Consistent Video Quality Control in Scalable Video Coding Using Dependent Distortion Quantization Model

Junhui Hou, *Student Member, IEEE,* Shuai Wan, *Member, IEEE,* Zhan Ma, *Member, IEEE,* and Lap-Pui Chau, *Senior Member, IEEE*

***Abstract*—Consistent video quality control is an important and practical issue for video streaming applications. Several algorithms have been developed in the literature that aims to maintain consistent quality through the entire video sequence, whereas most of them focus on non-scalable video coding. In this paper, we propose an algorithm for consistent quality control for H.264/AVC based scalable video coding (SVC), relying on a dependent distortion-quantization (D-Q) model. Such a dependent D-Q model is developed to capture the distortion behavior of frames at enhancement layers (ELs) by exploring the correlation between two successive layers. Experimental results demonstrate that the proposed model can accurately estimate the distortion of each frame at different layers in the quality, temporal, and spatial scalability. The proposed model is applied in SVC where the quantization parameter of each frame at EL is carefully selected to achieve consistent video quality given the distortion constraint. Meanwhile, model parameters are initialized using the content features extracted from the underlying video sequences, and updated using the encoded data at the frame level. Simulations show that the proposed scheme enables more stable video quality with the PSNR keeping close to the target value (i.e., small PSNR variation).**

***Index Terms*—Consistent video quality control, dependent distortion-quantization model, scalable video coding, H.264/AVC.**

I. Introduction

H.264/AVC based scalable video coding (SVC) has been standardized [1] to enable video services for heterogeneous access networks and clients. In SVC, video signals can be encoded into a single full-resolution stream including a base layer (BL) and one or more enhancement layers (ELs) to provide spatial, temporal, quality and combined scalabilities [2]. The full-resolution bit stream can be adapted to meet diverse requirements from the underlying access network and end-user. It is therefore expected that SVC will be widely adopted in networked video applications without resorting to transcoding or storage of multiple compressed copies at different qualities for the same content.

On the other hand, subjects typically demand a constant video quality [4]. Due to the characteristics of the human visual system, it is annoying and unpleasant if the video being watched has quality fluctuations over the time. To address the problem of quality fluctuation, it is important for the encoder *to maintain a consistent video quality (or without noticeable quality fluctuation) over the entire video sequence*. There has been several related works for non-scalable video coding in the literature [5]–[10]. For example, in [5], a rate control scheme producing a consistent picture quality between consecutive frames was proposed for MPEG-2, which was achieved by a closed-form rate-distortion (R-D) model. A sequence-based bit allocation framework using a rate-complexity model was presented in [7] to achieve smooth video quality with less flickering and motion jerkiness. Wang *et al.* [8] used a two-pass encoding to achieve constant video quality with variable bit rate (VBR) for video storage applications using MPEG-2, whereas two-pass encoding is very computational demanding and is unsuitable for real-time applications. More recently, Huang *et al.*, [9] have proposed a trellis-based framework to alleviate temporal quality variation. Lee *et al.* [10] proposed a frame-level variable bit rate (VBR) encoding method for consistent picture quality with small buffering delay constraint of videos on demand (VOD) system.

Scalable video encoding has the similar problem for consistent quality control. For example, although the CGS (coarse-grain quality scalability) provides quality scalability, it cannot guarantee that the quality (e.g., the peak signal to noise ratio, PSNR) within a particular layer is consistent. Besides, hierarchical B-frames (H-B) [25] are adopted to achieve temporal scalability, which increase the quality fluctuation. On the other hand, there are few algorithms for consistent quality control dedicated to scalable video [11], [12], [13]. For instance, the reference software of SVC, i.e., JSVM, provides a FixedQP scheme which codes the video iteratively to achieve a constant distortion [11]. However, the corresponding complexity due to multiple encoding is extremely high, which makes it inappropriate for real-time applications. A closed form distortion model for spatial ELs is developed in [12], whereas it only allows to use the reference layer signal for prediction rather than adaptive intra and inter layer prediction. Such a restriction will reduce the R-D performance of the SVC [2]. Meanwhile, it also requires intensive computations because it traverses

Manuscript received September 25, 2012; revised February 6, 2013; accepted February 7, 2013. Date of current version December 10, 2013.

J. Hou was with the School of Electronic and Information, Northwestern Polytechnical University, Xi'an 710072, China, and is also with the School of Electrical and Electronics Engineering, Nanyang Technological University, 639798, Singapore (e-mail: houj0001@e.ntu.edu.sg).

S. Wan is with the School of Electronic and Information, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: swan@nwpu.edu.cn).

Z. Ma is with the Dallas Technology Laboratory, Samsung Electronics, Richardson, TX 75082 USA (e-mail: zhan.ma@ieee.org).

L.-P. Chau is with the School of Electrical and Electronics Engineering, Nanyang Technological University, 639798, Singapore (e-mail:elpchau@ntu.edu.sg).

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/TBC.2013.2289632

all possible 52 QPs to find the one producing the minimum distance from the targeted distortion. Quality control for fine grain scalability (FGS) in MPEG-4 has been studied in [13], but FGS has not been adopted into SVC for the current version.

Theoretically speaking, distortion in video coding comes from quantization. Thus, accurate distortion-quantization (D-Q) model is the key for consistent quality control. In this paper, we propose a dependent D-Q model for SVC EL encoding, where the distortion information from the co-located lower layer is used for accurate prediction. The proposed D-Q model is used to estimate the distortion of each frame at different quality, spatial and temporal ELs. It is applied to perform the consistent quality control for EL encoding, where accurate QP for each frame at EL is selected according to the target distortion. Please note that the parameters of the proposed model are first initialized using the content features extracted from the underlying original video sequences, and then updated using the real encoded data at the frame level. Thus, it is a single-pass encoding which reduces the complexity significantly compared with [11] and [12].

The rest of this paper is organized as follows. In Section II, we first give a brief overview of D-Q models for video coding, and then introduce the dependent D-Q model for SVC EL encoding. The model accuracy is verified in the same section. In Section III, we apply the dependent D-Q model to perform the consistent quality control with the model parameter initialized using the content features and frame-level updated using the previous encoded data. Simulation results and discussion are presented in Section IV. Section V concludes the work in this paper.

II. Dependent D-Q Model for SVC Enhancement Layer Encoding

A. D-Q Model for Video Coding: A Review

D-Q modeling is a challenging problem since the distortion behavior is difficult to capture accurately due to the non-stationary characteristics of the input image source and adaptive coding tools in the encoder. Typically, D-Q behavior is modeled through either analytical or empirical approaches, each of which has its own advantages and drawbacks. The analytical approach develops the D-Q models by assuming that the transform coefficients follow certain mathematical distributions, i.e., Gaussian, Cauchy, Laplacian or others. However, the model performance is bounded because the real source distribution is not always consistent with the ideal mathematical distribution. Empirical approach usually establishes the model for a given encoder by extensive experimental simulations, which usually yields better estimation of the D-Q curve.

1) *D-Q Model for Non-Scalable Video Coding:* There are several D-Q models developed for non-scalable coding in the literature [14]-[19], with different functional forms, including quadratic [14], linear [15], power function [16], and so on. One classic D-Q model is formulated in [14] as

$$D = \eta \cdot Q_s^2, \tag{1}$$

where D denotes the distortion of a reconstructed frame in terms of the mean squared error (MSE)[1], Q_s is the quantization stepsize (QS), and η is the model parameter. However, (1) has significant prediction errors at large QSs since it is derived under the high rate assumption. Wang *et al*. [15] proposed an empirical linear D-Q model for H.264/AVC. It is recognized as

$$D = \gamma \cdot Q_s, \tag{2}$$

where γ is the model parameter. Meanwhile, a power function model is derived under the assumption that discrete cosine transform (DCT) coefficients follow the Cauchy distribution [16], i.e.,

$$D = \xi \cdot Q_s^{\alpha}, \tag{3}$$

where ξ and α are model parameters. As we can see, (3) is a generalized functional form for (1) and (2) with $\alpha = 2$ and $\alpha = 1$, respectively. It is demonstrated by simulations that the linear model in (2) is already very accurate compared with other models [20]. Meanwhile, there is only one parameter in the linear model, requiring less complexity for its derivation.

Apart from these pixel-domain models, there are some other distortion models in the transform-domain [18], [19]. In [18], the distortion is modeled as an exponential function of ρ, where ρ refers to the percentage of zero transform coefficients. In [19], the D-Q model is written as follows:

$$D = \theta \cdot \mathrm{SATD}(Q_s) \cdot Q_s^p + D_{skip}(Q_s), \tag{4}$$

where θ is a content dependent parameter. Here, $p = 1$ for P and B frames while $p = 1.2$ for I frames. SATD is the sum of absolute transform differences (SATD) of the intra or inter prediction residual, which depends on Q_s. $D_{skip}(Q_s)$ is the distortion of the skip macroblocks. Despite the high accuracy, transform-domain models have the common drawback that it is difficult to be used in single-pass encoding since the parameter extraction is performed after transform and quantization. Such multiple-pass encoding is also not favored by real-time video applications.

According to the above analysis, in this paper, we choose the linear D-Q model in (2) to develop the dependent D-Q model for SVC EL encoding.

2) *D-Q Model for Scalable Video Coding:* For SVC, on the other hand, there are only few D-Q models [20], [22], [23]. Hu *et al.* [20] proposed to use the linear model in (2) with different model parameters to express the relationship between distortion and quantization at each temporal layer for temporal scalability. In [22], a logarithm distortion model at ELs for CGS is introduced as

$$\mathrm{PSNR} = b_1 \log_{10}(\mathrm{MAD}^{\beta} + 1) \cdot Q_p + b_2, \tag{5}$$

where b_1, b_2 and β are model parameters, Q_p stands for QP, and MAD is the mean absolute difference (MAD) of motion-compensated residual, which is predicted from the BL. However, MAD is not capturing the distortion very well. Besides, it can not be accurately predicted. These drawbacks will reduce the accuracy of the D-Q model in (5). Liu *et al.* [23] developed a dependent D-Q model for ELs regarding the spatial scalability by exploring the distribution of DCT coefficients. However, it is not appropriate for one-pass encoding since it requires off-line processing to derive the model parameters.

[1] D refers to MSE distortion throughout the paper.

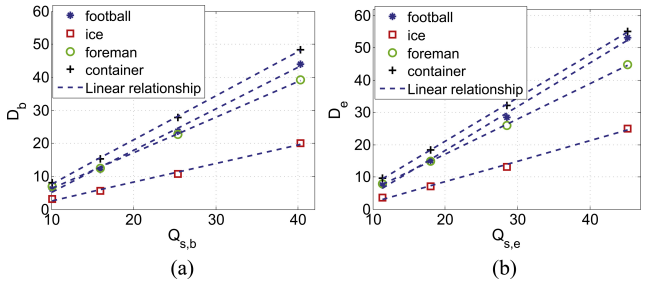

Fig. 1. Illustration of the linear relationship between D and QP for BL and EL, respectively, (a) BL, (b) EL.

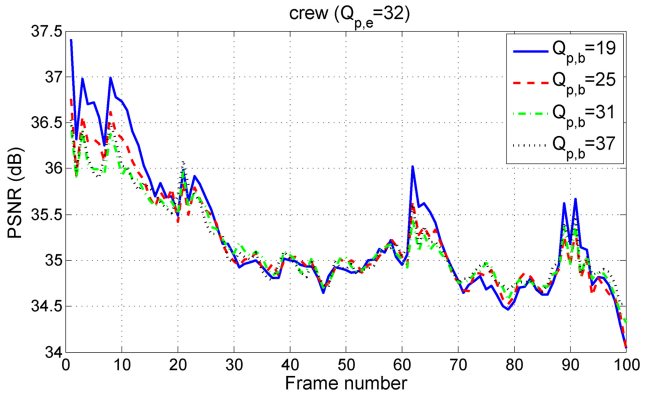

Fig. 2. Illustration of the impact of $Q_{p,b}$ on the quality of EL.

B. Dependent D-Q Model for ELs in SVC

SVC provides the quality, temporal, spatial and combined scalability. Specifically, temporal scalability is implemented by the H-B structure [25], either dyadic or non-dyadic, where frames are partitioned into a temporal BL and one or more temporal ELs. Spatial scalability, on the other hand, encodes the video with the lowest spatial resolution into the BL and those with the higher resolutions into ELs. At ELs, in addition to the normal prediction modes enabled by the non-scalable H.264/AVC, adaptive inter-layer prediction [2] is employed to further reduce the redundancy and improve the coding efficiency. Quality scalability, which can be seen as a special example of the spatial scalability, has the identical video frame size at each layer and typically uses a smaller quantizer at a higher layer to achieve signal amplitude granularity. There are two types of quality scalability referred to as CGS and medium-grain-quality (MGS), respectively. In this paper, we focus on CGS. More details regarding SVC can be found in [2] [3].

Based on our analysis in Sec. II-A and the illustration in Fig. 1, for a typical two-layer SVC encoding structure, i.e., one BL and one EL in quality or spatial scalability, the BL and EL distortion can be respectively expressed as

$$D_b = \gamma_b \cdot Q_{s,b}, \tag{6}$$

$$D_e = \gamma_e \cdot Q_{s,e}, \tag{7}$$

where D_b, γ_b and $Q_{s,b}$ are the distortion in terms of the MSE, content-dependent parameter and QS at BL, respectively; D_e, γ_e and $Q_{s,e}$ stand for the respective information for co-located frame at the EL.

Because of the inter-layer prediction mechanism, the quality of the BL reconstruction will influence the EL encoding. For instance, if we choose a smaller QP for the BL, successive EL picture quality will be much better compared with the case where we choose a larger QP for the BL, under the condition that the EL uses the same QP, as illustrated in Fig. 2. These observations suggest that the EL distortion behavior is closely dependent on the BL reconstruction. Therefore, Eq. (7) can be rewritten as

$$D_e = \gamma_e(Q_{s,b}) \cdot Q_{s,e}, \tag{8}$$

where $\gamma_e(Q_{s,b})$ is a function of $Q_{s,b}$, which reflects the impact of the reconstructed BL frames on the corresponding EL encoding.

For a typical SVC encoder, video frames at the BL are first encoded with appropriate information buffered for EL frame encoding such as intra reconstructed blocks, coding modes, motion information, as well as residual signals. Also, we have the complete distortion and rate information of the BL frame before encoding the EL. Therefore, BL encoding can be seen as the pre-encoding for the corresponding EL coding where BL information can be re-used. Combining Eqs. (6) and (8) with the relationship between QP and QS, i.e., $Q_s = 2^{(Q_p-4)/6}$, we have

$$D_e = f(Q_{p,b}) \cdot 2^{\Delta/6} D_b, \tag{9}$$

with $f(Q_{p,b}) = \gamma_e(Q_{s,b})/\gamma_b$ and $\Delta = Q_{p,e} - Q_{p,b}$.

It is difficult to find a closed form for $f(Q_{p,b})$ from theoretical analysis since complex prediction modes are involved. Hence, we choose to model $f(Q_{p,b})$ numerically. Videos were encoded using different BL QSs and different Δ. BL and EL distortions were collected and shown in Fig. 3. It is observed that $2^{\Delta/6} D_b$ and D_e are linearly correlated, and the slope of the linear function is nearly constant for different Δ .

Then, Eq.(9) can be formulated as

$$D_e = X_1 \cdot 2^{\Delta/6} D_b + X_0, \tag{10}$$

where X_1 and X_0 are both model parameters, which can be refined (or updated) along with the frame encoding using least-square-error (LSE) criteria.

Model (10) can be easily extended to support multiple EL encoding since the same encoding technologies, i.e., intra or inter-layer prediction, are employed in all ELs. For each pair of the two successive layers, the lower layer can be seen as the *virtual* BL while the higher layer is the *virtual* EL. Then, the extended model can be expressed as

$$D_k = X_{1,k} \cdot 2^{\Delta_k/6} D_{k-1} + X_{0,k}, \quad k = 1, 2, \ldots \tag{11}$$

where the k is the layer number, and $\Delta_k = Q_{p,k} - Q_{p,k-1}$.

C. Prediction Accuracy Verification in Real-time Coding

We conducted experiments to evaluate the model performance under both quality and spatial scalabilities with

TABLE I

PREDICTION ACCURACY COMPARISON OF DISTORTION MODELS UNDER QUALITY SCALABILITY. QP={38(BL), 32(EL-1), 26(EL-2)}

		IPPP				H-B (GOPsize:16)			
		EL-1		EL-2		EL-1		EL-2	
	sequence(CIF)	Proposed	Model in [22]	Proposed	Model in [22]	Proposed	Model in [22]	Proposed	Model in [22]
RMSE	ice	0.1016	0.3926	0.0742	0.5844	0.1326	0.2639	0.1182	0.3407
	bus	0.0844	0.3038	0.0687	0.2679	0.1490	0.1886	0.1107	0.1676
	city	0.0936	0.3695	0.1091	0.2473	0.1275	0.1490	0.1296	0.1580
	crew	0.1597	0.2295	0.1540	0.3270	0.1743	0.2894	0.1244	0.2255
	soccer	0.1394	0.2617	0.1627	0.2434	0.1017	0.1574	0.1037	0.1514
	football	0.2337	1.2703	0.2776	0.4246	0.1982	0.4249	0.2251	0.4547
	foreman	0.1057	0.1389	0.0863	0.2315	0.0883	0.1581	0.0992	0.2110
	mobile	0.0824	0.4291	0.0573	0.1059	0.1323	0.2593	0.1129	0.2101
Ave.		**0.1251**	0.4244	**0.1237**	0.3040	**0.1380**	0.2363	**0.1280**	0.2399

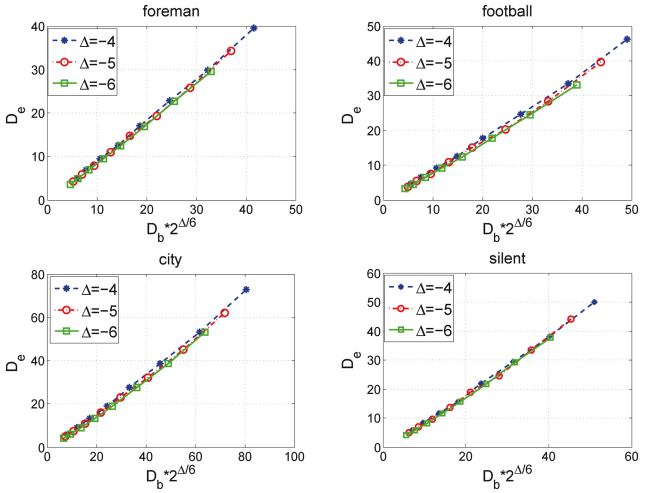

Fig. 3. Illustration of linear approximation between D_e and $2^{\Delta/6}D_b$ (Four different sequences at CIF resolution with 30 frames per second (fps) were encoded with two quality layers).

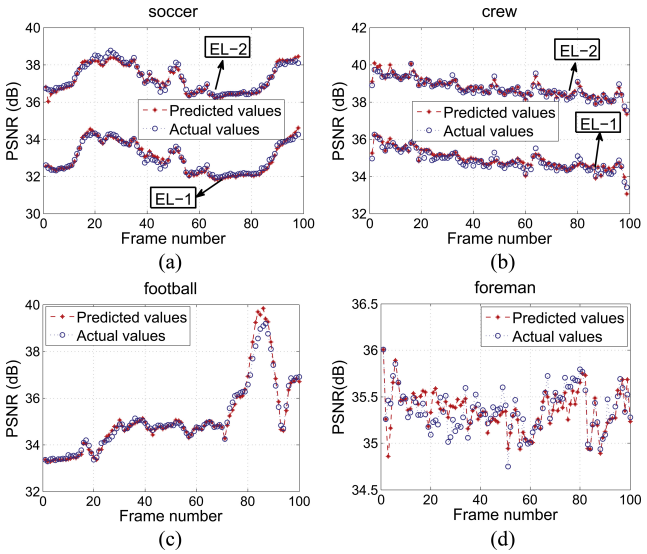

Fig. 4. Performance of the proposed distortion model for EL encoding, (a) Quality scalability, IPPP, (b) Quality scalability, H-B, (c) Spatial scalability, IPPP, (d) Spatial scalability, H-B.

different coding structures, i.e., low-delay IPPP and high efficiency H-B structures.

First, videos were encoded using quality scalability. Video sequences at the CIF resolution with 30 frames per second (fps) were encoded to a BL with QP = 38 and two ELs with QPs of 32 and 26. Model parameters were refined frame by frame along with the encoding using the actual encoded data of previous frames through LSE. For the IPPP structure, only the first frame is encoded as I frame, and the rest are all P frames. For the dyadic H-B structure, GOP (group of picture) length (denoted as GOPL) is 16. We have verified the performance of the proposed model in comparison to the work reported in [22] for EL encoding regarding quality scalability. Experimental results are plotted in Fig. 4(a)(b) at the frame basis. Meanwhile, Table I summarizes the root mean squared error (RMSE) between the actual distortion and the estimated values. These results show that the proposed model outperforms the model in [22] with smaller RMSE.

Moreover, Fig. 4(c)(d) and Table II also show the model's performance for spatial scalability, where sequences with QCIF and CIF resolutions were encoded with corresponding QPs of 26 and 31, respectively. As we can see, the proposed model still provides high accuracy for distortion estimation at the spatial EL.

TABLE II

PREDICTION ACCURACY OF THE PROPOSED MODEL UNDER SPATIAL SCALABILITY. QP={26(BL),31(EL-1)}

sequence	IPPP	H-B (GOPL:16)	sequence	IPPP	H-B (GOPL:16)
ice	0.1113	0.1251	football	0.2534	0.3342
silent	0.0448	0.0795	foreman	0.1200	0.1490
city	0.1608	0.2117	mobile	0.0855	0.1018
crew	0.1828	0.2287	soccer	0.2865	0.3185
Ave.	**0.1557/0.1936**				

III. CONSISTENT QUALITY CONTROL FOR SVC EL ENCODING

In streaming applications, it is annoying for humans to watch video with flickering. Therefore, one of the goals for streaming servers is to provide videos with smooth quality to end-users. In this section, we present a model-based quality

control algorithm for SVC EL encoder to keep a constant quality throughout the entire sequence, where the distortion of each frame at the EL is controlled by adjusting the QP using the proposed dependent D-Q model. The proposed scheme mainly includes two steps, i.e., calculating QP for the EL and updating model parameters frame by frame.

A. QP Derivation for EL Encoding

Given the target distortion at the EL and the available BL QP information, the QP value of the EL $Q_{p,e}$ can be calculated using (10) as

$$Q_{p,e} = Q_{p,b} + 6 \cdot \log_2\left(\frac{D_T - X_0}{X_1 D_b}\right), \tag{12}$$

where D_T is the target distortion in terms of the MSE. It is noted that sometimes rough QPs will be derived for successive frames due to the inaccurate estimation of model parameters. To avoid visual quality flickering among successive frames, $Q_{p,e}$ for current frame is bounded using

$$Q'_{p,e} = \max\{\min\{Q_{p,e}, Q^{n-1}_{p,e} + 2\}, Q^{n-1}_{p,e} - 2\}, \tag{13}$$

$$Q_{p,e} = \max\{\min\{Q'_{p,e}, 51\}, 0\}, \tag{14}$$

where $Q^{n-1}_{p,e}$ is the QP of previous encoded frame.

B. Model Parameters Initialization and Refinement

1) *Model Parameter Initialization:* X_0 and X_1 have to be estimated prior to using (12) for QP derivation. However, encoded information is not available for parameter prediction before encoding the first frame. Intuitively, for a given distortion requirement, a smaller QP is preferable for a video sequence with complex spatial contents. On the other hand, a video sequence with simple spatial textures may require a larger QP. This indicates that model parameters are content dependent. Inspired by this observation, we propose to predict the parameters using the content features from the original video signal.

Let $\mathbf{P} = [X_1 \quad X_0]^T$, $\mathbf{F}$ be the parameter and content feature vector, and $\mathbf{W}$ be the associated weighting matrix for prediction. The model parameters can be calculated as $\mathbf{P} = \mathbf{WF}$.

By analyzing the coding technologies for the I frame, e.g., prediction modes, through extensive simulations, we find that that four content features are sufficient to provide the excellent prediction performance. That is, $\mathbf{F} = [1 \quad V \quad M \quad G \quad S]^T$, where V denotes the averaged variances of 4x4 luminance blocks in the first frame; M is the mean of luminance values of all pixels; $G = \sum_{i=0}^{h-2}\sum_{j=0}^{w-2} g(I_{i,j})/(3(h-2)(w-2))$ is the mean of the differences between adjacent pixels with $g(I_{i,j}) = (|I_{i,j} - I_{i+1,j}| + |I_{i,j} - I_{i,j+1}| + |I_{i,j} - I_{i+1,j+1}|)$, and $S = \left(\sum_{k=1}^{n_{\mathrm{MB}}}\sum_{i=0}^{15}\sum_{j=0}^{15} |I^k_{i,j} - I^k_{dc}|\right)/(1000 n_{\mathrm{MB}})$ denotes the complexity measure of the first frame [21]. Here, w and h stand for the image width and height, respectively. $I(i, j)$ represents the pixel value at the (i, j)-th location within the frame, k means the order of MB, n_{MB} is the total macroblock number in a frame, and I^k_{dc} is the predicted value resulted from the INTRA16 DC mode. All these four features are calculated by analyzing the first frame from the original video sequence.

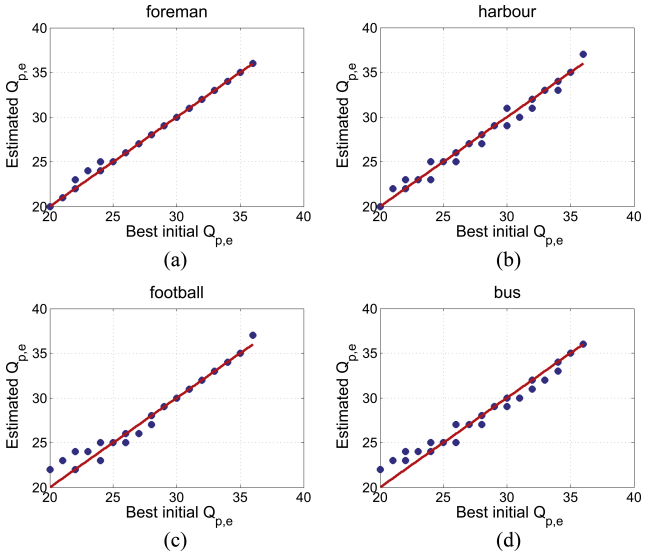

Fig. 5. The performance of proposed initial model parameters decision method to derive the initial QP.

We also have found that the weighting coefficient matrix

$$\mathbf{W} = \begin{bmatrix} 0.991 & 0.039 & -0.07 & -0.055 & 0.028 \\ -0.837 & -0.301 & 0.204 & 0.300 & -0.799 \end{bmatrix} \tag{15}$$

is consistent for all video contents, which is derived using training sequences with different characteristics. These training sequences are different from the sequences used for performance evaluation.

Experimental results with regard to different target PSNRs and $Q_{p,b}$ are shown in Fig. 5 to verify the performance of the proposed model parameter initialization for EL initial QP derivation. It is noted that the proposed scheme can estimate the initial QP for the EL efficiently.

2) *Model Parameter Refinement:* Model parameters in (12) should be updated frame by frame using the actual data of previously encoded frames under the least squared error criterion.

$$X_1 = \frac{n_w \sum \chi^i D^i_b \sum D^i_e - \sum \chi^i D^i_b \sum D^i_e}{n_w \sum (\chi^i D^i_b)^2 - (\sum \chi^i D^i_b)^2}, \tag{16}$$

$$X_0 = \frac{1}{n_w}\left(\sum D^i_e - X_1 \sum \chi^i D^i_b\right), \tag{17}$$

where $\chi^i = 2^{(Q^i_{p,e} - Q^i_{p,b})/6}$, and n_w denotes the number of previous frames before the current frame. n_w is 20 in this paper as suggested by [26].

Please note that the sliding-window and outlier removal scheme proposed in [26] can be integrated in this paper to further improve refinement efficiency.

IV. Performance Evaluation of the Proposed Consistent Quality Control

To evaluate the performance of the proposed algorithm for consistent quality control, it was implemented in the SVC reference software JSVM 9.19.13 [11]. To maximize the coding efficiency, the inter-layer prediction modes were set to

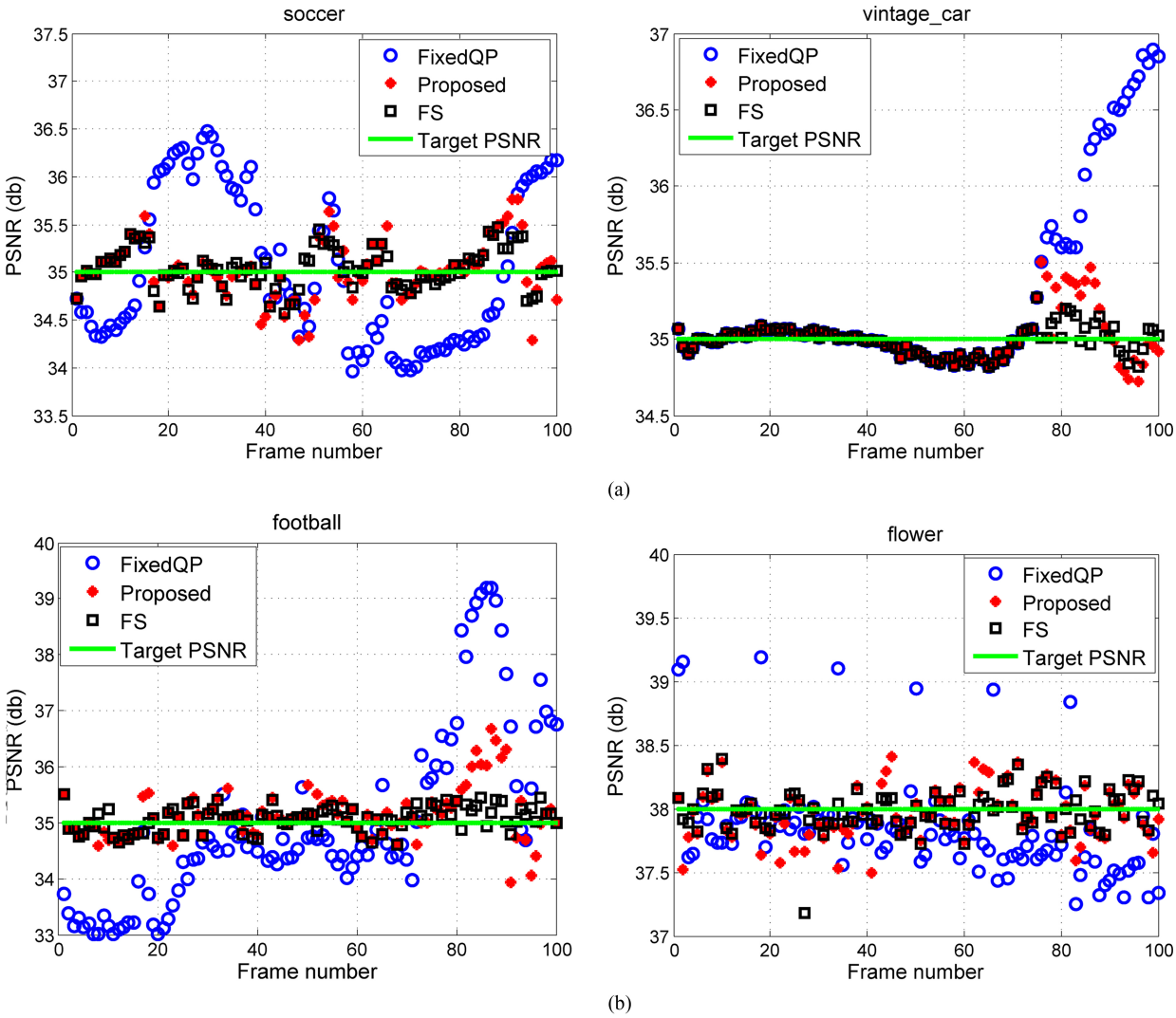

Fig. 6. Frame level EL PSNR of "FixedQP", "FS" (full search) and "Proposed" for CGS, (a) IPPP structure, (b) H-B structure.

"adaptive". The algorithm was tested under the regular IPPP and dyadic H-B coding structures, respectively. In both cases , test sequences were encoded with two quality layers. Similar experiments can be carried out for spatial scalability. Seven sequences with the CIF resolution at 30 fps were tested with various target PSNR PSNR_T, i.e., 41, 38, 35 and 32 in dB. For the ease of discussion, the BL was encoded with a constant QP. It is worthwhile to point out that the proposed scheme is still effective when the BL is controlled with existing constant quality control algorithms. To make QP difference between successive layers reasonable [3], different BL QPs were used for sequences with different contents.

It is noted that the techniques proposed in [8]–[9] cannot be properly applied to H.264/AVC based SVC since they are proposed specifically for non-scalable H.264/AVC, MPEG-2 or FGS. Therefore, for fair comparison, we choose the FixedQP scheme in JSVM and full search (FS) method as the benchmark. In the FixedQP scheme, a logarithmic search algorithm is adopted to find a QP for the entire sequence generating an average PSNR which is closest to the target PSNR. It is usually assumed that a constant QP for the entire video sequence typically yields good coding performance and uniform visual quality [7], [9]. As with the FS method, for each frame, a QP generating a PSNR which is closest to the target PSNR is decided by multiple encoding using all possible 52 QPs. Such FS method can provide the optimal result. The same encoding configurations were applied for comparison.

Fig. 6 provides frame by frame PSNRs for different sequences for both IPPP and H-B coding structures. Average performance of different algorithms are quantitatively summarized in Table III and IV where the average absolute difference μ_{PSNR} between target and actual PSNRs, and the variance of the absolute differences σ^2_{PSNR} are used as the measurement, i.e., $\mu_{\mathrm{PSNR}} = \frac{1}{N}\sum_{i=1}^{N} |\mathrm{PSNR}_e^i - \mathrm{PSNR}_T|$ and $\sigma^2_{\mathrm{PSNR}} = \frac{1}{N}\sum_{i=1}^{N} |\mathrm{PSNR}_e^i - \mathrm{PSNR}_T|^2 - \mu^2_{\mathrm{PSNR}}$.

Meanwhile, R-D performance is presented using both average PSNR decrease (Bjϕntegaard Delta-PSNR; in decibels), denoted as BD-PSNR, and average bit rate increase (Bjϕntegaard Delta-BR; in percentage), denoted as BD-BR [24]. The FixdedQP scheme serves as the reference. The results are listed in Table III, IV and V, respectively. Also, Fig. 7

TABLE III

AVERAGED μ_{PSNR} AND σ^2_{PSNR} UNDER IPPP CODING STRUCTURE. {PSNR_T = {41 dB, 38 dB, 35 dB, 32 dB}, $Q_{p,b}$={28,32,36,40 FOR SOCCER, FOOTBALL, CREW; 31,35,39,43 FOR ICE; 26, 30, 34, 38 FOR VINTAGE_CAR, NIGHT;}}

	FixedQP		FS				Proposed			
Sequence	μ_{PSNR}	σ^2_{PSNR}	μ_{PSNR}	σ^2_{PSNR}	BD-BR	BD-PSNR	μ_{PSNR}	σ^2_{PSNR}	BD-BR	BD-PSNR
soccer	0.6559	0.1313	0.1898	0.0182	-5.44	0.27	**0.2348**	**0.0418**	**-5.34**	**0.26**
football	0.9949	0.8209	0.2479	0.0381	3.16	-0.21	**0.3314**	**0.0876**	**2.44**	**-0.16**
crew	0.3691	0.0968	0.1652	0.0101	1.45	-0.07	**0.1916**	**0.0164**	**1.31**	**-0.06**
ice	0.2907	0.0405	0.1207	0.0066	0.012	-0.001	**0.1523**	**0.0111**	**0.30**	**-0.02**
vintage_car	0.4246	0.1762	0.1303	0.0095	-2.74	0.17	**0.2130**	**0.0255**	**-2.36**	**0.10**
night	0.2322	0.0110	0.1144	0.0042	1.44	-0.09	**0.1602**	**0.0105**	**1.16**	**-0.06**
Ave.	0.4946	0.2128	0.1613	0.0144	-0.353	0.01	**0.2138**	**0.0321**	**-0.41**	**0.02**

TABLE IV

AVERAGED μ_{PSNR} AND σ^2_{PSNR} UNDER H-B CODING STRUCTURE. {PSNR_T={41 dB, 38 dB, 35 dB, 32 dB}, $Q_{p,b}$={26,30,34,38 FOR SOCCER, FOOTBALL, VINTAGE_CAR; 28, 32, 36,40 FOR CREW; 31,35,39,43 FOR ICE; 24, 28, 32, 36 FOR FLOWER;}, GOPL =16}

	FixedQP		FS				Proposed			
Sequence	μ_{PSNR}	σ^2_{PSNR}	μ_{PSNR}	σ^2_{PSNR}	BD-BR	BD-PSNR	μ_{PSNR}	σ^2_{PSNR}	BD-BR	BD-PSNR
soccer	0.7553	0.1230	0.1785	0.0138	1.61	-0.09	**0.3370**	**0.0790**	**2.20**	**-0.12**
football	1.2153	0.8332	0.2396	0.0839	3.68	-0.23	**0.3886**	**0.2121**	**2.31**	**-0.14**
crew	0.4572	0.1085	0.1335	0.0086	3.14	-0.15	**0.1681**	**0.0177**	**3.07**	**-0.14**
ice	0.4313	0.0990	0.1427	0.0093	-0.30	0.02	**0.2210**	**0.0364**	**0.55**	**-0.03**
vintage_car	0.5440	0.1860	0.1602	0.0147	3.72	-0.18	**0.2564**	**0.0442**	**3.28**	**-0.16**
flower	0.3201	0.0871	0.1514	0.0113	5.51	-0.33	**0.2014**	**0.0253**	**5.38**	**-0.31**
Ave.	0.6205	0.2394	0.1676	0.0236	2.89	-0.16	**0.2621**	**0.0691**	**2.80**	**-0.15**

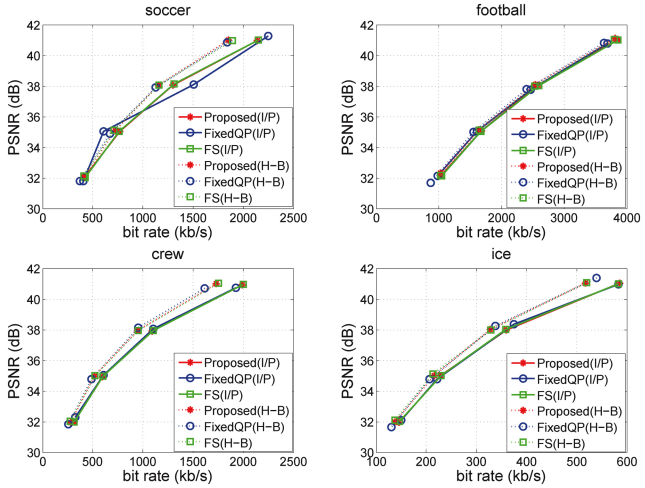

Fig. 7. Rate distortion curves comparison for both IPPP and H-B structures under different schemes.

plots the average PSNRs of the encoded frames evaluated at different bit rates.

From above extensive experimental results, we can derive some conclusions as follows:

(1) For sequences with intensity motion, scene change or encoded using the H-B structure, using the constant QP for the entire sequence is not suitable for consistent video quality control.

(2) Compared with the FixedQP scheme, the proposed algorithm decreases the PSNR variation over the overall sequence remarkably under the both coding structures, and the PSNR values of each frame is closer to the PSNR_T, with μ_{PSNR} reduced by 57.26%, and σ_{PSNR} reduced by 78.03%. On the other hand, the performance of the proposed method is very close to FS, which provides the optimal results relying on exhaustive search.

(3) Prior to encoding the first frame, the proposed method for model parameters prediction can give a suitable QP for the first I frame to approach the PSNR_T.

(4) The proposed algorithm has similar or superior R-D performance to the other schemes. In other words, it does not decrease the coding efficiency.

(5) Furthermore, the proposed method requires single-pass process, which significantly reduces the complexity compared with existing solutions. Thus, it is useful for practical applications.

TABLE V

PERFORMANCE COMPARISON BETWEEN PROPOSED ALGORITHM AND METHOD IN [12] IN TERMS OF AVERAGE RD GAIN OVER FIXEDQP

	Proposed method	Scheme in [12]
Average BD-PSNR	-0.23	-0.49
Average BD-BR	5.1	16.7

V. CONCLUSION

This paper presents an efficient algorithm for consistent video quality control for enhancement layer encoding of H.264/AVC based scalable video coding. With the knowledge of coded frames at the co-located lower layer, such as the distortion and quantization parameters, a dependent distortion-quantization model is employed for enhancement layer en-

coding which can well capture the distortion-quantization behavior. Model parameters are initialized using four content features extracted from the first frame before encoding, and then they are refined and updated frame by frame using encoded data from previous frames. Experimental results demonstrate that the proposed method yields consistent video quality comparable to that using the optimal full search method, with a significantly reduced complexity.

REFERENCES

[1] *Joint Draft ITU-T Rec. H.264-ISO/IEC 14496-10/Amd.3 Scalable Video Coding*, Standard, 2007.

[2] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[3] X. Li, P. Amon, A. Hutter, and A. Kaup, "Performance analysis of Iinter-layer prediction in scalable video coding," *IEEE Trans. Broadcast.*, vol. 57, no. 1, pp. 66–74, Mar. 2011.

[4] Y. Yu, J. Zhou, Y. Wang, and C.-W. Chen, "A novel two-pass VBR coding algorithm for fixed-storage application," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 345–356, Mar. 2001.

[5] S. H. Hong, S. J. Yoo, S. W. Lee, H. S. Kang, and S. Y. Hong, "Rate control of MPEG video for consistent picture quality," *IEEE Trans. Broadcast.*, vol. 49, no. 1, pp. 1–13, Mar. 2003.

[6] H. Xiong, J. Sun, S. Yu, J. Zhou, and C. Chen, "Rate control for real-time video network transmission on end-to-end rate-distortion and application-oriented QoS," *IEEE Trans. Broadcast.*, vol. 51, no. 1, pp. 122–132, Mar. 2005.

[7] B. Xie and W. Zeng, "A sequence-based rate control framework for consistent quality real-time video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 56–71, Mar. 2006.

[8] K. Wang and J. W. Woods, "MPEG motion picture coding with long-term constraint on distortion variation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 3, pp. 294–304, Mar. 2008.

[9] K.-L. Huang and H.-M. Hang, "Consistent picture quality control strategy for dependent video coding," *IEEE Trans. Image Process.*, vol. 18, no. 5, pp. 1004–1014, May 2009.

[10] H. Lee and S. Sull, "A VBR video encoding for locally consistent picture quality with small buffering delay under limited bandwidth," *IEEE Trans. Broadcast.*, vol. 58, no. 1, pp. 47–56, Mar. 2012.

[11] *Joint Scalable Video Model*, JSVM 9.19.13 Software Package, CVS server for the JSVM software, Mar. 2011.

[12] C.-W. Seo, J.-K. Han, and T. Q. Nguyen, "Rate control scheme for consistent video quality in scalable video codec," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2166–2176, Aug. 2011.

[13] M. Dai, D. Loguinov, and H. M. Radha, "Rate-distortion analysis and quality control in scalable internet streaming," *IEEE Trans. Multimedia*, vol. 8, no. 6, pp. 1135–1146, Dec. 2006.

[14] T. Berger, *Rate-Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1971.

[15] H. Wang and S. Kwong, "A rate-distortion optimization algorithm for rate control in H.264," in *Proc. IEEE ICASSP*, Apr. 2007, pp. 1149–1152.

[16] N. Kamaci, Y. Altunbasak, and R. M. Mersereau, "Frame bit allocation for the H.264/AVC video coder via cauchy-density-based rate and distortion Models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 8, pp. 994–1006, Aug. 2005.

[17] Y. Liu, Z. G. Li, and Y. C. Soh, "A novel rate control scheme for low delay video communication of H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 1, pp. 68–78, Jan. 2007.

[18] Z. He and S. K. Mitra., "A unified rate-distortion analysis frame work for transform coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 12, pp. 1221–1236, 2001.

[19] D.-K. Kwon, M.-Y. Shen, and C.-C. Jay Kuo, "Rate control for H.264 video with enhanced rate and distortion models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 5, pp. 517–529, May. 2007.

[20] S. Hu, H. Wang, S. Kwong, and C.-C. J. Kuo, "Rate control optimization for temporal-layer scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 8, pp. 1152–1162, Aug. 2011.

[21] H. Wang and S. Kwong, "Rate-distortion optimization of rate control for H.264 with adaptive initial quantization parameter determination," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 1, pp. 140–144, Jan. 2008.

[22] H. Mansour, P. Nasiopoulos, and V. Krishnamurthy, "Rate and distortion Modeling of CGS coded scalable video content," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 165–180, Apr. 2011.

[23] J. Liu, J. Cho, Z. Guo, and C.-C. J. Kuo, "Bit allocation for spatial scalability coding of H.264/SVC with dependent rate-distortion analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 7, pp. 967–981, Jul. 2010.

[24] G. Bjϕntegaard, "Calculation of average PSNR differences between RD curves," document VCEG-M33, Austin, TX, USA, Apr. 2001.

[25] H. Schwarz, D. Marpe, and T. Wiegand, *Hierarchical B Pictures*, document JVT-P014, Joint Video Team, Thailand, Jul. 2005.

[26] Z. Li, F. Pan, K. P. Lim, G. Feng, X. Lin, and S. Rahardja, "Adaptive basic unit layer rate control for JVT," Doc. JVT-G012-r1, Joint Video Team, Thailand, Mar. 2003.