Signal Processing: *Image Communication* 29 (2014) 1046–1062

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Window-based rate control for video quality optimization with a novel INTER-dependent rate-distortion model

Yuan Li [a], Huizhu Jia [a,*], Chuang Zhu [a], Mingyuan Yang [b], Xiaodong Xie [a], Wen Gao [a]

[a] National Engineering Laboratory for Video Technology, Peking University, Beijing 100871, China
[b] Beijing BOYA-HUALU Technology Inc, Beijing 100080, China

ARTICLE INFO

Article history:
Received 24 March 2014
Received in revised form
8 August 2014
Accepted 16 September 2014
Available online 23 September 2014

Keywords:
Rate control
Rate–distortion (R–D) model
Video coding
Consistent video quality
Sliding window

ABSTRACT

Most model-based rate control schemes use independent rate-distortion (R–D) models at macroblock (MB) level to represent the relationship among bit rate, distortion and encoding complexity. However the correlations between frames (INTER-dependency) are not well considered for distortion, bit allocation and quantization parameter (QP) decision. In this paper, a novel INTER-dependent R–D model is proposed based on the theoretical analysis of the relationship between the predicted residual of one frame and the distortion of its reference frame. To achieve both bit rate accuracy and consistent video quality, a window-based rate control scheme with two sliding windows is introduced. One window is to group certain previously encoded frames and current frame to control the bit rate and buffer delay; the other is to group certain future encoding frames to optimize the fluctuation of video quality. Furthermore, the optimization of Lagrange multiplier is also discussed under the INTER-dependent situation. Experimental results demonstrate that the proposed window-based rate control scheme with INTER-dependent R–D model can achieve accurate target bit rate and improve PSNR performance, meanwhile the variation of PSNR is the smallest compared with other three benchmark algorithms. This one-pass rate control scheme is highly practical for the real-time video coding applications.

1. Introduction

Rate control is essential for the real applications of the modern video coding standards such as MPEG-2 [1], H.264/AVC [2] and AVS [3]. Various video coding applications introduce strict bit rate constraint to the bit stream due to the limited transmission bandwidth or storage size. Rate control scheme is responsible for achieving the bit budget by adjusting the quantization parameters (QPs) or trading the compressed video quality in a video encoder. Beside rate, other optimizations are also needed to be considered such as system latency, buffer occupation and smoothness in objective or subjective video quality.

To achieve the bit rate constraint, two essential steps are adopted in a typical rate control scheme, which are frame bit allocation and QP decision. The former is used to allocate bit quota among different video frames, which usually takes the buffer latency, video quality and coding complexity into consideration. The latter, after allocation of the frame bit budget, decides the QP of a frame or each MB of the frame to achieve the bit budget accurately. R–D model is widely applied in this step to represent the relationship among bit rate, distortion and QP.

Many rate control schemes towards these two problems are proposed and developed in the literature. For the frame

* Corresponding author.

http://dx.doi.org/10.1016/j.image.2014.09.004
0923-5965/

bit allocation, the state-of-the-art works can be classified into three categories. The first category allocates equal or nearly equal bits among different video frames. In [4], Ribas-Corbera and Lei used a nearly constant frame bit target to ach;ieve the low buffer delay, meanwhile avoiding the underflow of the buffer. He [5] et al. adopted the similar near-constant frame bit allocation and smoothed the rate variation by adjusting the distortion in a small range. These methods can maintain the small fluctuation of the encoder buffer. However, the source video complexity is not well considered in these schemes, which will cause the fluctuation of the video quality especially when high motion occurs or scene changes. The second category assumes that the video content is stationary among different group of pictures (GOP). Under this assumption, equal bits are allocated to each GOP. Within a GOP, the frame bit allocation scheme assigns a fixed weighting factor according to the frame type. The typical one of this category is TM5 [6] for MPEG-2, it also considers the fullness of the compressed bit buffer. The JM [7] for H.264/AVC uses the similar mechanism for frame bit allocation, meanwhile taking the hypothetical reference decoder (HRD) [8] buffer into consideration to further regulate the frame bit budget. Since the assumption of these methods is not always true, the fluctuation of the video quality is unavoidable. The third category is aimed to achieve smooth video quality. The basic idea of these methods is to allocate more bits to the high-complexity frames and less bits to the low-complexity frames. In [9], Xie and Zeng proposed a sequence-based bit allocation scheme by tracking the non-stationary characteristics in a video sequence. Xu [10] et al. proposed a window model about the picture quality and the buffer occupancy. By applying window-level bit allocation, the tradeoff between quality smoothness and buffer smoothness can be achieved. The "*forward*" rate control scheme, which means that allocating frame bits based on the characteristics of the current frame or future frames via certain pre-analysis, is widely used in these methods. Our work, towards the one-pass real-time encoding application with smooth video quality, also belongs to this category.

For QP decision, various R–D models are proposed in the literature. Some of these R–D models [4,5,11–14] assume that the video coding units are independent with each other. Under this assumption, Chiang and Zhang [11] proposed a quadratic R–D model to calculate the target bit rate for each frame, which was adopted in both MPEG-4 and H.264/AVC. In [4], a MB-level R–D model was used to choose the QP, meanwhile the Lagrange optimization was introduced to minimize distortion. The rate control scheme was adopted by H.263. He et al. [5] proposed a linear ρ-domain R–D model, which used the percentage of zero coefficients after quantization to approximate the bit rate. To tackle the inherent dilemma between rate control and R–D optimization (RDO) in H.264/AVC, Ma et al. [12] used the true quantization step size to establish the R–D model and proposed a rate control scheme with partial two-pass process at MB level [13] proposed an enhanced R–D model, which modeled the source bits as the function of the quantization step size and the complexity of coded 4×4 blocks. In [14], a linear model was formulated to describe the relationship between the total amount of bits for both texture and non-texture information and the QP. There are also a number of investigations for R–D models in the next generation video coding standard—High Efficiency Video Coding (HEVC). Based on the quadratic R–D model in [11], Choi et al. [27] proposed a pixel-based unified rate-quantization (URQ) model, which employed a mean of absolute difference (MAD) factor to predict the texture complexity. This rate control algorithm was adopted in the HEVC test model reference software version 6.0 (HM6.0) [28]. In [29], a rate control scheme using a linear R-λ model was proposed, which showed smaller bit rate errors than the URQ model and was adopted in HM10.0 [30]. Considering the quadtree coding structure in HEVC, Seo et al. [31] proposed a rate control scheme with a new R–D model based on the Laplacian function to minimize the fluctuation of video quality. In [32], a frame-level rate control scheme based on texture and nontexture rate models was proposed, which considered the different statistical characteristics of transform coefficients depending on the depth levels of coding units (CUs). A better R–D performance could also be achieved compared to the previous methods. However, in the more general case, the coding units may not be coded independently, especially when the INTER-dependency is taken into consideration. Here INTER-dependency means that both distortion and bit rate of the current encoding inter frame (either P or B frame) are highly correlated with the distortion of its reference frame, because of the prediction process between the inter frame and its reference frame. To tackle the rate control problem with INTER-dependent characteristics, Ramchandran et al. [15] provided a trellis-based solution for an arbitrary set of QPs for each coding unit. The computational complexity grew exponentially with the increase of dependent frame numbers. In [16], Lin and Ortega used interpolation to establish the approximated R–D curves. The spline interpolation and piecewise linear interpolation were adopted for I frames and P frames respectively. Liu et al. [17] analyzed the dependent temporal-spatial bit allocation problem and proposed two iteration algorithms to reduce the computational complexity. In scalable video coding, Liu and Kuo [18] proposed a GOP-based distortion model for different temporal layers according to the dependency between the base layer and the enhancement layer. The algorithms of [16–18] need to encode the source video several times, which are not suitable for real-time applications.

The above-mentioned R–D models are established by heuristic analyses and statistical examinations. However, the theoretical INTER-dependent R–D model among different coding units needs to be further developed. In this paper, we first analyze the INTER-dependent problem and establish the relationship between the residual of one frame and the distortion of its reference frame. Based on this analysis, we derive the INTER-dependent distortion-quantization (D–Q) model and rate-quantization (R–Q) model via the study of the spatial-domain residual and the transform-domain residual. Then a window-based rate control scheme is proposed with the complexity-based frame bit allocation and video quality optimization. Furthermore, the optimization of Lagrange multiplier is also discussed under the INTER-dependent situation. Experimental results demonstrate that the proposed window-based rate control scheme with INTER-dep-

endent R–D model can achieve accurate target bit rate and improve PSNR performance, meanwhile the variation of PSNR is the smallest compared with other three benchmark algorithms. This one-pass rate control scheme is highly practical for the real-time video coding application.

The rest of this paper is organized as follows. The INTER-dependency problem is analyzed in Section 2. Based on the analysis, a novel INTER-dependent R–D model is derived in Section 3. Section 4 represents the window-based rate control scheme with complexity-based frame bit allocation and video quality optimization. The experimental results and discussions are shown in Section 5 and Section 6 and we give the conclusion in Section 7.

2. INTER-dependency problem

Inter-frame prediction used in video coding increases the compression performance dramatically, meanwhile causing the dependency problem in R–D based rate control. Both distortion and bit rate of an inter frame (either P or B frame) will be affected by the QP variation of its reference frame (either I or P frame). To demonstrate this dependency problem, we take two frames as an example, in which the second dependent frame references the first independent frame. In this situation, the rate control issue can be formulated as minimizing the total distortion under the bit rate constraint. With the traditional coding-unit-independent assumption, the formulation is

$$\min_{Q_1,Q_2}(D_1(Q_1)+D_2(Q_2))$$
$$\text{such that } (R_1(Q_1)+R_2(Q_2)) \le R_{budget} \tag{1}$$

where Q_1, $D_1(Q_1)$, $R_1(Q_1)$ are the QP, distortion and bit rate of the first frame which is the reference frame, Q_2, $D_2(Q_2)$, $R_2(Q_2)$ are the corresponding parameters of the second frame which is predicted from the first frame. Actually, in the encoding process, the distortion and bit rate of the second frame have strong dependency with its reference frame since adopting different QPs for the first frame will generate different reconstructed frames, which act as the references for the second frame. Considering this inter-frame dependency, the rate control problem becomes more complicated. The formulation (1) can thus be rewritten as

$$\min_{Q_1,Q_2}(D_1(Q_1)+D_2(Q_1,Q_2))$$
$$\text{such that } (R_1(Q_1)+R_2(Q_1,Q_2)) \le R_{budget} \tag{2}$$

where $D_2(Q_1,Q_2)$ and $R_2(Q_1,Q_2)$ represent that the distortion and bit rate of the second frame are dependent on both the QP of the first frame as well as the second frame.

To address this dependency problem, a trellis-based solution is used in [15]. However, the real bit rate and distortion need to be obtained first. R–D model based solutions are proposed in [16,18], which need multi-pass encoding to derive the R–D models. These solutions are not suitable for real-time video coding applications since the "*forward*" bit allocation is usually needed in rate control scheme. To further reduce the computational complexity, we aim to establish the R–D model using the spatial-domain information which can be easily obtained in the pre-analysis process before the actual encoding is performed [5,10].

In the inter-frame coding process, the residual pixels of the second frame, which directly contribute to the bit rate and the distortion, are generated by the subtraction of the original pixels of the second frame, represented as $org_{2,(i,j)}$, and its reference pixels from the first frame. The reference pixels in the first frame are the reconstructed pixels which contain the distortion due to the quantization. Since the motion search is adopted in the inter-frame coding process which generates the motion vector (MV) for the $org_{2,(i,j)}$, represented by $MV(i,j)=(x_i,y_j)$, the reference pixel in the first frame corresponding to $org_{2,(i,j)}$ should be represented as $rec_{1,(i+x_i,j+y_j)}$. We use *MAD* at frame level to represent the residual information in spatial domain as

$$MAD_2 = \frac{1}{M\times N}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left|org_{2,(i,j)} - rec_{1,(i+x_i,j+y_j)}\right| \tag{3}$$

where MAD_2 is the real *MAD* of the second frame generated in the encoding process, M, N are the frame width and height respectively. The $rec_{1,(i+xi,j+yj)}$ can be calculated by the subtraction of the original pixels and the error which is the distortion represented by $err_{1,(i+x_i,j+y_j)}$.

$$rec_{1,(i+x_i,j+y_j)} = org_{1,(i+x_i,j+y_j)} - err_{1,(i+x_i,j+y_j)} \tag{4}$$

From (3) and (4), we can get that

$$\begin{aligned} MAD_2 &= \frac{1}{M\times N}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left|org_{2,(i,j)} - \left(org_{1,(i+x_i,j+y_j)} - err_{1,(i+x_i,j+y_j)}\right)\right| \\ &\approx \frac{1}{M\times N}\left(\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left|org_{2,(i,j)} - org_{1,(i+x_i,j+y_j)}\right| + \sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left|err_{1,(i+x_i,j+y_j)}\right| - \beta\right) = MAD\_O_2 \\ &\quad + \frac{1}{M\times N}\left(\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left|err_{1,(i+x_i,j+y_j)}\right| - \beta\right) \end{aligned} \tag{5}$$

where MAD_O_2 represents the MAD between the original pixels of the second frame and the first frame, which can be easily obtained by the pre-analysis. According to the approximation that we used in (5), β is positive and should be subtracted by the sum of MAD_O_2 and err_1. The second term of (5) is related to the distortion which is usually represented by the mean squared error (MSE) as

$$\begin{aligned} MSE_1 &= \frac{1}{M\times N}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(org_{1,(i,j)} - rec_{1,(i,j)})^2 \\ &= \frac{1}{M\times N}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} err^2_{1,(i,j)} \end{aligned} \tag{6}$$

Since the second item of (5), which only contains the distortion of the referenced pixels but not the whole frame (it should be noticed that the sum of $err_{1,(i+x_i,j+y_j)}$ for iterators i and j could not traverse the pixels of whole frame because of the impact by the $MV(i,j)$), is partial to (6), we can use (6) as the approximation of it. That is

$$\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left|err_{1,(i+x_i,j+y_j)}\right| \approx \alpha \times \sqrt{M\times N\times MSE_1} \tag{7}$$

where α is the parameter that has a direct relationship with the MAD_ref_1/MAD_1, MAD_ref_1 represents the MAD of

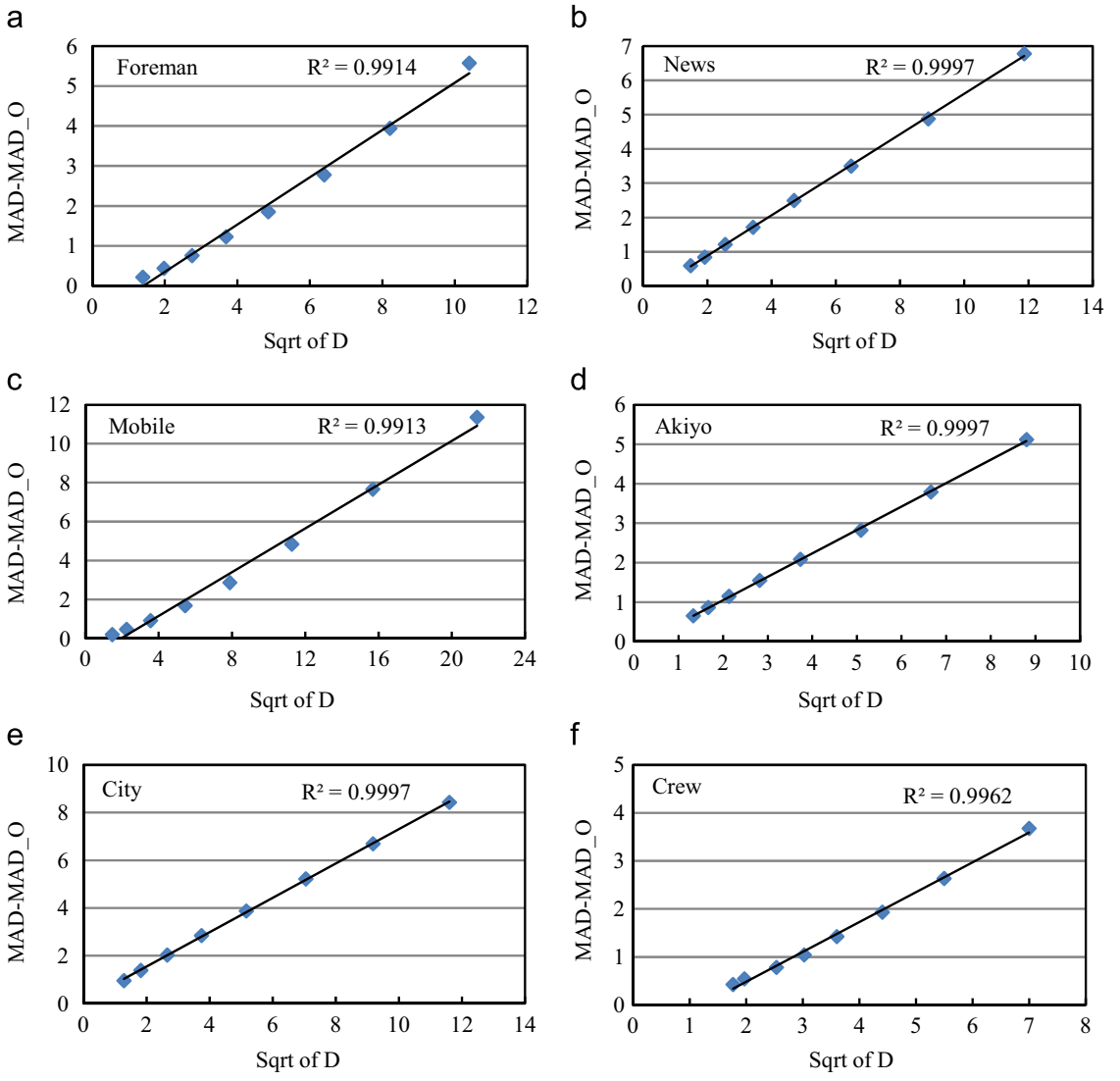

Fig. 1. Relationship between the distortion of the reference frame and the dependent residual. (a) "*Foreman*", CIF format, second frame, QP from 18 to 46; (b) "*News*", CIF format, third frame, QP from 18 to 46; (c) "*Mobile*", CIF format, sixth frame, QP from 18 to 46; (d) "*Akiyo*", CIF format, fifth frame, QP from 18 to 46; (e) "*City*", 720P format, third frame, QP from 18 to 46; and (f) "*Crew*", 720P format, second frame, QP from 18 to 46.

the referenced pixels in the first frame. Then, from (5) and (7), we can get that

$$MAD_2 = MAD\_O_2 + k \times \sqrt{D_1} + t \tag{8}$$

where k and t are model parameters. We test some video sequences with variable QPs to verify this relationship. The experimental results on the CIF and 720P sequences are shown in Fig. 1. The correlation between the square root of distortion of the reference frame and the subtraction of *MAD* and *MAD_O* is larger than 0.99, which shows a good linear relation between the two.

From (8), the relationship between the residual of the second frame in spatial domain and the distortion of the first frame is established. We can get the real complexity information (MAD_2) from the pre-analysis (MAD_O_2) plus the distortion of the reference frame (D_1) without the actual encoding. Combining the D–Q model of the complexity and the distortion and the R–Q model of the complexity and the bit rate which will be described in the following subsection, the "*forward*" bit allocation (which means assign target frame bits for current frame and future frames before encoding) and video quality optimization can be easily achieved according only to the spatial-domain residual information.

3. INTER-dependent rate-distortion model

In this section, we establish the so-called INTER-dependent R–D model, which is based on the theoretical analysis on the relationship between the spatial-domain residual and the transform-domain residual.

3.1. INTER-dependent D–Q model

Setting up the INTER-dependent D–Q model contains two steps: (1) setting up the relationship between the distortion, quantization step size and the true MAD. (2) combining it into (8) to get the INTER-dependent D–Q model.

Modern hybrid video coding standards adopt discrete cosine/sine transform (DCT/DST) to convert the predicted residual block, represented by S, to a transformed matrix, represented by X, then use the quantization and entropy coding to achieve compression. The DCT process can be expressed as the following [19]:

$$X = ASA^T \tag{9}$$

where T denotes transposition and for the case of DCT,

$$A(k,n) = \begin{cases} \frac{1}{\sqrt{N}}, k = 0, & 0 \le n \le N-1 \\ \sqrt{\frac{2}{N}} \cos \frac{\pi(2n+1)k}{2N}, & 1 \le k \le N-1, \ \ 0 \le n \le N-1 \end{cases} \tag{10}$$

where N is 4 in H.264/AVC.

Since the pixel values of S which is the input of the DCT can be approximated by a Laplacian distribution with a zero mean and a separable covariance $r(m,n) = \sigma_S^2 \rho^{|m|} \rho^{|n|}$ [20], the variance of the (u,v)th DCT coefficient $\sigma_X^2(u,v)$ can be represented as [21]

$$\sigma_X^2(u,v) = \sigma_S^2 [ARA^T]_{u,u} [ARA^T]_{v,v} \tag{11}$$

where

$$R = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} \tag{12}$$

ρ is the correlation coefficient and $[\bullet]_{u,u}$ denotes the (u,u)th component of the matrix. With $\rho = 0.6$ as a typical value [20], we can get that

$$\sigma_X^2(u,v) = \sigma_S^2 C(u,v) \tag{13}$$

where $C(u,v)$ is a matrix with constants. With the Laplacian distribution and zero mean, the variance of S can be approximated by $\sigma_S = \sqrt{2} MAD$ [20]. This will lead to $\sigma_X(u,v) = \sqrt{2C(u,v)} MAD$. Assuming that (u,v)th transformed coefficient $X(u,v)$ is Laplacian distributed as [22]

$$f_{X(u,v)}(x) = \frac{\lambda(u,v)}{2} e^{-\lambda(u,v)|x|} \tag{14}$$

where f denotes the probability density function (PDF) and

$$\lambda(u,v) = \frac{\sqrt{2}}{\sigma_{X(u,v)}} = \frac{1}{\sqrt{C(u,v)} MAD} \tag{15}$$

With PDF in (14), we can calculate the distortion by summing up the distortion at each quantization interval as [23]

$$D(u,v) = \int_{-(Q-\gamma Q)}^{Q-\gamma Q} x^2 f_{X(u,v)}(x) dx + 2 \sum_{n=1}^{\infty} \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} \times (x - nQ)^2 f_{X(u,v)}(x) dx \tag{16}$$

where Q is the quantization step size, γQ denotes the rounding offset and γ is between (0,1), which is 1/6 for H.264/AVC [2]. Substituting (14) into (16), we can get [23]

$$D(u,v) = \frac{\lambda Q e^{\gamma \lambda Q}(2 + \lambda Q - 2\gamma \lambda Q) + 2(1 - e^{\lambda Q})}{\lambda^2 (1 - e^{\lambda Q})} \tag{17}$$

where λ represents the $\lambda(u,v)$ in (15).

With (15) and (17), the relationship between distortion, quantization step size and MAD is established. However, this model is too complicated. A simpler and approximate model needs to be developed. Our approximation is based on two useful observations. First, we focus on the second item in the numerator of (17) and the denominator of (17), where $(1 - e^{\lambda Q})$ can be removed and we can get that $D(u,v)$ has a relationship with $2/\lambda^2$. From (15) we already obtain that λ is inversely proportional to MAD. Thus D has a directly relationship with MAD^2. Second, it has been stated that the distortion has an approximate exponential relation with QP [12], which is $D = MSE = (255^2)/10^{((lQP+b)/10)}$. With the relationship between quantization step size and QP in H.264/AVC [2], which is $Q = 2^{(QP-4)/6}$, it can be derived that D has a linear relation with Q. Based on these observations, we propose a simple yet accurate distortion model as

$$D = \alpha(MAD^2 + Q) + \beta \tag{18}$$

where α and β are model parameters. Substituting (8) into (18), we can further get the INTER-dependent D–Q model. Here, an approximate formulation which is easy for applying is given as

$$D_2 = a(Q_2 + MAD\_O_2^2 + k^2 D_1) + b \tag{19}$$

where a and b are model parameters, k is the same as in (8).

To verify the accuracy of the INTER-dependent D–Q model (19), we tested some video sequences with variable resolutions and QPs. The model accuracy is represented as

$$\text{Accuracy} = \left(1 - \frac{|\text{Estimated value} - \text{Actual } E|}{\text{Actual } E}\right) \times 100\% \tag{20}$$

Table 1
Verification of the INTER-dependent D–Q model.

Format	Sequence	QP	Est. D	Act. D	Accuracy (%)
CIF	Foreman	26	8.55	8.48	99.15
		34	26.22	24.32	92.22
	News	26	6.88	6.54	94.66
		34	24.74	22.09	88.04
	Mobile	30	32.69	29.68	89.85
		38	140.81	126.96	89.09
	Akiyo	30	8.52	7.90	92.17
		38	27.64	25.98	93.62
720P	City	26	8.17	7.06	84.24
		34	28.65	26.64	92.44
	Crew	26	5.36	6.43	83.38
		34	12.63	12.98	97.31
	Night	30	16.49	14.61	87.16
		38	51.10	48.08	93.71
	Harbour	30	18.26	16.33	88.20
		38	62.69	58.31	92.48
Average			–	–	**91.11**

Est. D: Estimated distortion.
Act. D: Actual distortion.

Table 2
Accuracy comparison of different D–Q models.

Format	Sequence	QP	ρ-domain (%)	Cauchy-based (%)	Proposed (%) (18)
CIF	Foreman	26	94.16	97.82	98.43
		34	92.84	94.21	93.94
	News	26	89.65	92.46	91.63
		34	86.47	90.15	87.41
	Mobile	30	82.58	85.31	86.46
		38	89.67	92.87	94.74
	Akiyo	30	88.21	89.64	90.45
		38	90.16	92.69	93.86
720P	City	26	87.49	88.78	88.85
		34	91.63	94.65	95.57
	Crew	26	86.79	87.19	89.20
		34	92.81	95.87	96.26
	Night	30	89.47	96.81	97.36
		38	90.59	97.49	98.28
	Harbour	30	91.53	96.97	97.23
		38	92.74	97.38	98.54
Average			**89.80**	**93.14**	**93.64**

Table 1 shows the detailed results. The average accuracy of the INTER-dependent D–Q model is 91.11%.

We also make the comparison of model accuracy with other existing D–Q models, which are ρ-domain D–Q model in [35] as $D(\rho)=\sigma^2 e^{-\omega(1-\rho)}$, where ω is parameter and σ^2 is the variance of transformed coefficients, and the Cauchy-density-based D–Q model in [24] as $D(Q)=mQ^{\delta}$, where m and δ are model parameters depend on the picture content. Note that these two D–Q models do not consider any dependency between inter frames, thus we use (18) which is also for the independent frames to do the comparison. The model accuracy is also represented by (20). The experimental results are shown in Table 2. Both Cauchy-based D–Q model and our proposed model have better accuracy than ρ-domain D–Q model. Besides, the computational complexity of our D–Q model is lower than Cauchy-based D–Q model because of the linear relationship.

3.2. INTER-dependent R–Q model

The relationship between bit rate and quantization step size under the independent coding assumption has been studied in the literature. By assuming that the transformed coefficients are Laplacian [22] or Cauchy [24] distributed, the bit rate can be derived from calculating the entropy of the quantized DCT coefficients. However, these R–Q models are complicated and not suitable for rate control applications. To reduce the computational complexity, many R–Q models based on the relationship between bit rate and coding complexities are proposed for rate control scheme [4,11,12], where the coding complexity is usually represented by MAD or sum of absolute difference (SAD). For the low computational complexity purpose, our investigation of the INTER-dependent R–Q model is based on the linear relationship between bit rate, SAD and quantization step size, which is widely used in the state-of-the-art works [12,14,33,34] as

$$R=a_1\frac{SAD}{Q}+b_1 \tag{21}$$

where SAD equals to $M\times N\times MAD$, a_1 and b_1 are the model parameters. The SAD in (21) is obtained in the actual frame coding process. By substituting (8) into (21), we can get the INTER-dependent R–Q model directly as

$$R_2=a_1\frac{SAD\_O_2+k\sqrt{D_1}+t}{Q_2}+b_1 \tag{22}$$

where SAD_O is the SAD between original frames similar to MAD_O. However, the complex term of $\sqrt{D}$ will also be introduced. To further simplify the R–Q model, we tested some sequences to statistically analyze (21) and the following R–Q model.

$$R=a_2\frac{SAD\_O}{Q}+b_2 \tag{23}$$

where a_2 and b_2 are model parameters. Fig. 2. shows the R–Q curves of (21) and (23) for different test sequences. From the figure, an approximate linear relationship is observed between R and SAD_O/Q in (23), similar to that in (21). With the R–Q model (23), the bit rate can be estimated only by the SAD_O and quantization step size, while the distortion effect of the reference frame can be neglected. In other words, the complex INTER-dependent R–Q issue is converted into a simple independent issue.

The implication contains two aspects. First, in high bit rate situation (QP is small), the distortion of the reference frame is negligible so that $\sqrt{D_1}$ in (22) can be removed. (22) degenerates to (23) which is still a linear relationship. Second, in low bit rate situation (QP is large), $\sqrt{D_1}$ in (22) cannot be neglected. From (18), D_1 has a linear relationship with Q_1 so that $\sqrt{Q_1}$ is introduced into the numerator of (22). Assuming $Q_1\approx Q_2$ for consistent video quality, we can get that the increase speed of numerator in (22) is slower than that of denominator because of the $\sqrt{Q}$ in the

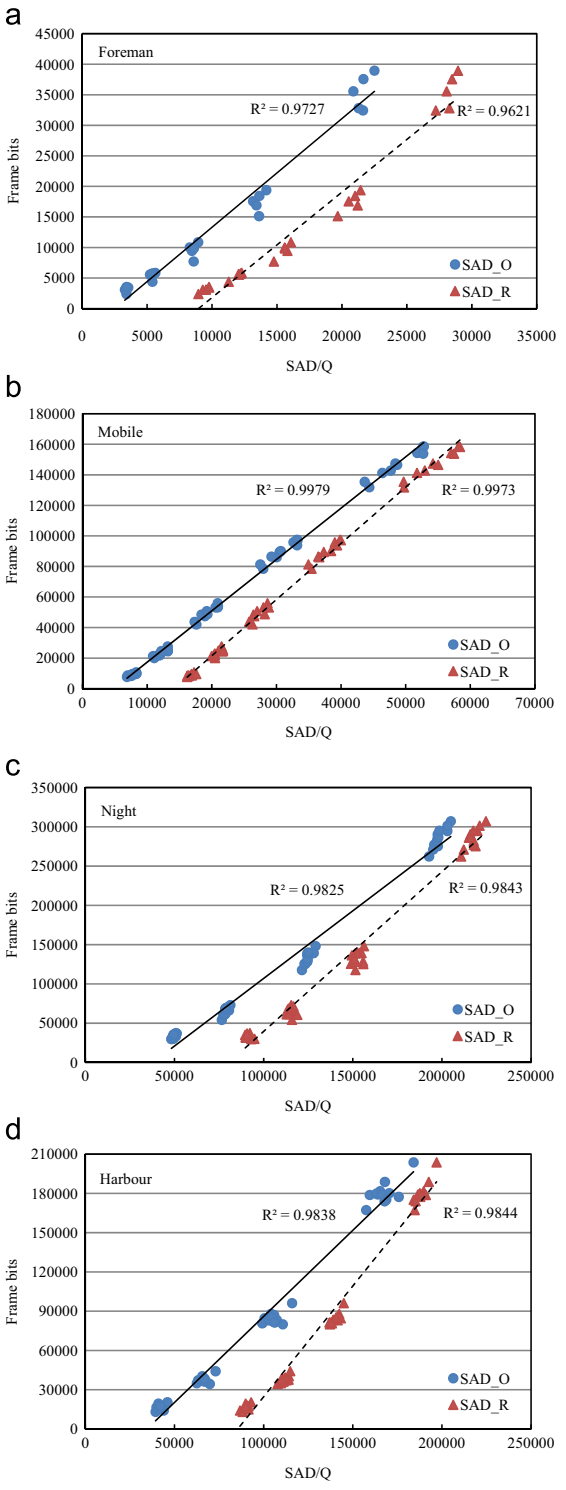

Fig. 2. *R–Q* curves of (21) and (23) from different sequences. (a) "*Foreman*", CIF format; (b) "*Mobile*", CIF format; (c) "*Night*", 720P format; and (d) "*Harbour*", 720P format.

numerator and Q in the denominator. However, in low bit rate, the header bits occupy a significant portion of the total bits, and the percentage of header bits increases as the Q becomes larger [13]. This will compensate for the slow increase speed of $\sqrt{Q}$ and the linear relationship will also be held as in (23).

The accuracy of INTER-dependent R–Q model is also verified using (20) and the results are listed in Table 3. The average accuracy of the INTER-dependent R–Q model is 91.45%.

The comparison of model accuracy between different R–Q models is also made in our work. The selected anchors include ρ-domain R–Q model in [5] as $R(\rho)=\theta(1-\rho)$, where θ is a model parameter and ρ is the percentage of zero coefficients among quantized transformed residuals, and the quadratic R–Q model used in H.264/AVC as $R(Q)=c_1MAD/Q+c_2MAD/Q^2$, where c_1 and c_2 are model parameters. Considering the independent assumption used in these anchor R–Q models, we use (21) to make the comparison. The model accuracy is calculated by (20). Table 4 shows the experimental results. It can be observed that the quadratic R–Q model has a better accuracy than linear R–Q models (both ρ-domain model and model (21)). However, the quadratic model has higher computational complexity when used for bit allocation. If certain frames are grouped to allocate bits for each one, solving the summation of linear model will be much easier than the summation of quadratic model. Thus, we adopt the linear R–Q model for the balance of accuracy and computational complexity.

4. Window-based rate control scheme

In this section, the proposed window-based rate control with INTER-dependent R–D model is introduced, which contains window-level bit rate constraint, complexity-based frame bit allocation with video quality optimization and QP decision.

4.1. Window-level bit rate constraint

In typical constant-bit-rate (CBR) rate control schemes, the first step is to allocate bits for a group of pictures to satisfy the bit rate constraint. GOP-based method is widely used which allocates equal bits for each GOP, such as TM5 for MPEG2 and JM for H.264/AVC. In Xu's work [10], as an improvement, a "*jumping*" window is proposed to allocate bits for a group of adjacent frames within the window, which means that adjacent windows are all independent and do not have any overlap. Both of these methods are under the assumption that the video content characteristics are stationary in different GOPs/windows, which is usually untrue in real life. The fluctuation of the video quality will occur when the bit budgets are not well allocated for difficult scene/frames and easy scene/frames among different GOPs/windows (in addition, the two schemes cannot guarantee a CBR when random access is observed). To avoid this situation, a sliding window, so called window-R, is proposed to allocate bits for several adjacent frames. Window-R, with the size of L, consists of the $L-1$ previously encoded frames plus the current frame, which means that the current frame is the last frame in the window. Let W_R be the total bits of window-R, then W_R can be obtained by (24).

$$W_R = L\frac{R_C}{F} \tag{24}$$

Table 3
Verification of the INTER-dependent R–Q model.

Format	Sequence	QP	Frame no.	Est. R	Act. R	Accuracy (%)
CIF	Foreman	26	10	18,946	20,352	93.09
		34	25	4536	5088	89.15
	News	26	34	1779	1568	86.54
		34	56	2479	2704	91.68
	Mobile	30	17	44,368	42,144	94.72
		38	38	9367	8920	94.99
	Akiyo	30	47	1364	1536	88.80
		38	68	924	768	79.69
720P	City	26	5	18,6542	177,120	94.68
		34	18	20,875	22,904	91.14
	Crew	26	13	227,631	249,304	91.31
		34	29	28,974	27,232	93.60
	Night	30	22	102,768	105,880	97.06
		38	36	31,567	35,280	89.48
	Harbour	30	37	161,687	171,080	94.51
		38	56	44,585	41,592	92.80
Average			**–**	–	–	**91.45**

Est. R: Estimated rate.
Act. R: Actual rate.

Table 4
Accuracy comparison of different R–Q models.

Format	Sequence	QP	ρ-domain (%)	Quadratic in JM (%)	R–Q model (%) (21)
CIF	Foreman	26	90.67	92.14	89.56
		34	91.23	91.82	88.11
	News	26	89.37	86.35	85.76
		34	91.64	88.67	86.32
	Mobile	30	85.74	90.68	88.72
		38	87.35	91.34	92.48
	Akiyo	30	89.36	90.35	89.61
		38	88.71	91.17	91.35
720P	City	26	89.37	87.02	85.67
		34	90.82	95.87	89.73
	Crew	26	87.92	92.07	90.64
		34	90.87	92.67	92.75
	Night	30	89.76	94.43	90.22
		38	88.64	93.68	91.36
	Harbour	30	92.01	92.64	88.71
		38	90.73	91.21	90.65
Average			**89.64**	**91.38**	**89.48**

where R_C denotes the bit rate of CBR and F denotes the frame rate. The bit budget for the current frame, denoted as R_T, is calculated as follows

$$R_T = W_R - \sum_{i=0}^{L-2} R_{r,i} \tag{25}$$

where $R_{r,i}$ represents the real bits for previously encoded ith frame in window-R. This window is sliding frame by frame to allocate the bits for each frame with a fixed window size of L. By this mechanism, the bits for any consecutive L frames are restricted as W_R. It also can be demonstrated that this method reduces the delay of the buffer, which is between the encoder and the transmission channel [25]. Window-R itself will not introduce any extra delay.

4.2. Frame bit allocation

Frame bit allocation needs not only to consider encoding complexity of each frame, but also to consider the fluctuation of the video quality across the entire encoded stream. Maintaining the near consistent video quality can be formulated as (26), which aims to minimize the variance of frame

distortion for the whole sequence.

$$\min(\mathrm{var}(D)), \text{where } \mathrm{var}(D)=\frac{1}{n}\sum_{i=0}^{n-1}(D_i-\overline{D})^2 \tag{26}$$

where D_i is the distortion of ith frame in the sequence and $\overline{D}$ denotes the average distortion of the n frames in the sequence. However, the global video characteristic is not available in one-pass real-time encoding process. The sub-optimal solution is adopted which is to minimize the variance of distortion for all previously encoded frames in respect of the current frame as shown in (27),

$$var(D_C)=\frac{1}{N}\left(\sum_{i=0}^{N-2}(D_{p,i}-\overline{D_N})^2+(D_C-\overline{D_N})^2\right) \tag{27}$$

where D_C denotes the target distortion of current frame, $D_{p,i}$ $i=0,1,\ldots,N-2$ denotes the distortion of the previously encoded frames, $\overline{D_N}$ denotes the average distortion including all encoded frames and the current frame, which can be represented by (28),

$$\overline{D_N}=\frac{1}{N}\left(\sum_{i=0}^{N-2}D_{p,i}+D_C\right) \tag{28}$$

Then substituting (28) into (27), we can get

$$\mathrm{var}(D_C)=\frac{N-1}{N^2}D_C^2-\frac{2D_S}{N^2}D_C+\frac{N^2+N-1}{N^3}D_{SS}+\frac{1-2N}{N^3}D_S^2 \tag{29}$$

where

$$D_S=\sum_{i=0}^{N-2}D_{p,i},\quad D_{SS}=\sum_{i=0}^{N-2}D_{p,i}^2 \tag{30}$$

To minimize $\mathrm{var}(D_C)$, let

$$\frac{\partial \mathrm{var}(D_C)}{\partial D_C}=0 \tag{31}$$

then we get

$$D_C=\frac{D_S}{N-1}=\frac{1}{N-1}\sum_{i=0}^{N-2}D_{p,i}=\overline{D_{N-1}} \tag{32}$$

From (32), we can conclude that, to minimize the variance of distortion for all previously encoded frames and the current encoding frame, the distortion of current frame should be equal to the average distortion of the previously encoded frames.

Based on this conclusion about the fluctuation of video quality reduction, another window, so called window-D, consisting of consecutive M frames is introduced to allocate the frame bit. The first frame of window-D is the current frame, while other $M-1$ frames are future successive frames to be encoded. Pre-analysis is used to get the complexity (MAD_O at frame level after prediction) of each frame in window-D. For INTER frames, only 16×16 motion search is used, while for INTRA frames, only few types of prediction (such as horizontal, vertical and diagonal) are used. The computational complexity of the pre-analysis is much lower compared to the complete encoding process. Window-D is also sliding frame by frame, together with window-R.

After the characteristics of the M frames in window-D are obtained, we can use the INTER-dependent D–Q model to derive the relationship of quantization step size between these frames. From (32), it can be further derived that, to minimize the variance of the distortion including the M unknown distortions, the M unknown distortions should be equal to each other, which is

$$D_0=D_1=\cdots=D_{M-1} \tag{33}$$

where D_i denotes the distortion of ith frame in window-D and D_0 is for the current frame. Substituting (33) into the D–Q model (19), we can get the relationship between the quantization step size of ith frame Q_i and D_0 as

$$Q_i=\left(\frac{1}{a}-k^2\right)D_0-MAD\_O_i^2-\frac{b}{a} \tag{34}$$

It should be noticed that D_0 can always be obtained by (18) in respect of Q_0. If the window is the first window of the video sequence, we can use $MAD_0\approx MAD\_O_0$ to calculate D_0. Otherwise, MAD_0 can be obtained by (8) since its reference frame has already been encoded. Thus, we can get that

$$D_0=\alpha(MAD_0^2+Q_0)+\beta \tag{35}$$

Substituting (35) into (34), we can get the relationship between Q_i and Q_0 as

$$Q_i=\theta_i Q_0+\tau_i \tag{36}$$

where θ_i and τ_i are decided by the model parameter of (18,19) and MAD_O_i.

4.3. QP decision

With this relationship about the quantization step size between frames in window-D, complexity-based frame bit allocation will be introduced together with the INTER-dependent R–Q model. First, we derive the total bit budget of window-D, which is denoted as W_D. Considering time instance t, window-R consists of $L-1$ already encoded frames and the current frame, while window-D consists of the current frame and future $M-1$ frames. After M frame time (at time instance $t+M$), all M frames in window-D at time t are already contained in window-R with the window-R sliding M times, meanwhile the first M frames in window-R at time t are excluded from it, which means that the bit budget for window-D at time t should be equal to the total bits of the first M frames in window-R also at time t. Then W_D can be represented as follows,

$$W_D=\sum_{i=0}^{M-1}R_{r,i} \tag{37}$$

Then, we allocate frame bits using the INTER-dependent R–Q model. By summing up the R–Q model (23) for each frame in window-D, we can get that

$$W_D=\sum_{i=0}^{M-1}\left(a_2\frac{SAD\_O_i}{Q_i}+b_2\right) \tag{38}$$

Substituting (36) into (38), the relationship between W_D and Q_0 can be obtained. Since it is a high order equation and gets harder to resolve with the increasing of the window size M, we use the average quantization step size $\overline{Q}$ in window-D replacing Q_i to obtain a simple yet efficient

solution. Then (38) becomes

$$\overline{Q} = a_2 \sum_{i=0}^{M-1} SAD\_O_i/(W_D - Mb_2) \tag{39}$$

From (36) and (39), the quantization step size of the first frame in window-D is derived as

$$Q_0 = \left(M\overline{Q} - \sum_{i=0}^{M-1} \tau_i\right) / \sum_{i=0}^{M-1} \theta_i \tag{40}$$

Q_0 is the quantization step size of the current frame which takes into account both the fluctuation of video quality and the encoding complexity.

The above-mentioned quantization step size Q_0 is derived from the window-D for consistent video quality purpose, which can be renamed as Q_D. Meanwhile, from window-R, we can also obtain the quantization step size of the current frame. Using the bit budget R_T from (25), a quantization step size for bit rate constraint can be derived from (23), which is denoted as Q_T. Similar to the video quality optimization in window-D, we also use the conclusion (32) to smooth the video quality in window-R. Let D_C denote the distortion of the current frame, to minimize the variance of distortion in window-R, the following equation can be get from (32)

$$D_C = \frac{1}{L-1} \sum_{i=0}^{L-2} D_{r,i} \tag{41}$$

where $D_{r,i}$ represents the real distortion of previously encoded ith frame in window-R. Then the quantization step size Q_C is derived from (19) as follows

$$Q_C = \frac{D_C - b}{a} - MAD\_O_C^2 - k^2 D_{r,L-2} \tag{42}$$

Then we use the average of Q_T and Q_C to represent the quantization step size from window-R, which is denoted as Q_R. The generation of Q_R considers the bit rate constraint and the smooth video quality in window-R. After that, to balance the bit rate constraint and the fluctuation of video quality, the final quantization step size Q_F of current frame can be derived as follows

$$Q_F = \delta Q_R + (1-\delta) Q_D \tag{43}$$

where δ denotes a weighting factor, which is set to 0.5 in our study. The larger δ makes more accurate bit rate and lower buffer latency, while being smaller brings more consistent video quality.

4.4. Optimization of RD performance

To further improve the RD performance, the Lagrange multiplier should also be adjusted under the INTER-dependent situation. We also take two frames into consideration for simplifying the problem as in Section 2. The formulation (2) can be rewritten as (44) by introducing the Lagrange multiplier.

$$\min\{J_{id}\}, \text{where } J_{id} = (D_0 + D_1) + \lambda_{id}(R_0 + R_1) \tag{44}$$

where λ_{id} denotes the Lagrange multiplier in the INTER-dependent environment and J_{id} is the Lagrange cost function, D_0 and R_0 are the distortion and bit rate of the current frame, D_1 and R_1 are the distortion and bit rate of the next frame which references the current frame. It has been proved that the solution of unconstrained problem (44) is the solution of (2) as well [15]. Since our window-based rate control scheme operates frame by frame, the target of the Lagrange optimization is to find the relationship between the Lagrange multiplier λ_{id}, quantization step size Q of current frame and correlations of adjacent frames. Considering that the R–D curve is convex, and both R and D are differentiable everywhere, we can get the solution of (44) by setting its derivative to zero, which is

$$\frac{dJ_{id}}{dR_0} = \frac{d(D_0 + D_1)}{dR_0} + \lambda_{id}\frac{d(R_0 + R_1)}{dR_0} = 0 \tag{45}$$

From Section 3.2 we know that the bit rate of one frame has a weak relationship with the bit rate of its reference frame. So we can get that $dR_1/dR_0 = 0$. Then (45) becomes

$$\lambda_{id} = -\frac{d(D_0 + D_1)}{dR_0} \tag{46}$$

Substituting the R–D model (19,23) into (46), we can get

$$\lambda_{id} = -\frac{\partial(D_0 + D_1)/\partial Q_0}{\partial R_0/\partial Q_0} = \frac{a + (a(\partial Q_1/\partial Q_0) + ak^2(\partial D_0/\partial Q_0))}{a_2 \bullet SAD\_O_0 \bullet Q_0^{-2}} \tag{47}$$

Here we focus on how the correlation of adjacent frames affects the Lagrange multiplier, therefore decouple the relationship between Q_0 and Q_1 derived from the constraint of constant video quality. The final form of λ_{id} can be represented as

$$\lambda_{id} = \frac{a + a^2k^2}{a_2 \bullet SAD\_O_0} \bullet Q_0^2 \tag{48}$$

where a and a_2 are model parameters similar with (19) and (23), k denotes the impact of the distortion of the reference frame to the next frame as discussed in Section 2.

From (48), some implications of the relationship about Lagrange multiplier, quantization step size and correlations between adjacent frames can be further studied. Firstly, the most popular and efficient RDO used in H.264 [2] derives the Lagrange multiplier λ_{org} from

$$\lambda_{org} = c \bullet Q^2 \tag{49}$$

where c is a constant and experimentally set by 0.85 [26]. This optimization is under the assumption of independent coding unit (either frame or MB). Our proposed λ_{id}, which is for the INTER-dependent situation, is actually an extended form of λ_{org}. Both INTER-dependency (parameter k) and coding complexity (SAD_O) are introduced in the determination of the Lagrange multiplier besides the quantization step size. Then, we can see that λ_{id} is inversely proportional to SAD_O, which means a higher coding complexity (usually occurs at scene change or high motion) makes a smaller λ_{id} and vice versa. Since the Lagrange multiplier represents a tradeoff between the distortion and bit rate, a small λ_{id} implies that the distortion should be more concerned than bit rate in the coding process under this kind of situation. Furthermore, the parameter k, which represents the impact of distortion between one frame and its reference frame, also affects the λ_{id} with a nearly quadratic relationship. A larger k makes a larger λ_{id}, which means the bit rate becomes more important since the

correlation of the adjacent frames is high (maybe stable scenes or slow motion).

After the λ_{id} is determined by (48), it can be used in the encoding process. However, there are some practical aspects about the implementation which is worth discussing. Firstly, when using (48) at MB level RDO, *MAD_O* is adopted instead of *SAD_O* to obtain the similar order of magnitude comparing with λ_{org}. Secondly, according to (48), λ_{id} is affected by many parameters, which increases the probability of its fluctuation. The fluctuation of λ_{id} should be avoided to guarantee that neither distortion nor bit rate will be emphasized too much. Considering the good performance of λ_{org} in H.264, we use λ_{org} to limit λ_{id} in the real application. The range of λ_{id} is $(B_{low}\bullet\lambda_{org}, B_{up}\bullet\lambda_{org})$, where $B_{low}=0.8$ and $B_{up}=2$ are obtained experimentally. Thirdly, the parameters in (48) are updated after each frames encoded. Five previously encoded frames are used for estimation by averaging them to further reduce the fluctuation.

4.5. Model parameter update

Model parameters have a significant impact on the bit rate accuracy and the consistent video quality in rate control during the encoding process. In our proposed rate control scheme, the model parameters are updated frame by frame using linear regression with the real data after each frame encoded. The detailed updating steps are listed as follows.

Step 1. Updating parameters k and t in (8) using MAD_O_2, real data of MAD_2 and D_1;
Step 2. Updating parameters a and b in (19) using Q_2, MAD_O_2, real data of D_1 and D_2, and k from step 1;
Step 3. Updating parameters a_2 and b_2 in (23) using *SAD_O*, Q and real data of R;
Step 4. Updating parameters in (48) using a, a_2 and k from above steps.

It should be noticed that the real data of five recent frames are used in the linear regression process. The final updated new parameters are clipped to the range of $[0.5\bullet parameter_{old}, 2.0\bullet parameter_{old}]$ to prevent the abrupt fluctuation.

5. Experimental results

The proposed rate control scheme is implemented on the JM18.5 of H.264/AVC. Several video sequences are tested using the configuration as follows: IPPP coding structure with GOP length of 15, 2 reference frames, RDO on and CABAC, 30f/s frame rate. The frame number of the window-R is set to 30, while the number of window-D is set to 10. Tested video sequences include CIF, 720P and 1080P format. To compare the performance between our rate control scheme and the state-of-the-art works, we also use the algorithm in JM [7], Xie's work [9] and Xu's work [10] as benchmarks.

The rate control accuracy is represented by the bit rate mismatch between the target bit rate R_{target} and the actual bit rate R_{actual} as follows.

$$Err = \frac{|R_{target} - R_{actual}|}{R_{target}} \times 100\% \tag{50}$$

Table 5
Comparison of the bit rate accuracy among different rate control schemes.

Format	Sequence	R_{target} (kbps)	RC in JM		Xie's		Xu's		Proposed	
			R_{actual}	*Err* (%)	R_{actual}	*Err* (%)	R_{actual}	*Err* (%)	R_{actual}	*Err* (%)
CIF	Foreman	1000	1004.26	0.43	994.35	0.56	997.24	0.28	1002.04	0.20
		500	498.61	0.28	507.45	1.49	498.28	0.34	499.38	0.12
	News	1000	1003.95	0.40	992.86	0.71	998.34	0.17	998.97	0.10
		500	504.31	0.86	504.67	0.93	501.48	0.30	501.21	0.24
	Mobile	1000	1005.21	0.52	991.34	0.87	998.67	0.13	997.86	0.21
		500	502.89	0.58	503.97	0.79	498.65	0.27	500.83	0.17
	Akiyo	1000	1003.51	0.35	993.87	0.61	997.87	0.21	996.86	0.31
		500	502.23	0.45	503.42	0.68	498.65	0.27	500.91	0.18
720P	City	8000	8015.64	0.20	7962.07	0.47	7975.61	0.30	7976.32	0.30
		5000	5020.13	0.40	4983.51	0.33	5010.34	0.21	5013.97	0.28
	Crew	8000	8050.37	0.63	7940.91	0.74	8031.84	0.40	7970.08	0.37
		5000	5021.62	0.43	5042.13	0.84	4986.38	0.27	5016.97	0.34
	Night	8000	8020.32	0.25	7950.89	0.61	7986.54	0.17	7989.65	0.13
		5000	5014.87	0.30	4972.65	0.55	5014.36	0.29	5010.37	0.21
	Harbour	8000	8018.52	0.23	7959.97	0.50	7984.64	0.19	8012.67	0.16
		5000	5011.64	0.23	4987.61	0.25	5013.24	0.26	4987.34	0.25
1080P	Blue_sky	12,000	12,063.54	0.53	12,105.34	0.88	11,952.48	0.40	12,031.64	0.26
		8000	7982.1	0.22	8049.87	0.62	7986.34	0.17	7978.35	0.27
	Mobcal_ter	12,000	11,963.1	0.31	12,084.62	0.71	12,022.21	0.19	11,984.67	0.13
		8000	7977.08	0.29	8029.67	0.37	7996.34	0.05	8015.31	0.19
Average			–	**0.39**	–	**0.68**	–	**0.24**	–	**0.22**

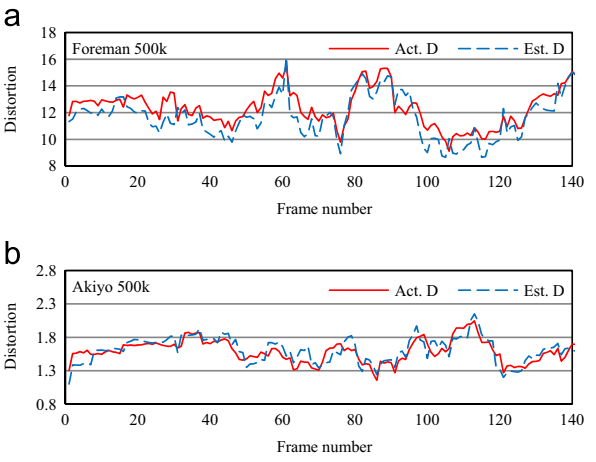

Fig. 3. Accuracy of INTER-dependent D–Q model (19) during encoding process. (a) "*Foreman*"; (b)"*Akiyo*". Act. D: Actual distortion after each frame encoded. Est. D: Estimated distortion via (19) before encoding.

The result of the bit rate accuracy can be seen in Table 5. In Xie's work, the target frame bit is derived by meeting the consistent video quality and the buffer constraint, where the bit rate constraint of GOP or "window" structure is not considered. The assumption of this scheme is that meeting the buffer constraint will automatically guarantee the convergence of the actual bit rate to the target bit rate, as long as the buffer size is negligible with respect to the total size of the compressed bit stream. However, this frame bit allocation scheme is not accurate enough and the bit rate accuracy is inferior to other three methods as shown in Table 5. Both Xu's work and our proposed scheme achieve better bit rate accuracy than the algorithm in JM by using more reasonable encoding complexity obtained from the pre-analysis, while the latter only uses predicted encoding complexity instead, which has certain mismatch at high motion scenes, e.g. "*Foreman*" and "*Night*".

The accuracy of proposed INTER-dependent D–Q model (19) during encoding process with window-based rate control is also tested. The detailed results are shown in Fig. 3. The model parameters are initialized experimentally and updated frame by frame after encoding using the method represented in Section 4.5. From Fig. 3 we can see that the estimated distortions via model (19) before encoding are highly matched with the actual distortions after encoding. With the updating of model parameter during encoding process, the proposed model can represent the relationship between distortion and quantization step size from various video contents either the high motion scenes (Fig. 3(a) "*Foreman*") or smooth scenes (Fig. 3(b) "*Akiyo*").

The performance of the rate control scheme is represented by R–D performance. From Fig. 4, we can see that both our scheme and Xu's work has higher R–D performance than other two algorithms for that the content complexity is considered in frame bit allocation, meanwhile certain characteristics of further frames are also involved to make up a group for allocating bit quota to each frame. It should be pointed out that the encoding complexity used in these two methods, either MAD in our work or the percentage of zeros

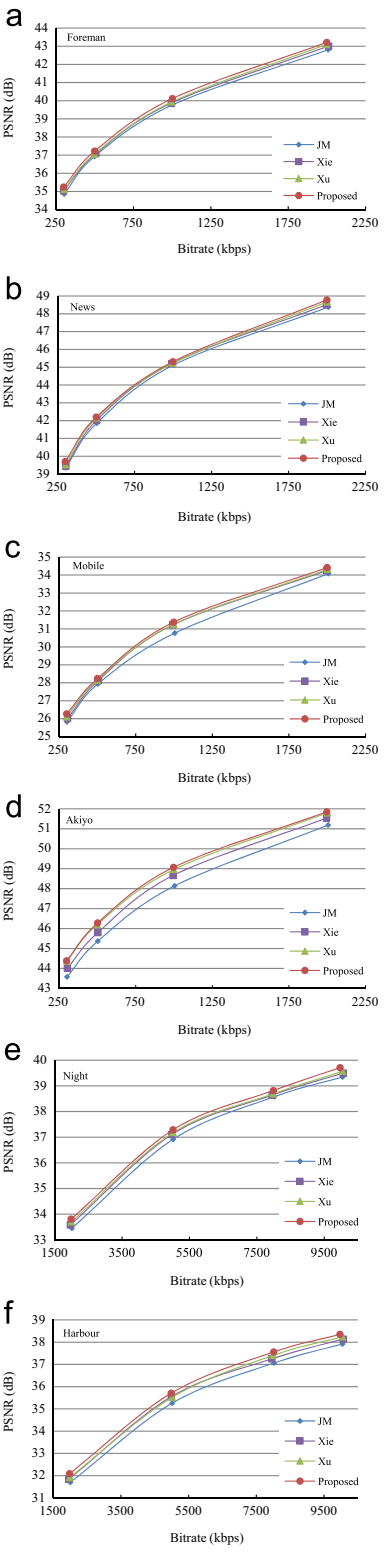

Fig. 4. Comparison of R–D performance among different rate control schemes. (a) "*Foreman*"; (b)"*News*"; (c) "*Mobile*"; (d) "*Akiyo*"; (e) "*Night*"; and (f) "*Harbour*".

among the quantized coefficients (ρ-domain) in Xu's work, is obtained from the original video frames. This encoding complexity has some mismatch with the real complexity

which is generated in the encoding process because of the distortion caused by the quantization. In other words, this complexity is INTER-independent. By further considering the INTER-dependency among frames and using the INTER-dependent R–D model, the R–D performance of our work is better than Xu's work in "*Foreman*", "*News*", "*Night*" and "*Harbour*" sequences, which is up to 0.16 dB in "*Foreman*" sequence as shown in Table 6. The optimization of Lagrange multiplier represented in Section 4.4 also leads to a gain of R–D performance about 0.05∼0.11dB. The detailed data is shown in Table 7. In Xie's work, only MAD of current frame is used for allocating frame bits. Its R–D performance is not as good as the just-mentioned two methods. The algorithm in JM allocates frame bits only according to the frame type within a GOP, which makes I frames get inappropriate bits budgets. For example, in "*Foreman*" sequence (Fig. 5(a)) I frames have less PSNR than other frames for the insufficient bit budgets, while in "*Mobile*" sequence (Fig. 5(b)), the PSNR of I frames are much higher than other frames for the extra bit budgets. This scheme causes the R–D performance not good enough.

The fluctuation of the video quality is measured by the variance of the PSNR (V_{PSNR}). The detailed experimental results are listed in Table 6. Our proposed rate control scheme achieves the least V_{PSNR} mainly because of the INTER-dependent R–D model and the mechanism of window-D for video quality optimization. Among the three benchmark methods, Xu's work gives the best performance of V_{PSNR} since the window model is used to control the QP variation. However, the variance of PSNR is not always coincident with the variance of QP. Therefore, our algorithm which is toward optimizing the fluctuation of PSNR directly has better

Table 6
Comparison of the PSNR and the variance of PSNR among different rate control schemes.

Format	Sequence	R_{target} (kbps)	RC in JM		Xie's		Xu's		Proposed	
			PSNR	V_{PSNR}	PSNR	V_{PSNR}	PSNR	V_{PSNR}	PSNR	V_{PSNR}
CIF	Foreman	1000	39.78	1.56	39.87	0.46	39.92	0.36	39.94	0.21
		500	36.93	2.23	37.09	0.48	37.06	0.31	37.22	0.23
	News	1000	45.14	1.67	45.21	0.75	45.13	0.29	45.24	0.21
		500	41.86	6.05	42.03	0.48	42.19	0.59	42.30	0.11
	Mobile	1000	30.78	7.08	31.12	1.05	31.07	1.13	31.13	0.27
		500	27.74	0.56	28.17	0.39	27.96	0.39	28.01	0.20
	Akiyo	1000	48.14	2.79	48.67	0.38	48.96	0.21	48.97	0.11
		500	45.37	7.97	45.80	0.38	46.20	0.24	46.10	0.21
720P	City	8000	38.63	4.79	38.71	0.76	38.84	0.49	38.96	0.25
		5000	37.38	2.73	37.52	0.57	37.64	0.42	37.72	0.21
	Crew	8000	41.80	3.83	41.93	0.47	42.08	0.36	42.12	0.19
		5000	40.75	2.48	40.97	0.75	41.02	0.39	41.13	0.23
	Night	8000	38.59	2.54	38.64	0.62	38.69	0.49	38.77	0.11
		5000	36.93	1.33	37.10	0.49	37.08	0.34	37.18	0.12
	Harbour	8000	37.07	1.75	37.25	0.34	37.40	0.40	37.44	0.16
		5000	35.26	0.71	35.55	0.34	35.52	0.22	35.61	0.16
1080P	Blue_sky	12,000	43.23	4.22	43.31	1.24	43.42	0.76	43.51	0.35
		8000	42.23	2.95	42.36	0.97	42.41	0.53	42.49	0.32
	Mobcal_ter	12,000	35.69	1.97	35.87	0.53	35.96	0.29	36.05	0.20
		8000	34.71	1.62	34.85	0.49	34.87	0.16	34.92	0.14
Average			**38.90**	**3.04**	**39.11**	**0.60**	**39.17**	**0.42**	**39.24**	**0.20**

Table 7
Comparison of PSNR between proposed rate control with and without λ optimization.

Format	Sequence	R_{target} (kbps)	Proposed RC		Format	Sequence	R_{target} (kbps)	Proposed RC	
			W/O λ	W λ				W/O λ	W λ
CIF	Foreman	1000	39.85	39.94	720P	City	8000	38.87	38.96
		500	37.15	37.22			5000	37.64	37.72
	News	1000	45.19	45.24		Crew	8000	42.01	42.12
		500	42.22	42.30			5000	41.04	41.13
	Mobile	1000	31.03	31.13		Night	8000	38.69	38.77
		500	27.96	28.01			5000	37.10	37.18
	Akiyo	1000	48.90	48.97		Harbour	8000	37.34	37.44
		500	46.02	46.10			5000	35.53	35.61
Average			**39.79**	**39.86**	**Average**			**38.53**	**38.62**

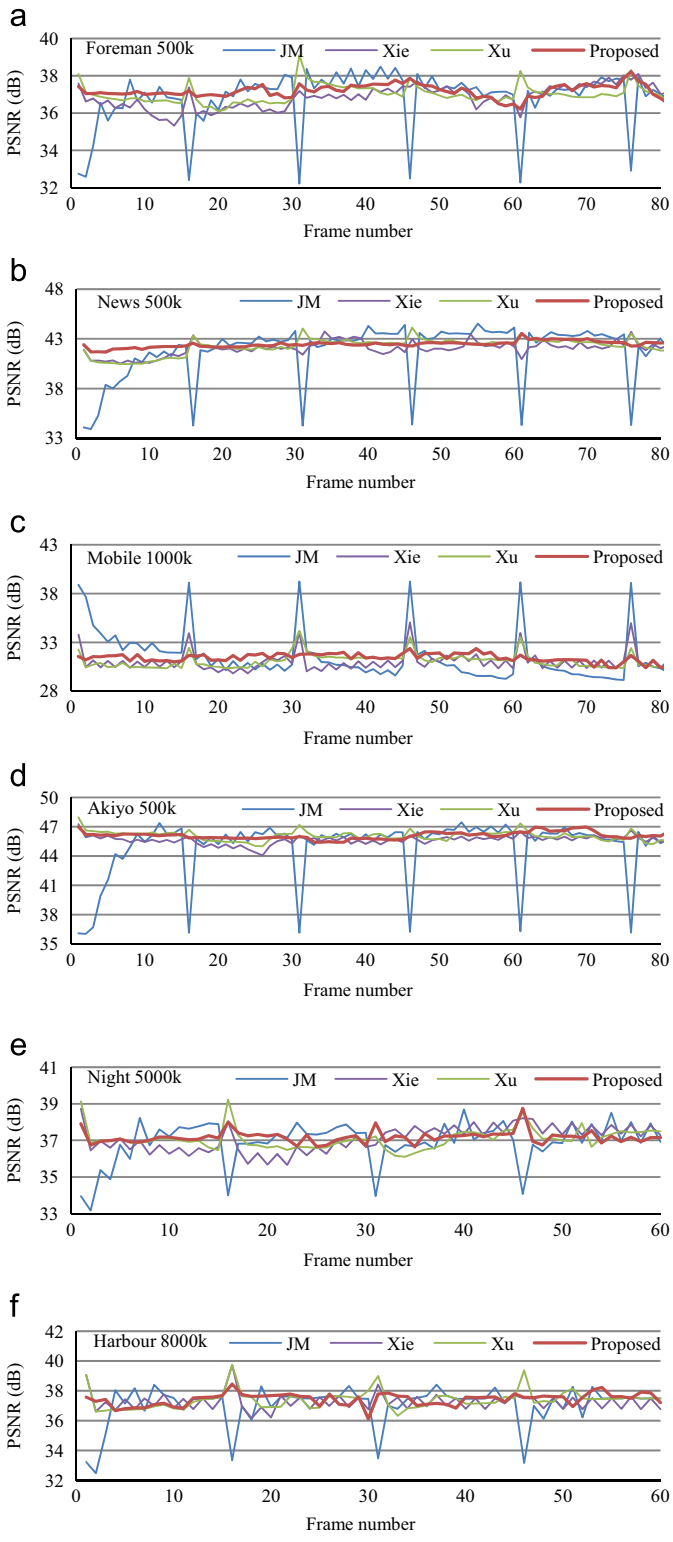

Fig. 5. Comparison of video quality fluctuation among different rate control schemes. (a) "*Foreman*", 0–80 frames; (b) "*News*", 0–80 frames; (c) "*Mobile*", 0–80 frames; (d) "*Akiyo*", 0–80 frames; (e) "*Night*", 0–60 frames; and (f) "*Harbour*", 0–60 frames.

Table 8
Comparison of initial QP and average QP between different rate control schemes.

Format	Sequence	R_{target} (kbps)	RC in JM		Xie's		Xu's		Proposed	
			IQP	AQP	IQP	AQP	IQP	AQP	IQP	AQP
CIF	Foreman	500	35	27.75	27	28.54	26	27.36	28	27.77
	News	500	35	23.82	23	23.93	23	23.71	24	23.68
	Mobile	1000	25	34.87	29	32.57	32	31.68	33	32.73
	Akiyo	500	35	19.55	21	22.13	20	21.37	22	22.05
720P	Night	5000	35	28.72	26	27.55	25	26.35	27	27.24
	Harbour	8000	35	28.73	27	28.68	27	28.51	29	28.43

IQP: Initial QP for the first frame.
AQP: Average QP for the whole video sequence.

performance than Xu's algorithm especially in the complex scenarios, e.g. "*Mobile*", "*Night*" and "*Harbour*" sequences. The PSNR of each frame can be seen in Fig. 5. Moreover, the mechanism of "*jumping*" window in Xu's work cannot well handle the fluctuation of video quality among the boundary of two adjacent windows and thus violating the CBR rule, while our sliding window performs better in this situation and guarantee a CBR characteristics across the entire stream for any given observation point (frame). Xie's method uses MAD to track the nonstationary characteristics of video sequence to allocate frame bits, whose result of V_{PSNR} is better than the algorithm in JM which has no optimization for the fluctuation of video quality.

Besides, the initial QP for the first frame is also quite important. It has been claimed that the best initial QP which provides the best consistent video quality (smallest standard deviation of all PSNRs for whole video sequence) is very close to the average QP for the whole video sequence [36]. We can use this conclusion as the criterion to judge the performance of the initial QP selection. The detailed results are listed in Table 8. The rate control algorithm in JM chooses the initial QP just according to the *bpp* (bits per pixel), without the consideration of the distortion. It causes the improper video quality for the first frame compared with the following frames, e.g. the PSNR of first frame in "*Mobile*" sequence (Fig. 5(c)) is much higher than other frames, while other sequences show opposite results (Fig. 5(a)–(f)). The initial QP of JM is also far from the average QP from Table 8. In Xie's work, the bit allocation for the first frame is set as 40% of the available encoder buffer to adapt to different bit rates or frame rates. Without the distortion optimization, this scheme still could not reach the best initial QP. The result is better than JM, but not as good as Xu's work and our work. Both Xu's work and our work take the first frame and several consecutive frames as a group to allocate frame bits according to the R–Q model, while encoding complexities from pre-analysis are also used. The difference is our proposed scheme use D–Q model to track the video quality, while Xu's work use ΔQ to make the PSNR smooth. Table 8 shows that our work can get the better initial QP than Xu's work.

We also test CBR for random access and the results are shown in Fig. 6. Here random access is defined as the sum of bits for any consecutive frames, which is aimed to simulate the real application environment of the transmission channel.

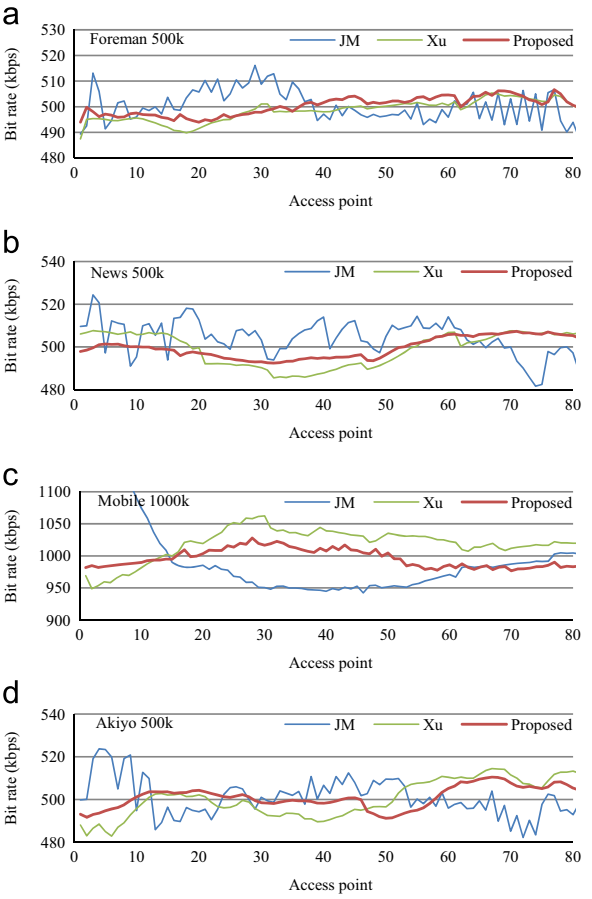

Fig. 6. Comparison of CBR for random access among different rate control schemes. (a) "*Foreman*", 0–80 frames; (b) "*News*", 0–80 frames; (c) "*Mobile*", 0–80 frames; and (d) "*Akiyo*", 0–80 frames.

The number of consecutive frames is set to 30 since the frame rate is 30 f/s. Both Xu's work and our proposed scheme are tested for the using of window mechanism. JM is also listed as benchmark. The results in Fig. 6 demonstrate that our proposed scheme has less fluctuation of CBR for random access than Xu's work. This is because we use the sliding window to keep the bit rate constraint rather than the "*jumping*" window

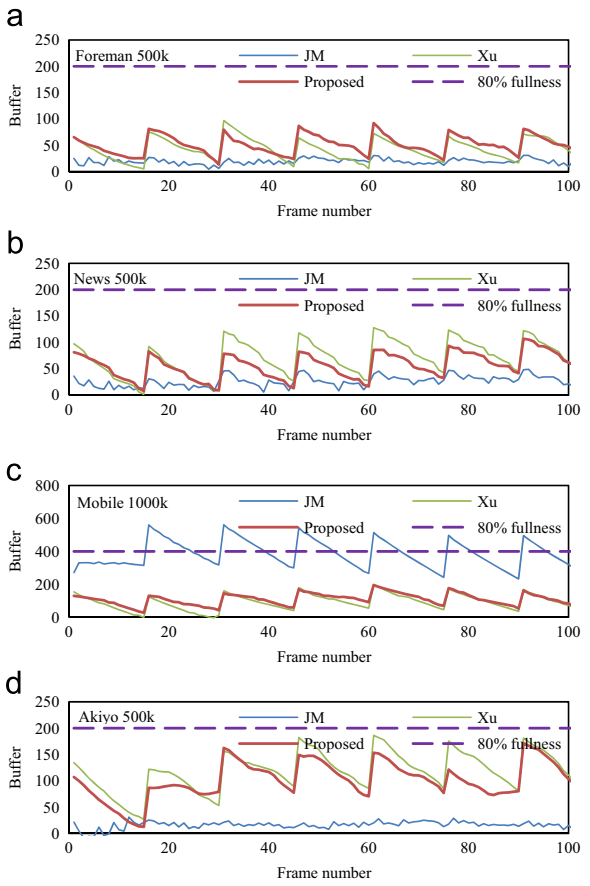

Fig. 7. Encoder buffer status among different rate control schemes, the size of buffer is set as 0.5 s bit rate. (a) "*Foreman*", 500 kbps; (b) "*News*", 500 kbps; (c) "*Mobile*", 1000 kbps; and (d) "*Akiyo*", 500 kbps.

in Xu's work. Both of these two methods have better performance than JM for random access, while there has no consideration for random access in JM. Besides, both bit rate of Xu's work and our scheme have the similar envelopes, the reason is that both of these two methods adopt the complexity-based frame bit allocation.

The encoder buffer status of different rate control schemes during encoding process are shown in Fig. 7. The buffer size is set as 0.5 s bit rate. The dash line for 80% buffer size represents a threshold of potential frame dropping, which means that the next frame will be dropped if the current buffer fullness exceeds this threshold. From Fig. 7, we can see that, the buffer occupations of JM rate control are smooth in most cases due to its frame bit allocation according to frame type. It also means that JM rate control does not use the tolerance of encoder buffer well to assign frame bits according to the encoding complexity among various scenes. Both Xu's work and our work take full advantage of the flexibility provided by the encoder buffer, which is allocating more bits to the more complicated scenes to maintain the consistent video quality. The situation of exceeding 80% buffer size also occurs in the "*Mobile*" sequence of 1000 kbps for JM rate control (Fig. 7(c)). That is because the frame bit allocation does not consider the encoding complexity. On the contrary, Xu's work and our work do not have this situation. It should be noticed that the encoder buffer size also has an effect on the smoothness of video quality. A larger buffer can tolerate more fluctuation of bit rate, thus smoother video quality can be achieved and vice versa. At this point, our proposed rate control scheme is scalable to smooth video quality for different applications.

6. Future work

In this paper, we focus on the frame-level bit allocation and QP decision to reach the bit rate accuracy and consistent video quality. The derivation of INTER-dependent R–D model is also based on the statistical characteristics at frame level. However, considering the different characteristics at MB level, how to extend the INTER-dependent R–D model to MB level is still need to be studied. Besides, most recently works mainly concern about the R–Q model at MB level for accurate bit rate [5,12,14], however the bit allocation considering D–Q model for consistent video quality at MB level is also need to be further discussed.

For HEVC, our proposed window-based rate control scheme with the consideration of consistent video quality can be used directly since the structure of frame level and GOP level at HEVC does not change much than H.264/AVC. However, the INTER-dependent R–D model needs to be adjusted according to the new quadtree-based coding unit (CU) structure. Resent rate control works on HEVC [31,32] have found that the PDF of transform coefficient highly depends on the depths of CU, which will affect the derivation of R–D model. Our future work will also target at the INTER-dependent R–D model and rate control on HEVC.

7. Conclusion

In this paper, we first introduce the concept of INTER-dependency and analyze the INTER-dependent problem and establish the relationship between the residual of one frame and the distortion of its reference frame. Based on this analysis, we derive the INTER-dependent D–Q model and R–Q model via the study of the spatial-domain residual and the transform-domain residual. Then a window-based rate control scheme is proposed with the complexity-based frame bit allocation and video quality optimization. Furthermore, the optimization of Lagrange multiplier is also discussed under the INTER-dependent situation. Experimental results demonstrate that the proposed window-based rate control scheme with INTER-dependent R–D model can achieve accurate target bit rate and improved PSNR performance, meanwhile the variation of PSNR is the smallest compared with other three benchmark algorithms. This one-pass rate control scheme is highly practical for the real-time video coding applications.

Acknowledgment

This work is partially supported by grants from the National Natural Science Foundation of China under Contract no. 61171139, the Major National Scientific Instrument and Equipment Development Project of China under Contract no. 2013YQ030967, and the National High

Technology Research and Development Program of China (863 Program) under Contract no. 2012AA011703.

References

[1] ISO/IEC video recommendation ITU-T H.262, Generic Coding of Moving Pictures and Associated Audio Information, MPEG2, 1995.
[2] ISO/IEC JVT, ITU-T VCEG, ISO/IEC 14496-10 or ITU-T Rec. H.264, Doc. JVT-B118r8, MPEG4/H.264, Feb. 1, 2002.
[3] Information technology advanced coding of audio and video part2: Video. Standard Draft, (AVS), AVS-P2, Mar. 2005.
[4] J.R. Corbera, S. Lei, Rate control in DCT video coding for low-delay communications, IEEE Trans. Circuits Syst. Video Technol. 9 (1) (1999) 172–185.
[5] Z. He, Y.K. Kim, S.K. Mitra, Low-delay rate control for DCT video coding via ρ-domain source modeling, IEEE Trans. Circuits Syst. Video Technol. 11 (8) (2001) 928–940.
[6] ISO/IEC, Coding of moving pictures and associated audio, Test Model 5, (MPEG2) MPEG, JTC1/SC29/WG11, 1994.
[7] Joint model reference software, JVT of ISO/IEC MPEG and ITU-T VCEG.
[8] T. Wiegand, G. Sullivan, A. Luthra, Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H.264ISO/IEC 14 496-10 AVC), Document JVT-G050r1, Geneva, 2003.
[9] B. Xie, W. Zeng, A sequence-based rate control framework for consistent quality real-time video, IEEE Trans. Circuits Syst. Video Technol. 16 (1) (2006) 56–71.
[10] L. Xu, D. Zhao, X. Ji, L. Deng, S. Kwong, W. Gao, Window-level rate control for smooth picture quality and smooth buffer occupancy, IEEE Trans. Image Process. 20 (3) (2011) 723–734.
[11] T. Chiang, Y.Q. Zhang, A new rate control scheme using quadratic rate distortion model, IEEE Trans. Circuits Syst. Video Technol. 7 (1) (1997) 287–311.
[12] S.W. Ma, W. Gao, Y. Lu, Rate-distortion analysis for H.264/AVC video coding and its application to rate control, IEEE Trans. Circuits Syst. Video Technol. 15 (12) (2005) 1533–1544.
[13] D. Kwon, M. Shen, C.C. Jay Kuo, Rate control for H.264 video with enhanced rate and distortion models, IEEE Trans. Circuits Syst. Video Technol. 17 (5) (2007) 517–529.
[14] Y. Liu, Z.G. Li, Y.C. Soh, A novel rate control scheme for low delay video communication of H.264/AVC standard, IEEE Trans. Circuits Syst. Video Technol. 17 (1) (2007) 68–78.
[15] K. Ramchandran, A. Ortega, M. Vetterli, Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders, IEEE Trans. Image Process. 3 (5) (1994) 533–545.
[16] L.J. Lin, A. Ortega, Bit-rate control using piecewise approximated rate-distortion characteristics, IEEE Trans. Circuits Syst. Video Technol. 8 (4) (1998) 446–459.
[17] S. Liu, C.C.J. Kuo, Joint temporal-spatial bit allocation for video coding with dependency, IEEE Trans. Circuits Syst. Video Technol. 15 (1) (2005) 15–26.
[18] J. Liu, Y. Cho, Z. Guo, C.C. Jay Kuo, Bit allocation for spatial scalability coding of H.264/SVC with dependent rate-distortion analysis, IEEE Trans. Circuits Syst. Video Technol. 20 (7) (2010) 967–981.
[19] H.S. Malvar, A. Hallapuro, M. Karczewicz, L. Kerofsky, Low-complexity transform and quantization in H.264/AVC, IEEE Trans. Circuits Syst. Video Technol. 13 (7) (2003) 598–603.
[20] I.M. Pao, M.T. Sun, Modeling DCT coefficients for fast video encoding, IEEE Trans. Circuits Syst. Video Technol. 9 (6) (1999) 608–616.
[21] A.K. Jain, Fundamentals of Digital Image Processing, Prentice-Hall, Englewood Cliffs, NJ, 1989.
[22] E. Lam, J. Goodman, A mathematical analysis of the DCT coefficient distributions for images, IEEE Trans. Image Process. 9 (10) (2000) 1661–1666.
[23] X. Li, N. Oertel, A. Hutter, A. Kaup, Laplace distribution based Lagrangian rate distortion optimization for hybrid video coding, IEEE Trans. Circuits Syst. Video Technol. 19 (2) (2009) 193–205.
[24] N. Kamaci, Y. Altunbasak, 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. 15 (8) (2005) 994–1006.
[25] Y. Li, H.Z. Jia, C. Zhu, M. Li, X. D. Xie, W. Gao, Low-delay window-based rate control scheme for video quality optimization in video encoder, in: Proceedings of the ICASSP, Florence, 2014, pp. 7333–7337.
[26] G.J. Sullian, T. Wiegand, Rate-distortion optimization for video compression, IEEE Signal Process. Mag. 15 (6) (1998) 74–90.
[27] H. Choi, J. Nam, J. Yoo, D. Sim, I. V. Bajic, Rate control based on unified RQ model for HEVC, JCTVC-H0213, 8-th JCTVC Meeting, San Jose, CA, USA, February 2012.
[28] JCT-VC of ISO/IEC MPEG and ITU-T VCEG, HM Reference Software 6.0 [Online], Available: ⟨https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-6.0/⟩.
[29] B. Li, H. Li, L. Li, J. Zhang, Rate control by R-lambda model for HEVC, JCTVC-K0103, in: 11th JCTVC Meeting, Shanghai, China, October 2012.
[30] JCT-VC of ISO/IEC MPEG and ITU-T VCEG, HM Reference Software 10.0 [Online], Available: ⟨https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-10.0/⟩.
[31] C. Seo, J. Moon, J. Han, Rate control for consistent objective quality in high efficiency video coding, IEEE Trans. Image Process. 22 (6) (2013) 2442–2454.
[32] B. Lee, M. Kim, T.Q. Nguyen, A frame-level rate control scheme based on texture and nontexture rate models for High Efficiency Video Coding, IEEE Trans. Circuits Syst. Video Technol. 24 (3) (2014) 465–479.
[33] J. Dong, N. Ling, A context-adaptive prediction scheme for parameter estimation in H.264/AVC macroblock layer rate control, IEEE Trans. Circuits Syst. Video Technol. 19 (8) (2009) 1108–1117.
[34] Y. Zhang, W. Yuan, S. Lin, A new rate control scheme for H.264/AVC, in: Proceedings of the IEEE International Conference on Digital Telecommunications, Cap Esterel, Côte d'Azur, France, August 2006, pp. 13–18.
[35] Z. He, S.K. Mitra, Optimum bit allocation and accurate rate control for video coding via ρ-domain source modeling, IEEE Trans. Circuits Syst. Video Technol. 12 (10) (2002) 840–849.
[36] M. Yang, J.C. Serrano, C. Grecos, MPEG-7 descriptors based shot detection and adaptive initial quantization parameter estimation for the H.264/AVC, IEEE Trans Broadcast. 55 (2) (2009) 165–177.