}$$

where J is called the RD cost and λ is known as the Lagrange multiplier that controls the tradeoff between R and D.

The way in which the distortion D is defined can have a great impact on the perceptual quality of the encoded video. Recently, a lot of work has been done to develop objective quality assessment measures, which provide more reliable predictions of perceptual image quality than mean squared error (MSE) and PSNR [22]–[24]. In this paper, we employ a structural similarity (SSIM) index-based quality measure [25], [26]. SSIM has been widely applied in various image/video processing areas due to its excellent compromise between quality evaluation accuracy and computational efficiency. It is also proved to be more effective in quantifying the suprathreshold compression artifacts, such as artifacts that distort the structure of an image [27]. As a result, it has been incorporated into key coding modules to improve the compression efficiency, including motion estimation, mode selection, and rate control [28]–[38]. SSIM-based RDO schemes were presented in [28]–[30] to improve the coding efficiency of intra frames. Along this vein, perceptual RDO schemes for inter-frame prediction and mode selection based on SSIM were further developed in [31]–[33]. To adapt the input video properties, SSIM-based mode selection and MB level rate control methods were proposed in [34]–[38], which employed a rate-SSIM curve to describe the relationship between SSIM and rate. The Laplacian distribution-based rate and distortion models that apply a reduced-reference quality measure to approximate the SSIM index were established in [39] and [40], and SSIM-based RDO coding technique was presented. In [41], it is shown that the main difference between SSIM and MSE may be well accounted for by a locally adaptive divisive normalization process, which leads to a series of divisive normalization-based video coding schemes [42]–[45] on the platforms of H.264/AVC and HEVC.

In this paper, we propose a perceptual two-pass VBR scheme based on the SSIM-inspired divisive normalization video coding mechanism. In particular, adaptive GoP, frame, and CTU level rate control schemes are proposed by transforming the prediction residuals into a perceptually uniform space. The contributions of this paper are as follows.

1) Based on the SSIM-inspired divisive normalization, the prediction residuals are transformed into a perceptually uniform space in HEVC, within which we perform the GoP, frame, and CU level rate control.
2) At the GoP level, the RD performance is optimized by dynamically balancing the λ value of each GoP, which is derived adaptively by statistical perceptual distortion and rate models. In this manner, the perceptually more important GoPs are coded with more bits and vice versa, leading to better RD performance.
3) At the frame level, the sum of absolute transformed differences (SATD) in divisive normalization domain is applied to model the frame complexity, and the encoding QP for each frame is adaptively derived based on the assigned coding bits.

II. Divisive Normalization-Based Perceptual Video Coding

Following the divisive normalization framework [42]–[44], [46]–[50], the discrete cosine transform (DCT) transform coefficient of a residual block C_k is normalized with a positive normalization factor f:

$$C(k)' = C(k)/f. \tag{3}$$

As such, the quantization process of the normalized residuals for a given predefined Q_s can be formulated as

$$\begin{aligned} Q(k) &= \text{sign}\{C(k)'\}\text{round}\left\{\frac{|C(k)'|}{Q_s} + p\right\} \\ &= \text{sign}\{C(k)\}\text{round}\left\{\frac{|C(k)|}{Q_s \cdot f} + p\right\} \end{aligned} \tag{4}$$

where p is the rounding offset in the quantization.

At the decoder, the dequantization and the reconstruction of $C(k)$ are performed as

$$\begin{aligned} R(k) &= R(k)' \cdot f = Q(k) \cdot Q_s \cdot f \\ &= \text{sign}\{C(k)\}\text{round}\left\{\frac{|C(k)|}{Q_s \cdot f} + p\right\} \cdot Q_s \cdot f. \end{aligned} \tag{5}$$

This implies that the QPs for each CU can be adaptively adjusted according to the divisive normalization process. The factor f, which accounts for the perceptual importance, is derived from the SSIM index in the DCT domain [51]

$$\begin{aligned} \text{SSIM}(\mathbf{x}, \mathbf{y}) &= \left(1 - \frac{(X(0) - Y(0))^2}{X(0)^2 + Y(0)^2 + N \cdot C_1}\right) \\ &\quad \times \left(1 - \frac{\frac{\sum_{k=1}^{N-1}(X(k) - Y(k))^2}{N-1}}{\frac{\sum_{k=1}^{N-1}(X(k)^2 + Y(k)^2)}{N-1} + C_2}\right) \end{aligned} \tag{6}$$

where X and Y represent the DCT coefficients of the original and distorted blocks, respectively, N denotes the size of the block, and C_1 and C_2 are constants according to the definition of SSIM index [25]. Assuming that each CU contains l DCT blocks (such as 4 × 4), the normalization factors for dc and ac coefficients are, therefore, defined as

$$f_{\mathrm{dc}} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{X_i(0)^2+Y_i(0)^2+N\cdot C_1}}{\mathbb{E}\left(\sqrt{X(0)^2+Y(0)^2+N\cdot C_1}\right)} \tag{7}$$

$$f_{\mathrm{ac}} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1}(X_i(k)^2+Y_i(k)^2)}{N-1}+C_2}}{\mathbb{E}\left(\sqrt{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1}+C_2}\right)} \tag{8}$$

where $\mathbb{E}(\cdot)$ denotes the expectation operation over the whole frame.

Following the divisive normalization process, a new distortion model that is consistent with the residual normalization process is defined to replace the conventional distortion measures, such as SAD and MSE. In particular, the distortion model is defined as the SSD between the normalized DCT coefficients. Therefore, based on (2), the RDO problem is given by

$$\begin{aligned}\min\{J\} \text{ where } J &= \sum_{i=1}^{l}\sum_{k=0}^{N-1}(C_i(k)' - R_i(k)')^2 + \lambda_c \cdot R \\ &= \frac{\sum_{i=1}^{l}(X_i(0)-Y_i(0))^2}{f_{\mathrm{dc}}^2} \\ &\quad + \frac{\sum_{i=1}^{l}\sum_{k=1}^{N-1}(X_i(k)-Y_i(k))^2}{f_{\mathrm{ac}}^2} + \lambda_c \cdot R\end{aligned} \tag{9}$$

where λ_c indicates the Lagrange multiplier defined in HEVC codec, which is usually specified by the predefined quantization step Q_s and modified by the reference level. As the distortion model calculates the SSD between the normalized original and distorted DCT coefficients, the Lagrange multiplier defined in HEVC λ_c is still applied in the divisive normalization process.

The divisive normalization process transfers the perceptual importance to the transform coefficients so that the coefficients with lower normalization factors correspond to higher perceptual importance, and vice versa. However, there are several critical limitations in the previous method.

1) The previous divisive normalization scheme only considers the perceptual characteristics within one frame. However, as the content of video sequences evolve over time, how to achieve divisive normalization across the whole video sequences needs to be addressed.
2) The previous divisive normalization scheme did not consider the constraint on the permissible coding bits. A convenient way to implement divisive normalization is to adjust the QP values. However, the QP values will in turn determine the coding rate. As a result, precise rate control is difficult if the previous direct divisive normalization method is used. Therefore, a practical problem is how to achieve efficient VBR rate control within the perceptually uniform space specified by divisive normalization.
3) The previous divisive normalization scheme is not standard compatible. In other words, the decoder has to be changed in order to decode the bitstream. This is a very significant drawback that may hinder the practical adoption of the divisive normalization idea.

In this paper, all of these issues have been addressed. As such, the RD performance can be significantly improved, and also accurate rate control can be achieved in the VBR coding scenario.

III. Two-Pass VBR Coding

In practical applications, the video content is usually nonstationary, as it may frequently vary from one scene to another or from one frame to another. Intuitively, the QPs should be different across GoPs and frames, so that more bits can be allocated to the frames with higher complexity or perceptual importance. It is widely recognized that maintaining a constant MSE/PSNR does not ensure constant visual quality, as they perform poorly on cross-content visual quality prediction. In Section II, the divisive normalization scheme targets at transforming the prediction residuals into a perceptually uniform space within a frame. To extend this philosophy to a video sequence level, a VBR coding scheme is introduced, which aims to generate variable rate output bitstreams subject to the constraints on the dynamic ranges of bitrate and buffer size. Compared with CBR coding, the advantage of VBR lies in that it can further improve the overall coding efficiency. However, in order to optimally allocate the bit budget into different GoPs and frames, the encoder needs access to the statistics of each frame before the actual encoding. As a result, look-ahead processing is adopted to meet this requirement [18]. In particular, we perform actual encoding with a constant QP in the first pass to collect the statistics.

The flowchart of the two-pass rate control algorithm is presented in Fig. 1. The first pass encoding is performed with a fixed QP, and the statistics are recorded for the second pass. Before the second pass encoding, the scene complexity model is employed for GoP bit allocation by establishing the Laplacian distribution-based R-Q and D-Q relationships. Subsequently, the optimal number of coding bits is assigned to each GoP, which is further adjusted during the second pass encoding process. When encoding each frame, the frame-level R-Q model is established by the frame complexity estimation method in the divisive normalized domain, and the corresponding QP is finally obtained for each frame. At the CU level, each CU is then encoded with the derived QP and divisive normalization factor.

To be consistent with the default HEVC settings, in this paper, the GoP sizes in low-delay (LD) and random access (RA) configurations are 4 and 8 frames, respectively. Such GoP structure can also be termed as rate-GoP [11], [52], which allows a flexible hierarchical reference structure to improve the coding performance. In particular, the frames that will get more referenced are better when coded with lower QP values to ensure the overall coding performance. The reference structures in LD and RA settings are demonstrated in Fig. 2,

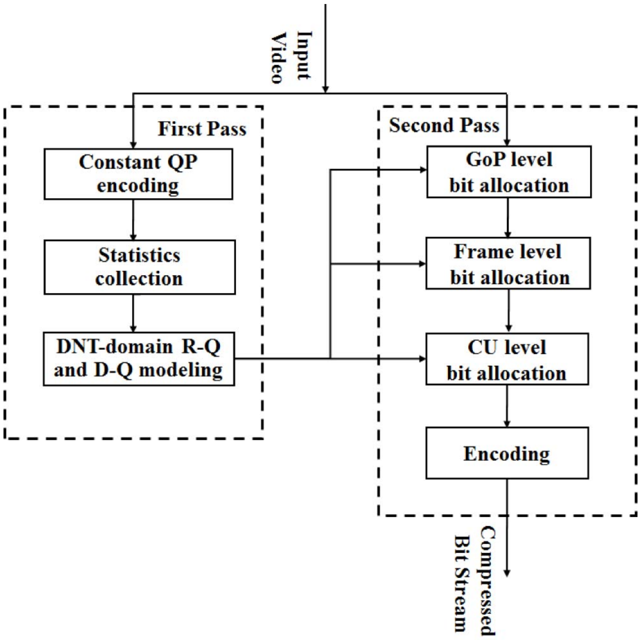
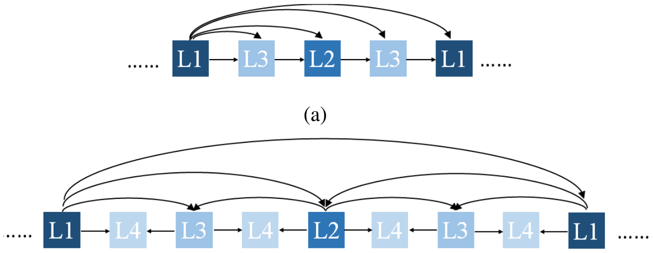


Fig. 1. Flowchart of the proposed two-pass rate control algorithm.


...... L1 L3 L2 L3 L1 ......
..... L1 L4 L3 L4 L2 L4 L3 L4 L1 ......

(a)

(b)

Fig. 2. Hierarchical reference structures for (a) LD and (b) RA.

where L1 and L3/L4 refer to the most and the least important layers, respectively.

A. GoP Level Bit Allocation

First, we treat a GoP as an individual time segment for bit allocation. In this paper, we formulate the optimal bit allocation for perceptual VBR coding as

$$\min\{D\} \text{ subject to } \sum_{i=1}^{n} R_i \leq R_c \tag{10}$$

where R_i represents the coding rate for each GoP, and R_c is the constraint on the total permissible rate.

In the literature, the distortion criteria can be classified into two categories: minimum distortion variance [53]–[55] and minimum average distortion (minAVG) [56]–[58]. In general, most minAVG methods can allocate more coding bits to frames with higher complexity. However, perceptual cues are not considered. Here, the overall quality of the whole video is defined as the average distortion in terms of the SSIM-based divisive normalized MSE for each GoP. As such, the CU level divisive normalization principle is naturally extended to GoP level, resulting in a perceptually uniform space over the whole video sequence. In particular, the minAVG criterion is formulated as

$$\min\{D\} \text{ where } D = \sum_{i=1}^{n} D_i \tag{11}$$

where D_i denotes the MSE in the divisive normalization domain for the ith GoP. Since the statistics of the input video can be obtained in the first pass coding, the perceptually uniform space is constructed at the scale of the whole sequence, within which the expectation quantity $\mathbb{E}(\cdot)$ in (8) is estimated. According to the divisive normalization process, the frames with the same MSE in the pixel domain may produce different D_i values, as smaller normalization factors are assigned to the more important frames from a perceptual optimization perspective.

Assume that the Lagrange multiplier of the ith GoP is λ_i. Theoretically, the optimal λ_i is obtained by calculating the derivative of the RD cost J_i with respect to R_i, then setting it to zero. From (2), it is formulated as

$$\frac{dJ_i}{dR_i} = \frac{d(D_i + \lambda R_i)}{dR_i} = \frac{dD_i}{dR_i} + \lambda_i = 0 \tag{12}$$

which leads to

$$\lambda_i = -\frac{dD_i}{dR_i}. \tag{13}$$

To achieve the minimization of D, the optimal strategy is to maintain a constant level of λ_i for all GoPs [57]

$$\lambda_1 = \lambda_2 = \cdots = \lambda_i = \lambda_n. \tag{14}$$

A brief proof of this solution in the scenario of VBR coding is shown in the Appendix. The philosophy behind this approach is that for all GoPs, regardless of the content, the slope of the R-D curve should be the same. In other words, on the same variation of bit rate, the change of distortion for each GoP should be approximately equal to each other.

To find the optimal λ value, we start with an initial guess and iteratively adjust it until the best λ^* is obtained with the constraint $\sum_{i=1}^{n} R_i(\lambda^*) = R_c$ [59]. It is noted that λ here is not λ_{HEVC} as specified by the encoder. For example, in HM codec, λ is only determined by the frame type, QP, and frame level, regardless of the properties of the video content. In view of various video contents, the λ derivation should be adapted to the properties of the input sequences (statistical properties of residuals, structural information, and so on) [39], [60]. For the same QP value with different residual energies, the optimal λ spans a wide range [42].

To derive λ-Q and R-Q relationship for each GoP, statistical models of both rate and distortion should be established. In video coding, Laplace distribution, which is a special case of the Generalized Gaussian distribution, shows a good trade-off between model precision and computational complexity. The density of the transformed residuals x in divisive normalization domain that is modeled with Laplace distribution is given by

$$f_{\text{Lap}}(x) = \frac{\Lambda}{2} \cdot e^{-\Lambda \cdot |x|} \tag{15}$$

where Λ is called the Laplacian parameter.

Considering the quantization process with quantization step Q[1] and rounding offset γ, the distortion and rate can be modeled as [39], [60]

$$D = \alpha\cdot\left(\int_{-(Q-\gamma Q)}^{(Q-\gamma Q)} x^2 f_{\text{Lap}}(x)dx + 2\sum_{n=1}^{\infty}\int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q}(x-nQ)^2 f_{\text{Lap}}(x)dx\right)$$

$$R = \beta\cdot\left(-P_0\cdot\log_2 P_0 - 2\sum_{n=1}^{\infty}P_n\cdot\log_2 P_n\right) \qquad (16)$$

where α and β are control parameters to ensure the accuracy of the estimation, which are estimated by the true coding bits and distortion in the first pass. The probabilities of the transformed residuals that are quantized to the zeroth and nth quantization levels P_0 and P_n are computed based on the Laplace distribution

$$P_0 = \int_{-(Q-\gamma Q)}^{(Q-\gamma Q)} f_{\text{Lap}}(x)dx$$

$$P_n = \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} f_{\text{Lap}}(x)dx. \qquad (17)$$

Finally, closed-form solutions for (16) are derived as follows:

$$D = \frac{\alpha\Lambda Q\cdot e^{\gamma\Lambda Q}(2+\Lambda Q-2\gamma\Lambda Q)+2-2e^{\Lambda Q}}{\Lambda^2(1-e^{\Lambda Q})}$$

$$R = \frac{\beta}{\ln 2}\cdot\left\{-\left(1-e^{-(1-r)\Lambda Q}\right)\ln\left(1-e^{-(1-r)\Lambda Q}\right)+e^{-(1-\gamma)\Lambda Q}\times\left(\ln 2-\ln(1-e^{\Lambda Q})-\gamma\Lambda Q+\frac{\gamma Q}{1-e^{-\Lambda Q}}\right)\right\}. \qquad (18)$$

The final Lagrange multiplier can be determined by incorporating the closed-form solutions of R and D into (13).

Therefore, the efficient bit allocation is based upon the statistics collected from the first pass compression, including the following.

1) The number of coding bits of each frame.
2) The coding distortion of each frame, which is evaluated in terms of the divisively normalized MSE

$$\tilde{D} = \frac{\sum_{i=1}^{\kappa_b}\sum_{k=0}^{N-1}\frac{(X_i(k)-Y_i(k))^2}{f_i^2}}{\kappa_b\cdot N} \qquad (19)$$

where κ_b denotes the total number of blocks in each frame, and f_i denotes the locally adaptive divisive normalization factor to establish the perceptually uniform space across the whole sequence.
3) The QP of each frame.
4) The frame-level Laplacian parameter Λ that models the transformed residuals in the divisive normalization domain.

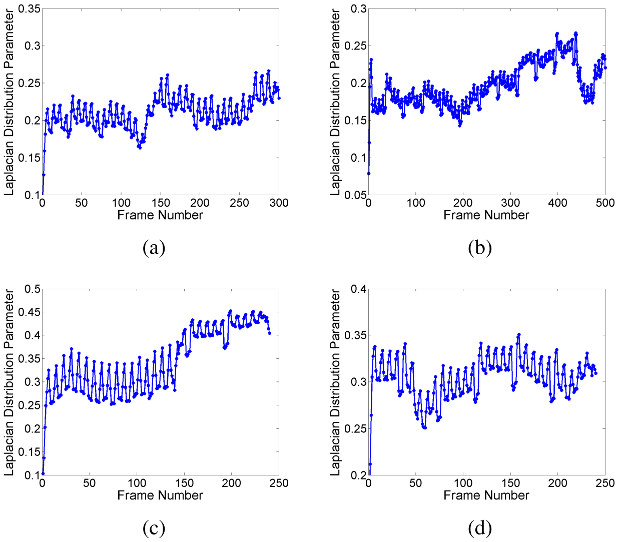


Fig. 3. Variations of laplapian distribution parameter. (a) *RaceHorses*@WQVGA. (b) *PartyScene*@WVGA. (c) *Kimono*@1080P. (d) *ParkScene*@1080P. Frame number is in coding order.

As the Laplacian parameter is computed in divisive normalized domain, this implies that the derivation of λ automatically considers the perceptual factors by computing it with the residual distribution in a perceptually uniform space. For example, there are two GoPs with the same prediction residual distribution but different perceptual importance, in which the first GoP is more important with a smaller normalization factor. This results in different Laplacian distributions, and finally leads to variations on encoding QPs.

It is observed that the distribution of the normalized transform coefficients in different scenes has different shapes. Moreover, the content of the same scene may also evolve over time. As a result, the variations of the Laplacian parameter Λ are very significant, as shown in Fig. 3. In general, a lower Λ value implies a more complex frame with larger energy of residuals. As different scenes represent different activities and motion features, the parameter Λ will vary from one scene or one GoP to another. The optimal Lagrangian multiplier derived from (13) is shown in Fig. 4, which confirms that λ_{opt} increases monotonically with Q but decreases with Λ. It should be noted that the same λ_{opt} but different Λ values correspond to different Q values.

Assuming that for GoP i and j, we have $\Lambda_i > \Lambda_j$ at QP values Q_i and Q_j. When Q_i=Q_j, the Lagrangian multiplier relationship for the low complexity GoP i and high complexity GoP j is then $\lambda_i < \lambda_j$. This indicates that for the same change of bit rate ΔR, we have $\Delta D_i < \Delta D_j$, where ΔD denotes the change of distortion. To achieve the optimal solution on the minimization of the overall distortion, more bits should be allocated to GoP j than GoP i, so that the overall distortion can be minimized. One feasible way of achieving this is to decrease Q_j and increase Q_i. As λ increases monotonically with Q, decreasing Q_j and increasing Q_i will narrow the gap between λ_i and λ_j, until the convergence point $\lambda_i = \lambda_j$. Otherwise, it is always beneficial to perform bit allocation to

[1] Here, Q specifies the quantization step of the entire GoP, and frame-level quantization step is obtained by altering it based on the reference level [4].

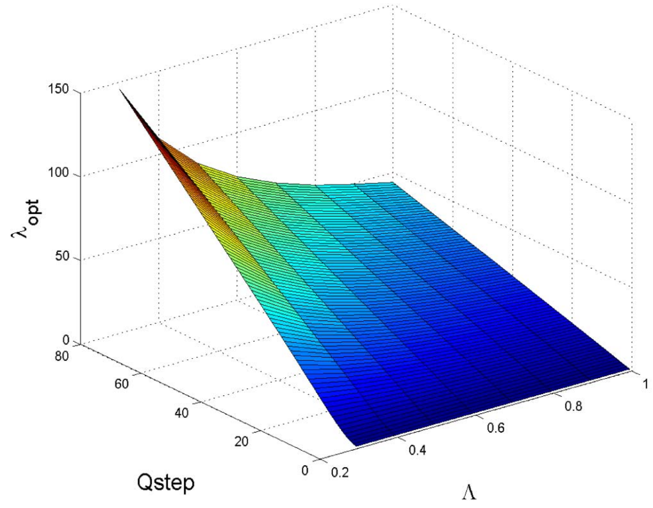


Fig. 4. Optimal λ as a function of Λ and Q [42].

achieve better overall quality. Therefore, with λ equalization, more coding bits can be automatically allocated to the GoP with higher priority, so that the perceptual quality of these frames and the whole video sequence are finally improved.

B. Frame-Level Rate Control

Given the bit rate distribution for each GoP, the task of the frame-level rate control is to derive an appropriate QP value for each frame. To maintain the hierarchical reference structure within each GoP, we adopt a similar strategy as in [14], such that more important frames according to the reference structure in each GoP can be allocated more bits. In particular, the frame-level bit allocation is performed as

$$\bar{R}_n = \frac{R_{\mathrm{GoP}} - R_{C\mathrm{GoP}}}{\sum_{i \in \Omega_{\mathrm{NC}}} \omega_i} \cdot \omega_n \tag{20}$$

where ω_n denotes the weight for the current nth frame and Ω_{NC} is the frame set of uncoded frames in the current GoP. R_{GoP} is the target bits for the current GoP and $R_{C\mathrm{GoP}}$ is the consumed bits for the already coded frames. The weights ω for each frame are defined as in [14]. It is worth noting that although the weights ω are the same as that defined in [14], because the derived R_{GoP} values for each GoP are different in [14] and the proposed rate control methods, the resulting $\bar{R}_n$ values are not identical. For example, for a sequence with 1000 frames, we may have 125 GoPs in RA setting, and the allocated bits for each GoP between the two schemes may be different. In particular, as the R_{GoP} value in the proposed scheme is obtained by considering the perceptual and the residual characteristics, more bits will be allocated into the perceptually more important GoPs so that better coding performance can be achieved.

In practice, a buffer constraint has to be applied to ensure that any burst-of-data are limited to a controllable degree [61]. In particular, assume that t_0 is the decoding delay in terms of the frame number and R_{avg} is the average bits allocated to each frame, and then the occupancy status of the decoder buffer at time instance t (in frames) is described as

$$B_t = \begin{cases} t \cdot R_{\mathrm{avg}} - \sum_{i=1}^{t-t_0} R_i, & \text{if } t \geq t_0 \\ t \cdot R_{\mathrm{avg}}, & \text{otherwise.} \end{cases} \tag{21}$$

To avoid underflow and overflow at the decoder, the buffer occupancy should satisfy

$$0 \leq B_t \leq B_c \tag{22}$$

where B_c is the buffer capacity. This would generate the upper and lower bounds of the coding bits for each frame. In practice, to meet this constraint, the number of target bits is clipped as follows:

$$R_n = \min\{R_{\mathrm{UB}}, \max\{\bar{R}_n, R_{\mathrm{LB}}\}\} \tag{23}$$

where R_{UB} and R_{LB} are the upper and lower bounds that are derived from (21) and (22). In this process, the buffer model may revise the target bits at the frame level, such that the number of the target coding bits may deviate from the desired amount. Fortunately, the buffer constraint is in effect only when buffer overflow or underflow occurs. If $R_{\mathrm{LB}} \leq \bar{R}_n \leq R_{\mathrm{UB}}$, then the buffer constraint is ineffective, and we have $R_n = \bar{R}_n$. In words, as long as the number of allocated bits is within a certain range, the buffer status would be safe.

After obtaining R_n, accurate and feasible rate model is required to automatically compute the QP given a target bit rate. Though (18) provides a solution in modeling the R-Q relationship, it is difficult to directly compute the QP from the input R. In this approach, both the frame complexity and perceptual importance should be considered in the R-Q model. Following the RD analysis in H.264/AVC [62], HEVC [6], [13], and TM5 [7], we apply SATD in divisive normalization domain for complexity modeling, which can be formulated as

$$R = \xi \cdot \chi / Q \tag{24}$$

where ξ is the model parameter and χ denotes the relative complexity computed by

$$\chi = \left(\frac{\Theta_n}{\Theta_{n-1}}\right)^{\eta} \cdot R'_{n-1} \cdot Q_{n-1}. \tag{25}$$

Here, n denotes the frame number of the currently to be encoded frame and R'_{n-1} is the actual number of coding bits of the previously encoded frame. Θ_n denotes the accumulated complexity obtained from the first to the current nth frames

$$\Theta_n = \frac{\sum_{i=0}^{n} 0.5^{n-i} \cdot \mathrm{DN\_SATD}_i}{\sum_{i=0}^{n} 0.5^{n-i}} \tag{26}$$

where $\mathrm{DN\_SATD}_i$ denotes the SATD in the divisive normalization domain. The parameter η is a constant and set to be 0.4. The relative weight of each frame $\mu_i = 0.5^{n-i} / \sum_{i=0}^{n} 0.5^{n-i}$ ensures that it decreases exponentially as the distance between the ith to the currently encoded nth frame increases. In Fig. 5, we plot the trend of μ_i when $n - i$ ranges from 0 to 10 for $n = 100\,000$. It is observed that the first few frames that are close to the nth frame play a key role in Θ_n, whereas the influences of the frames with long distances are negligible.

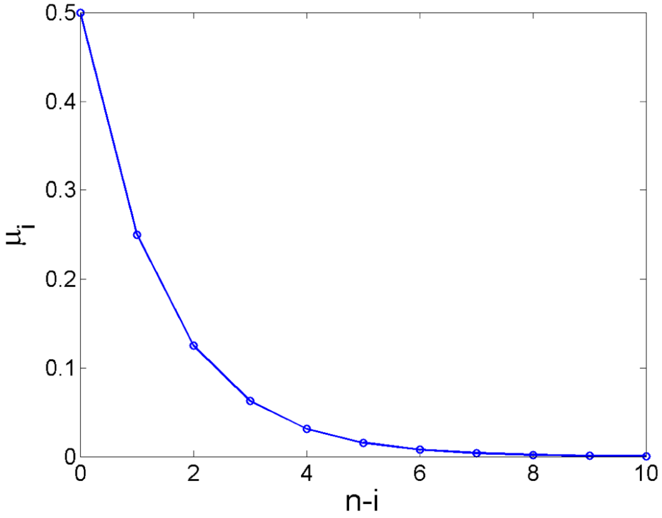


Fig. 5. Relationship between u_i and $n - i$.

In practice, the accumulated complexity can be obtained iteratively in the following fashion:

$$\Theta_n = \frac{\sum_{i=0}^{n} 0.5^{n-i} \cdot \mathrm{DN\_SATD}_i}{\sum_{i=0}^{n} 0.5^{n-i}} = \frac{\sum_{i=0}^{n-1} 0.5^{n-i} \cdot \mathrm{DN\_SATD}_i + \mathrm{DN\_SATD}_n}{\sum_{i=0}^{n-1} 0.5^{n-i} + 1}. \quad (27)$$

Therefore, after encoding the $(n-1)$th frame, the values of $\sum_{i=0}^{n-1} 0.5^{n-i} \cdot \mathrm{DN\_SATD}_i$ and $\sum_{i=0}^{n-1} 0.5^{n-i}$ are stored for the computation of Θ_n, such that Θ_n can be calculated iteratively with low complexity.

The parameter $\mathrm{DN\_SATD}_i$ estimates the perceptual complexity at the frame level by computing the SATD in the divisive normalization domain. As such, the residuals with more perceptual importance are amplified because of the lower divisive normalization factors. In Fig. 6, the mismatch of the generated bits and estimated bits with the divisive normalization domain SATD is demonstrated. It shows that the discrepancy per frame is relatively small, which verifies the accuracy of the rate model.

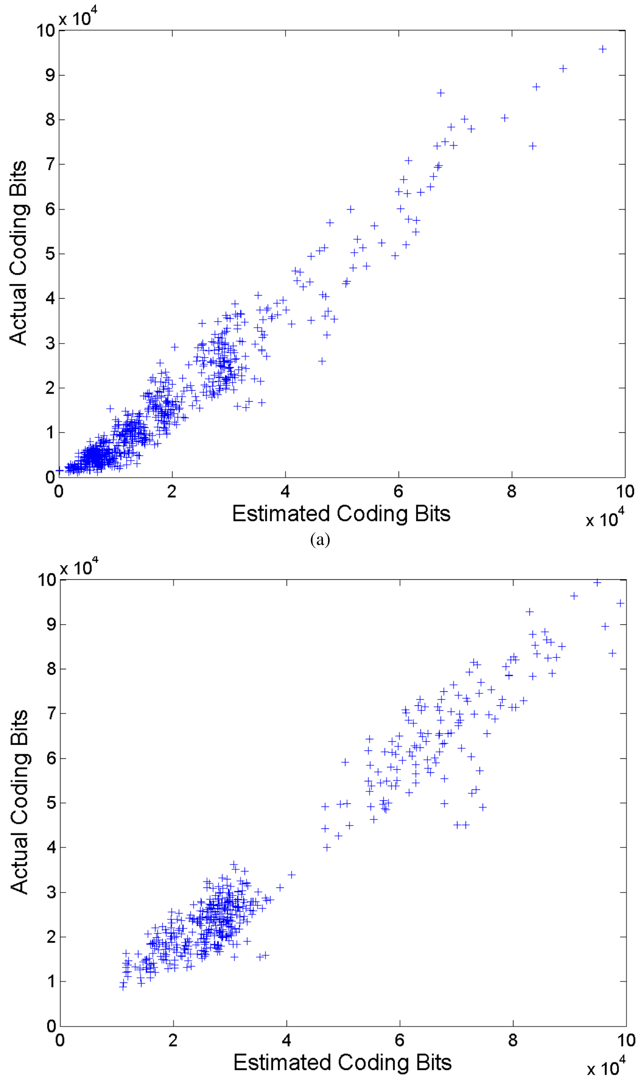
Finally, given the target bit rate R_n for the current frame, the corresponding quantization step is calculated as

$$Q_n = \frac{\xi \cdot \chi}{R_n}. \quad (28)$$

C. CU Level QP Adjustment

The CU level QP adjustment is performed by dynamically assigning each CU an appropriate ΔQP value according to its relative importance. Since the frame-level coding bits are derived in the perceptually uniform domain, it becomes natural to perform divisive normalization for each CU. In particular, the divisive normalization factor for each CU is given by

$$f = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1} 2X_i(k)^2}{N-1} + C_2}}{\mathbb{E}\left(\sqrt{\frac{\sum_{k=1}^{N-1} 2X(k)^2}{N-1} + C_2}\right)}. \quad (29)$$



(a)

(b)

Fig. 6. Actual versus estimated coding bits. (a) HEVC test sequences (Class D in CTC). (b) HEVC test sequences (Class E in CTC).

Again, as in (8), l denotes the number of DCT blocks in each CU, N denotes the size of the block, and X represents the DCT coefficients of the original blocks from the input frame. It is worth mentioning that to be compatible with the HEVC standard, only ac coefficients are applied to derive the divisive normalization factor. Moreover, the applied divisive normalization factor is slightly different from the one directly derived from the SSIM index because before coding the current frame, the distorted frame is not available. Therefore, the distorted signal Y is replaced by the original signal X in the calculation of the divisive normalization factors.

From (29), it is observed that the spatially adaptive divisive normalization factor is highly dependent on the content of the local CU, and further determines its relative perceptual importance. To make the scheme fully standard compatible, as specified by HEVC, the ΔQP at CU level is signaled into the bitstream. In particular, assuming that the derived QP from the target coding bits for the current nth frame is QP_n,

TABLE I

PERFORMANCE COMPARISON BASED ON THE R-λ METHOD [14] (LDB_MAIN)

Sequences (Seq1∼Seq7)	R_{target} (kbps)	Anchor				Proposed				ΔR^{*}	ΔR^{**}
		R_{actual} (kbps)	SSIM	MS-SSIM	$BitErr$	R_{actual} (kbps)	SSIM	MS-SSIM	$BitErr$		
Basketballpass@WQVGA	859.37	859.18	0.9461	0.9910	0.02%	859.19	0.9524	0.9926	0.02%	-9.25%	-11.34%
BlowingBubbles@WQVGA	393.53	393.45	0.8965	0.9786	0.02%	393.40	0.9053	0.9819	0.03%		
BQSquare@WQVGA	180.46	180.42	0.8222	0.9530	0.02%	180.43	0.8319	0.9579	0.01%		
RaceHorses@WQVGA	85.75	85.74	0.7404	0.9132	0.00%	85.77	0.7475	0.9180	0.03%		
Coastguard@CIF	642.42	642.37	0.9618	0.9919	0.01%	643.95	0.9580	0.9908	0.24%	-15.6%	-19.36%
Container@CIF	283.81	283.81	0.9281	0.9830	0.00%	283.73	0.9367	0.9857	0.03%		
Flower@CIF	126.44	126.44	0.8789	0.9678	0.00%	126.37	0.8920	0.9730	0.05%		
News@CIF	55.31	55.31	0.8066	0.9409	0.00%	55.30	0.8234	0.9494	0.02%		
Flowervase@WVGA	1527.82	1527.41	0.9384	0.9774	0.03%	1528.00	0.9547	0.9896	0.01%	-55.7%	-73.96%
Keiba@WVGA	664.59	664.42	0.9048	0.9564	0.03%	664.54	0.9414	0.9857	0.01%		
Mobisode@WVGA	304.07	304.07	0.8729	0.9279	0.00%	303.69	0.9085	0.9721	0.12%		
RaceHorses@WVGA	144.30	144.40	0.8436	0.9019	0.07%	144.28	0.8642	0.9465	0.02%		
Mobcal@720P Parkrun@720P Shields@720P	12618.54	12618.61	0.9299	0.9882	0.00%	12615.32	0.9390	0.9907	0.03%	-41.4%	-52.58%
	5116.98	5117.10	0.8942	0.9778	0.00%	5115.71	0.9080	0.9825	0.02%		
	2050.96	2051.03	0.8476	0.9593	0.00%	2050.79	0.8836	0.977	0.01%		
	810.00	805.72	0.7882	0.9297	0.53%	811.18	0.8142	0.9544	0.15%		
BigShip@720P Raven@720P ShuttleStart@720P	3141.34	3141.43	0.9570	0.9897	0.00%	3141.96	0.9605	0.9908	0.02%	-13.9%	-10.09%
	1172.40	1172.02	0.9346	0.9809	0.03%	1172.51	0.9392	0.9824	0.01%		
	464.30	463.57	0.8988	0.9636	0.16%	464.04	0.9051	0.9661	0.06%		
	185.17	185.17	0.8489	0.9293	0.00%	185.16	0.8578	0.9341	0.00%		
Sunflower@1080P	3854.38	3854.26	0.9521	0.9897	0.00%	3852.33	0.9561	0.9912	0.05%	-16.9%	-14.40%
Tractor@1080P	1711.08	1706.49	0.9293	0.9798	0.27%	1712.80	0.9359	0.9824	0.10%		
Kimono@1080P	793.85	790.10	0.8953	0.9612	0.47%	795.63	0.9050	0.9661	0.22%		
ParkScene@1080P	383.60	384.94	0.8536	0.9295	0.35%	383.89	0.8632	0.9369	0.08%		
Cactus@1080P BasketballDrive@1080P Crowd_run@1080P	18201.93	18201.98	0.9171	0.9863	0.00%	18202.30	0.9204	0.9876	0.00%	-20.1%	-20.93%
	8125.05	8125.07	0.8831	0.9756	0.00%	8127.74	0.8985	0.9811	0.03%		
	3850.69	3850.71	0.8404	0.9572	0.00%	3853.43	0.8547	0.9644	0.07%		
	1863.99	1863.99	0.7898	0.9282	0.00%	1864.80	0.7956	0.9333	0.04%		

* Rate reduction while maintaining SSIM.

** Rate reduction while maintaining MS-SSIM.

it is given by

$$\Delta \mathrm{QP} = Q\mathrm{step2QP}(\mathrm{QP2}Q\mathrm{step}(\mathrm{QP}_n)\cdot f) - \mathrm{QP}_n \tag{30}$$

where Qstep2QP(·) and QP2Qstep(·) refer to the mapping function between QP and the quantization step. This implies that the CUs that are less important are quantized more coarsely as compared to the more important CUs. In this manner, the bits from the regions that are perceptually less important, are borrowed and assigned to the regions with more perceptual relevance, leading to the perceptually uniform space within each frame. This provides the foundation of the proposed rate control algorithm, such that the optimization in GoP, frame, and CU levels is all achieved in the divisive normalization domain.

IV. VALIDATIONS

To validate the proposed scheme, we integrate it into the HEVC reference software HM13.0 [63]. All test video sequences are in the YCbCr 4:2:0 format. Two categorizes of video sequences are used in the experiment. First, we verify the proposed scheme to encode the video sequences that are concatenated by different video shots, which can better reflect the cross-content quality prediction ability of the employed measure. Subsequently, the performance of the proposed algorithm is evaluated on test sequences in HEVC common test condition (CTC) to further demonstrate the rate control performance.

A. Performance Evaluation of the Proposed Algorithm

The performance is evaluated in terms of Bjontegaard (BD)-Rate [64] and rate control accuracy. In particular, the rate control accuracy is measured in terms of the bit rate error

$$\mathrm{BitErr} = \frac{|R_{\mathrm{target}} - R_{\mathrm{actual}}|}{R_{\mathrm{target}}} \times 100\%. \tag{31}$$

The performance of the proposed scheme in terms of the BD-Rate and rate control accuracy is summarized in Tables I and II, where coding configurations RA Main (RA_Main) and LD_B Main (LDB_Main) are tested. Each test video is generated by concatenating three or four video shots with different statistical properties but the same frame rate. The test sequences cover various resolutions from WQVGA to 1080P. The names of these sequences are simplified as Seq1∼Seq7, following the order of Tables I and II. In these experiments, the rate control performance of the CBR R-λ model (anchor) in the HM software [14] and the proposed SSIM-motivated rate control strategy (proposed) are compared. Moreover, the CTU level rate control in HM software is also applied. Both SSIM and MS-SSIM [65] are used as distortion measures when calculating the BD-Rate. It can be observed that the proposed scheme can significantly improve the SSIM and Multi-scale SSIM (MS-SSIM) indices at the similar bit rate. When evaluating the performance with BD-Rate, on average in terms of SSIM, 24.7% bit rate reduction for LDB_Main and 21.7% bit rate reduction for RA_Main are observed. This is because of the unified construction of

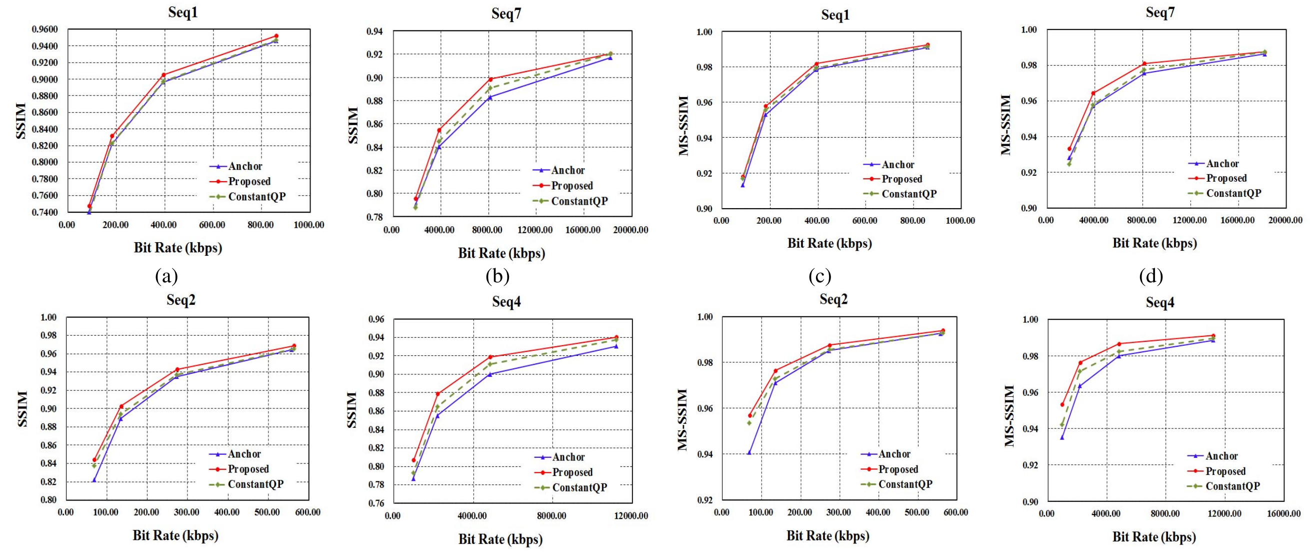
TABLE II

PERFORMANCE COMPARISON BASED ON THE R-λ METHOD [14] (RA_MAIN)

Sequence (Seq1∼Seq7)	R_{target} (kbps)	Anchor				Proposed				ΔR^{*}	ΔR^{**}
		R_{actual} (kbps)	SSIM	MS-SSIM	$BitErr$	R_{actual} (kbps)	SSIM	MS-SSIM	$BitErr$		
Basketballpass@WQVGA	756.67	757.57	0.9486	0.9913	0.12%	757.51	0.9527	0.9922	0.11%	-6.61%	-5.53%
BlowingBubbles@WQVGA	366.76	367.53	0.9039	0.9802	0.21%	367.66	0.9099	0.9817	0.25%		
BQSquare@WQVGA	180.46	181.07	0.8373	0.9581	0.34%	181.13	0.8449	0.9607	0.37%		
RaceHorses@WQVGA	90.63	91.02	0.7579	0.9222	0.43%	90.78	0.7615	0.9227	0.17%		
Coastguard@CIF	563.72	557.47	0.9644	0.9927	1.11%	563.27	0.9689	0.9940	0.08%	-16.0%	-20.13%
Container@CIF	272.79	272.70	0.9349	0.9852	0.03%	273.63	0.9430	0.9876	0.31%		
Flower@CIF	134.31	134.60	0.8893	0.9711	0.21%	135.05	0.9028	0.9765	0.55%		
News@CIF	68.39	68.65	0.8223	0.9407	0.38%	68.91	0.8445	0.9570	0.77%		
Flowervase@WVGA	1379.94	1386.57	0.9352	0.9761	0.48%	1379.14	0.9527	0.9879	0.06%	-48.0%	-63.66%
Keiba@WVGA	642.00	648.76	0.9055	0.9550	1.05%	643.59	0.9232	0.9787	0.25%		
Mobisode@WVGA	314.82	347.26	0.8876	0.9446	10.31%	315.76	0.9136	0.9738	0.30%		
RaceHorses@WVGA	156.15	178.95	0.8577	0.9238	14.60%	156.28	0.8678	0.9495	0.09%		
Mobcal@720P	11186.72	11186.81	0.9306	0.9886	0.00%	11190.06	0.9405	0.9911	0.03%	-29.9%	-42.96%
Parkrun@720P	4822.68	4822.69	0.9000	0.9800	0.00%	4827.57	0.9190	0.9866	0.10%		
Shields@720P	2179.14	2179.15	0.8551	0.9635	0.00%	2192.39	0.8788	0.9763	0.61%		
	974.10	967.76	0.7869	0.9353	0.65%	978.86	0.8070	0.9533	0.49%		
BigShip@720P	3002.67	3002.67	0.9583	0.9902	0.00%	3002.24	0.9616	0.9912	0.01%	-19.0%	-25.31%
Raven@720P	1283.18	1283.18	0.9368	0.9811	0.00%	1285.64	0.9427	0.9840	0.19%		
ShuttleStart@720P	584.52	584.84	0.9018	0.9603	0.05%	587.07	0.9132	0.9700	0.44%		
	271.36	271.57	0.8619	0.9316	0.07%	273.03	0.8722	0.9438	0.61%		
Sunflower@1080P	3706.52	3740.76	0.9539	0.9900	0.92%	3723.05	0.9572	0.9912	0.45%	-11.9%	-11.24%
Tractor@1080P	1744.67	1764.65	0.9345	0.9816	1.15%	1753.63	0.9396	0.9839	0.51%		
Kimono@1080P	863.28	873.39	0.9063	0.9669	1.17%	865.61	0.9113	0.9696	0.27%		
ParkScene@1080P	445.18	449.62	0.8682	0.9411	1.00%	447.47	0.8729	0.9452	0.51%		
Cactus@1080P	16504.51	16504.60	0.9164	0.9858	0.00%	16503.10	0.9240	0.9882	0.01%	-20.9%	-23.41%
BasketballDrive@1080P	7623.10	7639.32	0.8855	0.9756	0.21%	7622.85	0.8996	0.9812	0.00%		
Crowd_run@1080P	3731.18	3745.64	0.8445	0.9581	0.39%	3727.32	0.8588	0.9658	0.10%		
	1860.52	1873.15	0.7954	0.9309	0.68%	1860.38	0.8035	0.9375	0.01%		

[*] Rate reduction while maintaining SSIM.

[**] Rate reduction while maintaining MS-SSIM.



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

(e) (f) (g) (h)

Fig. 7. RD performance comparison in terms of SSIM and MS-SSIM for sequences in Tables I and II. (a)–(d) LDB_Main. (e)–(h) RA_Main.

the perceptually uniform space at GoP, frame, and CU levels, which jointly improve the coding performance at the expense of two-pass encoding. The bit rate errors for LDB_Main and RA_Main cases are all within 1%, enabling its applications in real scenarios.

The RD curves for the sequences in Tables I and II are provided in Fig. 7. It can be observed that the proposed algorithm has better R-D performance for both high and low bit rate coding. Moreover, the RD performance of the constant QP coding strategy is also illustrated, which usually lies between the proposed and R-λ methods. This further demonstrates the superior performance of the proposed scheme over both R-λ and the constant QP coding strategies.

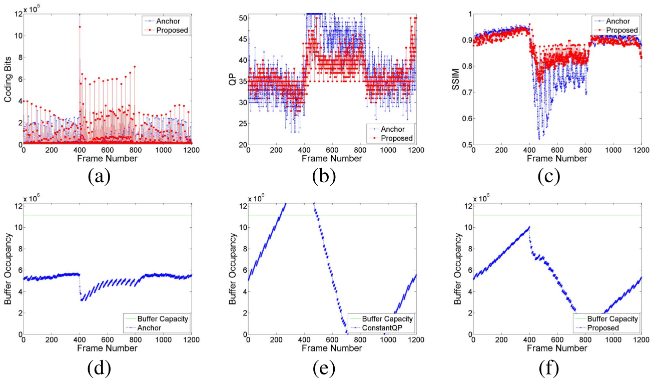
To further study the perceptual video quality-of-experience (QoE) of the proposed algorithm, an experiment is conducted to evaluate the RD performance in terms of the recently proposed SSIMplus index [66], [67]. The unique feature of SSIMplus is that it can provide device-adaptive,

TABLE III
RATE-DISTORTION PERFORMANCE EVALUATED IN TERMS OF SSIMplus

Sequence	Device	ΔR (LB_Main)	ΔR (RA_Main)
Seq1	Default	-7.38%	-2.36%
	Phone_iPhone6Plus	-6.68%	-3.50%
	Tablet_iPadAir2	-6.53%	-3.72%
	TV_F8500	-6.44%	-3.70%
	Monitor_27MP32HQ	-6.41%	-4.00%
	Laptop_MacBookPro	-6.46%	-3.81%
Seq2	Default	-12.52%	-17.34%
	Phone_iPhone6Plus	-10.23%	-11.96%
	Tablet_iPadAir2	-9.26%	-10.13%
	TV_F8500	-4.85%	-10.55%
	Monitor_27MP32HQ	-12.10%	-8.56%
	Laptop_MacBookPro	-9.11%	-10.01%
Seq3	Default	-48.20%	-45.61%
	Phone_iPhone6Plus	-46.64%	-42.62%
	Tablet_iPadAir2	-39.84%	-36.06%
	TV_F8500	-40.85%	-36.76%
	Monitor_27MP32HQ	-30.91%	-29.44%
	Laptop_MacBookPro	-37.97%	-34.72%
Seq4	Default	-41.43%	-33.34%
	Phone_iPhone6Plus	-42.06%	-31.96%
	Tablet_iPadAir2	-34.11%	-30.97%
	TV_F8500	-32.58%	-30.99%
	Monitor_27MP32HQ	-37.66%	-27.11%
	Laptop_MacBookPro	-42.57%	-30.39%
Seq5	Default	-11.24%	-11.61%
	Phone_iPhone6Plus	-11.90%	-13.45%
	Tablet_iPadAir2	-11.38%	-11.95%
	TV_F8500	-11.20%	-12.38%
	Monitor_27MP32HQ	-12.27%	-10.41%
	Laptop_MacBookPro	-11.51%	-11.84%
Seq6	Default	-7.47%	-4.80%
	Phone_iPhone6Plus	-6.55%	-5.21%
	Tablet_iPadAir2	-6.74%	-5.06%
	TV_F8500	-6.55%	-5.10%
	Monitor_27MP32HQ	-7.45%	-4.92%
	Laptop_MacBookPro	-6.84%	-4.84%
Seq7	Default	-18.57%	-21.43%
	Phone_iPhone6Plus	-14.00%	-17.56%
	Tablet_iPadAir2	-15.74%	-18.77%
	TV_F8500	-15.51%	-18.25%
	Monitor_27MP32HQ	-18.20%	-20.62%
	Laptop_MacBookPro	-16.04%	-18.82%

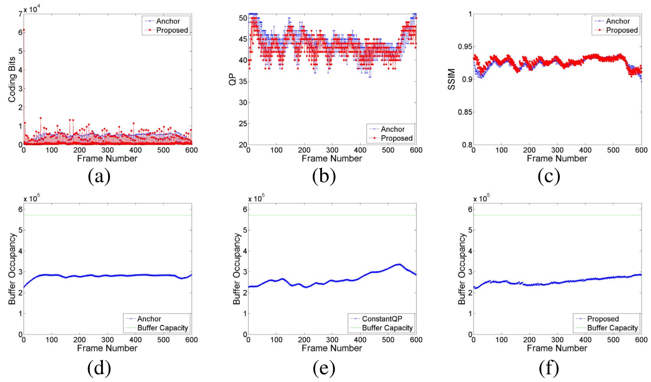
cross-resolution, and cross-content predictions of the perceptual quality in real-time, and therefore, the properties of display devices and viewing conditions are fully considered. In Table III, we demonstrate the RD performance of the proposed method in terms of different devices, including the default, iPhone, iPad, TV, monitor, and laptop. The results indicate that our method consistently improves the coding performance in different viewing environments.

To further demonstrate the performance of the proposed method, in Fig. 8, we provide the variations of the coding bits, QP, SSIM indices, and buffer status when the middle portion of the video is more complicated and perceptually important. In this scenario, if CBR coding is applied, the quality of the middle portion will be significantly decreased, as shown in Fig. 8(c) (anchor case). This may lead to perceptual quality variations and poor QoE. By contrast, the proposed VBR strategy improves the quality of the middle portion by allocating more coding bits. This is achieved by decreasing the QP values of corresponding GoPs, as shown in Fig. 8(b). Due to the constraint on the total permissible coding bits, the first and last portions, which are less complicated are hence allocated with fewer coding bits, as shown in Fig. 8(a). In Fig. 8(d)–(f), the corresponding buffer occupancies of the CBR, constant QP, and VBR strategies are demonstrated, and we can observe that both the anchor and proposed algorithms maintain that the buffer status is at a secure level. Moreover, for the proposed scheme, the low bit rate encoding of the first portion allows one to reduce the probability of rebuffing and stalling at the future complex portions such that the quality of the second video portion can be significantly improved. The SSIM indices as a function of the frame index are shown in Fig. 8(c). To quantitatively evaluate the variations, the standard deviations of the anchor and proposed schemes in terms of SSIM indices are computed, which are 0.0990 and 0.0443, respectively. As SSIM is able to efficiently predict the visual quality across different contents, lower SSIM difference between different scenes indicates lower video quality fluctuation. One can also discern that although our approach does not impose a smooth term in the quality evaluation, more bits are allocated into the middle portion, so that the reconstructed video is much smoother in quality with low SSIM variance. This originates from the divisive normalization-based rate control approach, which automatically allocates more bits to the areas that may create more perceptual distortion and therefore results in more consistent video quality over time.



(a) (b) (c)
(d) (e) (f)

Fig. 8. Demonstration of the bit consumption, buffer status, QP, and SSIM indices for Seq4 (RA_Main, target bitrate: 2179.1 kb/s). (a) Coding bits. (b) QP. (c) SSIM. (d) Buffer occupancy of CBR. (e) Buffer occupancy of constant QP. (f) Buffer occupancy of the proposed scheme.



(a) (b) (c)
(d) (e) (f)

Fig. 9. Demonstration of the bit consumption, buffer status, QP, and SSIM indices for KristenAndSara (LDB_Main, target bitrate: 111.8 kb/s). (a) Coding bits. (b) QP. (c) SSIM. (d) Buffer occupancy of CBR. (e) Buffer occupancy of constant QP. (f) Buffer occupancy of the proposed scheme.

TABLE IV

PERFORMANCE COMPARISON ON HEVC CTC SEQUENCES BASED ON THE R-λ METHOD [14]

	LDB_Main						RA_Main					
	$\Delta R_Y{}^{*}$	$\Delta R_U{}^{*}$	$\Delta R_V{}^{*}$	$\Delta R_Y{}^{**}$	$\Delta R_U{}^{**}$	$\Delta R_V{}^{**}$	$\Delta R_Y{}^{*}$	$\Delta R_U{}^{*}$	$\Delta R_V{}^{*}$	$\Delta R_Y{}^{**}$	$\Delta R_U{}^{**}$	$\Delta R_V{}^{**}$
Class A	-12.65%	-1.83%	-1.44%	-16.72%	-6.96%	-7.02%	-4.79%	-2.20%	-1.71%	-7.82%	-4.45%	-4.33%
Class B	-9.85%	-2.08%	-1.66%	-13.57%	-6.86%	-6.84%	-2.77%	-0.85%	0.14%	-4.78%	-3.07%	-1.82%
Class C	-10.83%	-5.17%	-3.46%	-14.23%	-7.52%	-5.65%	-6.12%	-2.67%	-1.29%	-8.80%	-3.91%	-2.63%
Class D	-13.07%	-7.30%	-4.11%	-15.35%	-9.97%	-7.01%	-8.28%	-3.15%	-1.94%	-9.82%	-4.28%	-3.19%
Class E	-10.79%	-4.25%	-7.23%	-13.80%	-5.01%	-7.74%						
Average	-11.39%	-4.02%	-3.30%	-14.72%	-7.35%	-6.81%	-5.33%	-2.14%	-1.12%	-7.62%	-3.88%	-2.92%

* Rate reduction while maintaining SSIM.

** Rate reduction while maintaining MS-SSIM.

TABLE V

PERFORMANCE COMPARISONS WITH CONSTANT QP ENCODING (LDB_MAIN)

	Constant QP VS R-λ						Constant QP VS Proposed					
	$\Delta R_Y{}^{*}$	$\Delta R_U{}^{*}$	$\Delta R_V{}^{*}$	$\Delta R_Y{}^{**}$	$\Delta R_U{}^{**}$	$\Delta R_V{}^{**}$	$\Delta R_Y{}^{*}$	$\Delta R_U{}^{*}$	$\Delta R_V{}^{*}$	$\Delta R_Y{}^{**}$	$\Delta R_U{}^{**}$	$\Delta R_V{}^{**}$
Class A	0.53%	0.41%	1.68%	-0.44%	1.74%	2.98%	-12.26%	-1.28%	0.31%	-17.18%	-5.47%	-4.33%
Class B	0.38%	3.39%	3.35%	2.58%	4.79%	4.70%	-9.62%	0.90%	1.20%	-11.25%	-2.51%	-2.62%
Class C	0.63%	0.17%	-0.26%	1.79%	0.81%	-0.08%	-10.02%	-5.11%	-3.30%	-12.52%	-6.88%	-5.49%
Class D	-1.65%	5.45%	4.17%	-0.86%	6.68%	4.18%	-14.43%	-2.16%	0.16%	-15.92%	-3.96%	-3.06%
Class E	3.06%	-0.33%	-0.04%	5.39%	-0.84%	-0.19%	-7.79%	-4.69%	-7.39%	-8.84%	-6.00%	-8.06%
Average	0.45%	2.00%	1.95%	1.55%	2.92%	2.56%	-10.91%	-2.19%	-1.37%	-13.27%	-4.79%	-4.44%

* Rate reduction while maintaining SSIM.

** Rate reduction while maintaining MS-SSIM.

TABLE VI

PERFORMANCE COMPARISONS WITH CONSTANT QP ENCODING (RA_MAIN)

	Constant QP VS R-λ						Constant QP VS Proposed					
	$\Delta R_Y{}^{*}$	$\Delta R_U{}^{*}$	$\Delta R_V{}^{*}$	$\Delta R_Y{}^{**}$	$\Delta R_U{}^{**}$	$\Delta R_V{}^{**}$	$\Delta R_Y{}^{*}$	$\Delta R_U{}^{*}$	$\Delta R_V{}^{*}$	$\Delta R_Y{}^{**}$	$\Delta R_U{}^{**}$	$\Delta R_V{}^{**}$
Class A	0.87%	2.76%	2.50%	0.70%	3.40%	3.14%	-3.96%	0.55%	0.79%	-7.19%	-1.21%	-1.36%
Class B	1.68%	3.75%	3.01%	3.00%	4.52%	3.56%	-1.13%	2.95%	3.24%	-1.88%	1.43%	1.76%
Class C	3.50%	2.58%	2.60%	4.66%	3.04%	2.51%	-2.85%	-0.17%	1.42%	-4.54%	-1.00%	-0.05%
Class D	2.22%	4.33%	4.40%	3.84%	5.35%	5.06%	-6.18%	1.10%	2.54%	-6.24%	0.96%	1.86%
Average	2.04%	3.38%	3.12%	3.05%	4.10%	3.57%	-3.39%	1.22%	2.07%	-4.78%	0.13%	0.62%

* Rate reduction while maintaining SSIM.

** Rate reduction while maintaining MS-SSIM.

For another example, which has low variation in content, we compare the VBR and CBR coding results in Fig. 9. As the content of the video does not show significant variation over time, the proposed method behaves similarly to the anchor approach. As such, the SSIM improvement mainly originates from the CU level divisive normalization approach, as the allocated bits for each GoP are quite similar. It is also observed that because the GoPs are coded with similar number of bits, the SSIM indices versus frame index is also very smooth for both the anchor and the proposed methods.

The rate control performances on test sequences in HEVC CTC are demonstrated in Table IV. In this experiment, the test video sequences are much simpler, as only one or two scenes are included. The bit rate reductions in Table IV illustrate that, on average, the proposed scheme achieves rate reductions of 11.4% for LDB_Main and 5.3% for RA_Main in terms of SSIM indices. For fixed MS-SSIM, the rate reductions for LDB and RA cases are 14.7% and 7.6%, respectively. The performance comparisons between constant QP coding and the rate control approaches including $R - \lambda$ and the proposed methods are demonstrated in Tables V and VI. It is observed that for HEVC CTC sequences, the $R - \lambda$ scheme cannot improve the coding performance in terms of either SSIM or MS-SSIM. By contrast, the proposed method significantly improves the performance compared with the constant QP coding configuration. Moreover, it is also noted that the bit rate reduction is not as significant as the test case when the video sequences contain large variations of content. In general, the two-pass rate control schemes work better for video sequences that include both simple and complex scenes.

B. Subjective Performance Evaluation

We further carried out two subjective quality evaluation tests based on a two-alternative-forced-choice method to verify this scheme. This method is widely adopted in psychophysical studies, where in each trial a subject is forced to choose the one he/she thinks to have better quality from a pair of video sequences. For each subjective test, we selected six pairs of sequences with different resolutions. Each pair is repeated four times in a random order. In the first test, the sequences are compressed by R-λ and the proposed methods at the same bit rate but with different SSIM levels. In the second test, the sequences were coded to achieve the similar SSIM levels (where the proposed scheme uses much lower bit rates).

TABLE VII

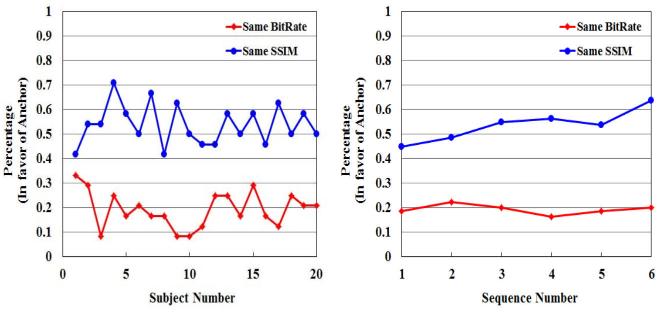
SSIM INDICES AND BIT RATES OF TESTING SEQUENCES USED IN THE SUBJECTIVE TEST (SIMILAR BIT RATE BUT DIFFERENT SSIM INDICES)

Sequences	Anchor		Proposed	
	SSIM	Bit rate (kbps)	SSIM	Bit rate (kbps)
Seq4(LDB_Main)	0.8476	2051.03	0.8836	2050.79
Seq6(LDB_Main)	0.8536	384.94	0.8632	383.89
PartyScene(LDB_Main)	0.7324	377.90	0.7511	376.87
Seq7(RA_Main)	0.7954	1873.15	0.8035	1860.38
Seq1(RA_Main)	0.9486	757.57	0.9527	757.51
BasketballDrill(RA_Main)	0.8577	587.45	0.8625	585.09

TABLE VIII

SSIM INDICES AND BIT RATES OF TESTING SEQUENCES USED IN THE SUBJECTIVE TEST (SIMILAR SSIM INDICES BUT DIFFERENT BIT RATES)

Sequences	Anchor		Proposed	
	SSIM	Bit rate (kbps)	SSIM	Bit rate (kbps)
Seq4(LDB_Main)	0.8476	2051.03	0.8495	1384.67
Seq6(LDB_Main)	0.8536	384.94	0.8537	333.98
PartyScene(LDB_Main)	0.7324	377.90	0.7341	318.69
Seq7(RA_Main)	0.7954	1873.15	0.7949	1640.15
Seq1(RA_Main)	0.9486	757.57	0.9484	696.86
BasketballDrill(RA_Main)	0.8577	587.45	0.8574	536.10



(a) (b)

Fig. 10. Subjective test results (in terms of the percentage in favor of the anchor). (a) Mean of preference for individual subject (1–20: subject number). (b) Mean of preference for individual sequence (1–6: sequence number).

Tables VII and VIII list all the test sequences as well as their SSIM indices and bit rates. In total, 20 subjects participated in the experiments.

The results of the subjective tests are reported in Fig. 10. In each figure, the percentage by which the subjects are in favor of the anchor against the proposed scheme is demonstrated. As can be observed, when the sequences are compressed with a similar bit rate, the subjects are inclined to select the proposed method for better video quality. On the contrary, for the similar quality case, it turns out that for almost all cases, the percentage is close to 50%. These results provide useful evidence that the proposed method improves the coding performance in terms of a better compromise between bit rate and subjective quality.

TABLE IX

COMPLEXITY EVALUATION OF THE PROPOSED SCHEME

Sequences	Anchor QP	LDB	RA
Seq1(WQVGA)	25	210.80%	214.20%
	40	213.80%	216.40%
Seq3(WVGA)	25	208.60%	212.60%
	40	211.00%	217.60%
Seq6(1080P)	25	211.40%	212.40%
	40	211.80%	211.40%
Average		**211.23%**	**214.10%**

C. Encoding Complexity Evaluation

We evaluate the complexity of the proposed scheme in terms of the actual encoding time. In particular, the computational complexity ΔT is evaluated as

$$\Delta T = \frac{T_{\text{pro}}}{T_{\text{org}}} \times 100\% \tag{32}$$

where T_{org} is the encoding time of the one-pass constant QP coding. T_{pro} is the encoding time of the proposed two-pass method with the target bit rate generated by the one-pass constant QP coding.

The computational complexity comparison is reported in Table IX in which both high bit rate and low bit rate coding are tested. The sequences from WQVGA to 1080P are evaluated in both RA and LDB cases. The test was carried out on an Intel 3.40-GHz Core processor with 12-GB RA memory. Compared with constant QP coding, on average the computation complexity of the proposed method is 211.2% for LDB and 214.1% for RA cases. In addition to the two-pass encoding, the added complexity overhead is mainly due to the calculation of the divisive normalization factor, GoP level bit allocation, and the frame/CU level QP values.

V. CONCLUSION

We propose an SSIM-motivated perceptual two-pass VBR rate control scheme for HEVC, aiming to optimize the overall quality of video sequences under the bit rate budget. The novelty of our approach lies in the hierarchical construction of a perceptually uniform space at GoP, frame, and CU levels based on the SSIM-inspired divisive normalization mechanism. The superior performance of the proposed scheme is demonstrated using the reference software HM whereby the proposed method achieves significantly higher coding efficiency. Visual quality improvement is also observed when compared with the conventional schemes.

APPENDIX

From (10) and (11), the constrained optimization problem for GoP level bit allocation can be formulated as follows:

$$\min\Big\{\sum_{i=1}^{n} D_i(R_i)\Big\} \text{ subject to } \sum_{i=1}^{n} R_i \le R_c \tag{33}$$

where i indicates the position of each GoP. This can be converted into an unconstrained problem by considering the

RD cost J_v and Lagrange multiplier λ_v of the whole video sequence

$$J_v = \sum_{i=1}^{n} D_i(R_i) + \lambda_v \left(R_c - \sum_{i=1}^{n} R_i \right). \tag{34}$$

Differentiating J_v with respect to R_i and λ_v, the optimal solution of the constrained problem is given by

$$\begin{aligned} \forall i, \quad \frac{d \sum_{i=1}^{n} D_i(R_i)}{d R_i} - \lambda_v \frac{d \sum_{i=1}^{n} R_i}{R_i} &= 0 \\ R_c - \sum_{i=1}^{n} R_i &= 0 \end{aligned} \tag{35}$$

leading to [59]

$$\frac{dD_1}{dR_1} = \frac{dD_2}{dR_2} = \cdots = \frac{dD_n}{dR_n} = \lambda_v \tag{36}$$

with constraint

$$\sum_{i=1}^{n} R_i = R_c. \tag{37}$$

Acknowledgment

The authors would like to thank the Associate Editor and anonymous reviewers for their valuable comments that significantly helped improve the quality of this paper.

References

[1] 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.

[2] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[3] I.-K. Kim, J. Min, T. Lee, W.-J. Han, and J. Park, "Block partitioning structure in the HEVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1697–1706, Dec. 2012.

[4] H. Li, B. Li, and J. Xu, "Rate-distortion optimized reference picture management for High Efficiency Video Coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1844–1857, Dec. 2012.

[5] S. Ma, W. Gao, and Y. Lu, "Rate-distortion analysis for H.264/AVC video coding and its application to rate control," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 12, pp. 1533–1544, Dec. 2005.

[6] S. Ma, J. Si, and S. Wang, "A study on the rate distortion modeling for High Efficiency Video Coding," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep./Oct. 2012, pp. 181–184.

[7] MPEG. *TM5*, accessed on May 2016. [Online]. Available: http://www.mpeg.org/MPEG/MSSG/tm5

[8] J. Ribas-Corbera and S. Lei, *Rate Control for Low-Delay Video Communications*, document TMN8, ITU-T, Video Codec Test Model, ITU-T/SG15, Portland, OR, USA, Jun. 1997.

[9] T. Chiang and Y.-Q. Zhang, "A new rate control scheme using quadratic rate distortion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 1, pp. 246–250, Feb. 1997.

[10] H. Choi, J. Nam, J. Yoo, D. Sim, and I. Bajić, "Rate control based on unified RQ model for HEVC," in *Proc. JCTVC-H0213 ITU-T MPEG*, San Jose, CA, USA, 2012, pp. 1–13.

[11] S. Wang, S. Ma, S. Wang, D. Zhao, and W. Gao, "Rate-GOP based rate control for High Efficiency Video Coding," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 6, pp. 1101–1111, Dec. 2013.

[12] Z. He and S. K. Mitra, "A linear source model and a unified rate control algorithm for DCT video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 11, pp. 970–982, Nov. 2002.

[13] J. Si, S. Ma, S. Wang, and W. Gao, "Laplace distribution based CTU level rate control for HEVC," in *Proc. Vis. Commun. Image Process. (VCIP)*, Nov. 2013, pp. 1–6.

[14] B. Li, H. Li, L. Li, and J. Zhang, "λ domain rate control algorithm for High Efficiency Video Coding," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3841–3854, Sep. 2014.

[15] B. Li, H. Li, L. Li, and J. Zhang, "Rate control by R-lambda model for HEVC," in *Proc. JCTVC-K0103, JCTVC ISO/IEC ITU-T, 11th Meeting*, Shanghai, China, 2012.

[16] M. Rezaei, M. M. Hannuksela, and M. Gabbouj, "Semi-fuzzy rate controller for variable bit rate video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 5, pp. 633–645, May 2008.

[17] W.-N. Lie, C.-F. Chen, and T. C.-I. Lin, "Two-pass rate-distortion optimized rate control technique for H.264/AVC video," *Proc. SPIE*, vol. 5960, p. 596035, Jul. 2006.

[18] J. Sun, Y. Duan, J. Li, J. Liu, and Z. Guo, "Rate-distortion analysis of dead-zone plus uniform threshold scalar quantization and its application—Part II: Two-pass VBR coding for H.264/AVC," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 215–228, Jan. 2013.

[19] D. Zhang, K. N. Ngan, and Z. Chen, "A two-pass rate control algorithm for H.264/AVC high definition video coding," *Signal Process., Image Commun.*, vol. 24, no. 5, pp. 357–367, 2009.

[20] R. Kamran, M. Rezaei, and D. Fani, "A frame level fuzzy video rate controller for variable bit rate applications of HEVC," *J. Intell. Fuzzy Syst.*, vol. 30, no. 3, pp. 1367–1375, 2016.

[21] 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.

[22] 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.

[23] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[24] Z. Wang and A. C. Bovik, "Modern image quality assessment," *Synth. Lect. Image, Video, Multimedia Process.*, vol. 2, no. 1, pp. 1–156, 2006.

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

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

[27] A. C. Brooks, X. Zhao, and T. N. Pappas, "Structural similarity quality metrics in a coding context: Exploring the space of realistic distortions," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 1261–1273, Aug. 2008.

[28] B. H. K. Aswathappa and K. R. Rao, "Rate-distortion optimization using structural information in H.264 strictly Intra-frame encoder," in *Proc. 42nd Southeastern Symp. Syst. Theory*, 2010, pp. 367–370.

[29] Z.-Y. Mai, C.-L. Yang, L.-M. Po, and S.-L. Xie, "A new rate-distortion optimization using structural information in H.264 I-frame encoder," in *Proc. ACIVS*, 2005, pp. 435–441.

[30] Z.-Y. Mai, C.-L. Yang, and S.-L. Xie, "Improved best prediction mode(s) selection methods based on structural similarity in H.264 I-frame encoder," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Oct. 2005, pp. 2673–2678.

[31] Z.-Y. Mai, C.-L. Yang, K.-Z. Kuang, and L.-M. Po, "A novel motion estimation method based on structural similarity for H.264 inter prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2. May 2006, pp. II913–II916.

[32] C.-L. Yang, R.-K. Leung, L.-M. Po, and Z.-Y. Mai, "An SSIM-optimal H.264/AVC inter frame encoder," in *Proc. IEEE Int. Conf. Intell. Comput. Intell. Syst.*, vol. 4. Nov. 2009, pp. 291–295.

[33] C.-L. Yang, H.-X. Wang, and L.-M. Po, "Improved inter prediction based on structural similarity in H.264," in *Proc. IEEE Int. Conf. Signal Process. Commun.*, vol. 2. Nov. 2007, pp. 340–343.

[34] Y.-H. Huang, T.-S. Ou, P.-Y. Su, and C. H. Chen, "Perceptual rate-distortion optimization using structural similarity index as quality metric," *Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1614–1624, Nov. 2010.

[35] H. H. Chen, Y. H. Huang, P. Y. Su, and T. S. Ou, "Improving video coding quality by perceptual rate-distortion optimization," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2010, pp. 1287–1292.

[36] P. Su, Y. Huang, T. Ou, and H. Chen, "Predictive Lagrange multiplier selection for perceptual-based rate-distortion optimization," in *Proc. 5th Int. Workshop Video Process. Qual. Metrics Consum. Electron.*, Jan. 2010.

[37] Y. H. Huang, T. S. Ou, and H. H. Chen, “Perceptual-based coding mode decision,” in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2010, pp. 393–396.

[38] T.-S. Ou, Y.-H. Huang, and H. H. Chen, “A perceptual-based approach to bit allocation for H.264 encoder,” *Proc. SPIE*, vol. 7744, p. 77441B, Jul. 2010.

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

[40] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, “Rate-SSIM optimization for video coding,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 833–836.

[41] D. Brunet, E. R. Vrscay, and Z. Wang, “On the mathematical properties of the structural similarity index,” *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1488–1499, Apr. 2012.

[42] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, “Perceptual video coding based on SSIM-inspired divisive normalization,” *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1418–1429, Apr. 2013.

[43] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, “SSIM-inspired divisive normalization for perceptual video coding,” in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 1657–1660.

[44] A. Rehman and Z. Wang, “SSIM-inspired perceptual video coding for HEVC,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2012, pp. 497–502.

[45] S. Wang, K. Gu, K. Zeng, Z. Wang, and W. Lin, “Objective quality assessment and perceptual compression of screen content images,” *IEEE Comput. Graph. Appl.*, to be published.

[46] J. M. Foley, “Human luminance pattern-vision mechanisms: Masking experiments require a new model,” *J. Opt. Soc. Amer.*, vol. 11, no. 6, pp. 1710–1719, 1994.

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

[48] D. J. Heeger, “Normalization of cell responses in cat striate cortex,” *Vis. Neurosci.*, vol. 9, no. 2, pp. 181–198, 1992.

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

[50] Q. Li and Z. Wang, “Reduced-reference image quality assessment using divisive normalization-based image representation,” *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 202–211, Apr. 2009.

[51] S. S. Channappayya, A. C. Bovik, and R. W. Heath, Jr., “Rate bounds on SSIM index of quantized images,” *IEEE Trans. Image Process.*, vol. 17, no. 9, pp. 1624–1639, Sep. 2008.

[52] R. Sjöberg and J. Samuelsson, *Absolute Signaling of Reference Pictures*, document JCTVC-F493, Joint Collaborative Team Video Coding (JCT-VC) ITU-T SG16 WP3 ISO/IEC JTC1/SC29/WG11 6th Meeting, Turin, Italy, 2011.

[53] J. Yang, X. Fang, and H. Xiong, “A joint rate control scheme for H.264 encoding of multiple video sequences,” *IEEE Trans. Consum. Electron.*, vol. 51, no. 2, pp. 617–623, May 2005.

[54] M. Tagliasacchi, G. Valenzise, and S. Tubaro, “Minimum variance optimal rate allocation for multiplexed H.264/AVC bitstreams,” *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1129–1143, Jul. 2008.

[55] Z. He and D. O. Wu, “Linear rate control and optimum statistical multiplexing for H.264 video broadcast,” *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1237–1249, Nov. 2008.

[56] M. Tiwari, T. Groves, and P. C. Cosman, “Competitive equilibrium bitrate allocation for multiple video streams,” *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 1009–1021, Apr. 2010.

[57] K. Ramchandran, A. Ortega, and M. Vetterli, “Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders,” *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 533–545, Sep. 1994.

[58] C. Pang, O. C. Au, F. Zou, J. Dai, X. Zhang, and W. Dai, “An analytic framework for frame-level dependent bit allocation in hybrid video coding,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 6, pp. 990–1002, Jun. 2013.

[59] A. Ortega, “Optimization techniques for adaptive quantization of image and video under delay constraints,” Ph.D. dissertation, Graduate School Arts Sci., Columbia Univ., New York, NY, USA, 1994.

[60] X. Li, N. Oertel, A. Hutter, and A. Kaup, “Laplace distribution based Lagrangian rate distortion optimization for hybrid video coding,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, pp. 193–205, Feb. 2009.

[61] J. Ribas-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.

[62] L. Merritt and R. Vanam. (2006). *x264: A High Performance H.264/AVC Encoder*. [Online]. Available: [Online]. Available: http://www.uta.edu/faculty/krrao/dip/Courses/EE5359/overview_x264_v8_5[1].pdf

[63] *HM 13.0 Software*. [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-13.0/

[64] G. Bjontegaard, *Improvements of the BD-PSNR model*, document ITU-T SG16 Q.6, VCEG-AI11, Berlin, Germany, Jul. 2008.

[65] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, vol. 2, pp. 1398–1402.

[66] A. Rehman, K. Zeng, and Z. Wang, “Display device-adapted video quality-of-experience assessment,” *Proc. SPIE*, vol. 9394, p. 939406, Mar. 2015.

[67] Z. Wang, *SSIMplus*, accessed on May 2016. [Online]. Available: https://ece.uwaterloo.ca/~z70wang/research/ssimplus/

Shiqi Wang (M’15) received the Ph.D. degree in computer application technology from Peking University, Beijing, China, in 2014.

He was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. He is currently with Nanyang Technological University, Singapore, as a Research Fellow. His research interests include video compression, image/video quality assessment, and image/video search and analysis.

Abdul Rehman received the Ph.D. degree in information and communication systems from University of Waterloo, Waterloo, ON, Canada.

He is currently the President of SSIMWave, Waterloo, a company, he co-founded in 2013, dedicated to delivering excellence in visual quality-of-experience (QoE). He leads the development of SSIMWave’s state-of-the-art video QoE measurement and optimization products geared toward the media, communication, and entertainment industry. His research interests include image and video processing, coding and quality assessment, and multimedia communications.

Kai Zeng received the B.E. and M.A.Sc. degrees in electrical engineering from Xidian University, Xi’an, China, in 2006 and 2009, respectively, and the Ph.D. degree in electrical and computer engineering from University of Waterloo, Waterloo, ON, Canada, in 2013.

He was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, from 2013 to 2015. He is currently the CTO of SSIMWave Inc., Waterloo. His research interests include computational video and image pattern analysis, multimedia communications, and image and video processing (coding, denoising, analysis, and representation), with an emphasis on image and video quality assessment and corresponding applications.

Jiheng Wang received the M.Math. degree in statistics and computing from University of Waterloo, Waterloo, ON, Canada, in 2011, where he is currently pursuing the Ph.D. degree in electrical and computer engineering.

From 2009 to 2010, he was a Research Assistant with the Department of Statistics and Actuarial Science, University of Waterloo. Since 2011, he has been a Research Assistant with the Department of Electrical and Computer Engineering, University of Waterloo. His research interests include 3D image and video quality assessment, perceptual 2D and 3D video coding, statistical learning, and dimensionality reduction.

Zhou Wang (F'14) received the Ph.D. degree in electrical and computer engineering from The University of Texas at Austin, Austin, TX, USA, in 2001.

He is currently a Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. He has authored more than 100 publications in his research fields with more than 30 000 citations (Google Scholar). His research interests include image processing, coding, and quality assessment; computational vision and pattern analysis; multimedia communications; and biomedical signal processing.

Dr. Wang received the 2015 Primetime Engineering Emmy Award, the 2014 NSERC E.W.R. Steacie Memorial Fellowship Award, the 2013 IEEE Signal Processing Best Magazine Paper Award, the 2009 IEEE Signal Processing Society Best Paper Award, the 2009 Ontario Early Researcher Award, and the International Conference on Image Processing 2008 IBM Best Student Paper Award (as a Senior Author). He has served as a Senior Area Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING since 2015, and an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY since 2016 and *Pattern Recognition* since 2006. He served as an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2009 to 2014, and the IEEE SIGNAL PROCESSING LETTERS from 2006 to 2010, and a Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING from 2013 to 2014 and from 2007 to 2009, the *EURASIP Journal of Image and Video Processing* from 2009 to 2010, and *Signal, Image and Video Processing* from 2011 to 2013.