VBR Rate Control for Perceptually Consistent Video Quality

Bo Han, Bingfeng Zhou,

Abstract — *In contrast with traditional VBR control strategies, such as constant QP and constant PSNR, we take the characteristics of human visual system into account and propose a two-pass VBR rate control for perceptually consistent video quality, mainly for DVD like video storage applications. We employ spatial-temporal complexity to evaluate the subjective visual quality and propose a statistical quality relation between the Qstep and complexity. We further extend it to a RDO based practical Q-Complexity model to effectively guide our second-pass VBR encoding. In addition, we adopt a fast iterative search technique and a run-time bit production control mechanism to satisfy the constraint of fixed storage capacity while producing consistent perceived quality. Our experimental results from various test clips have demonstrated its efficiency and reliability, indicating a noticeable improvement on video quality*[1]**.**

Index Terms — DVD, rate control, variable bit rate (VBR), consistent quality, rate distortion optimization (RDO), perceptual coding, HVS.

I. INTRODUCTION

With recent success of multimedia entertainment, digital techniques for transmitting and storing video content have become increasingly popular and important. For practical video encoders, rate control is an essential piece to maximize the visual quality while satisfying a set of real world constraints, such as bandwidth, decoding delay, buffer capacity, and computation complexity. In general, rate control methods are classified into two categories: the constant bit rate (CBR) and the variable bit rate (VBR). The CBR algorithm adopts uniform bit allocation to those different coding units, regardless of their characteristics, which results in low coding efficiency and fluctuating perceived quality. In contrast, the VBR algorithm is able to adjust the short-term bit rate dynamically according to characteristics of video content, for the purpose of long term consistent visual quality.

One of typical VBR applications is the digital versatile disk (DVD). Driven by the home entrainment and movie industry, DVD has become one of the most successful consumer electronics. Recently, the increasing popular blu-ray disks (BD) further bring DVDs into a new high definition era. This kind of video applications usually rely on a VBR coding scheme to achieve maximal, consistent visual quality across the entire sequence while under the constraint of fixed storage capacity. It is a hard sequence-level optimal bit allocation problem. In practice, a two-pass or multi-pass coding scheme is usually employed to solve the problem. After analyzing a video sequence in advance to track its content characteristics, we are able to collect the information and extract some specific rules to distribute bits appropriately among various video segments, with fewer bits used in less demanding passages and more bits used in difficult-to-encode passages. By means of that, the VBR encoder has the potential to improve coding efficiency and produce uniform visual quality, which is especially suitable for DVD like video storage applications.

Since humans are the final receivers of any video content, the ultimate target of any successful VBR coding scheme is to achieve perceptually consistent visual quality. However, in practice, visual quality is usually measured by the objective distortion metric, such as mean square error (MSE) and its derivative (PSNR). Based on the well defined *R-D* theory, most VBR encoders assume that maintaining constant distortion can achieve consistent visual quality. Thus, they adopt two popular methods of constant PSNR and constant quantization parameter (QP). However, a number of biological and psychological experiments [11], [13] show that the human visual system (HVS) is less sensitive to errors in regions where there are high spatial frequency patterns and fast temporal movements. Thus, complex video scenes are supposed to tolerate more distortion than easy ones without degrading the perceived visual quality. In short, neither PSNR nor QP is able to effectively measure the perceptual video quality. Therefore, we take the characteristics of HVS into account when allocating bits among different coding units, which gives the basic idea of our proposed work.

In this paper, we aim for a two pass VBR rate control to achieve the *perceptually consistent visual quality* rather than the traditional constant objective distortion. Thus, we take spatial temporal complexity into account when evaluating the subjective visual quality. Based on our statistical quality relation and the framework of rate-distortion optimization (RDO), we propose a practical *Q-Complexity* model for our second pass VBR encoding to effectively track characteristics of video content. In addition, we present a fast iterative search technique to satisfy the constraint of fixed storage capacity and adopt a run-time bit production control mechanism to

[1] This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 60573149.

B. Han and B.F. Zhou are with Institute of Computer Science & Technology, Peking University, China. (e-mail: hanbo @icst.pku.edu.cn, cczbf@pku.edu.cn).

Contributed Paper
Manuscript received September 24, 2008

compensate the model mismatch. Our experimental results from various test clips show its superiority over the previous work, indicating a noticeable improvement on both subjective and objective visual quality.

The rest of paper is organized as follows. In section 2, we give a quick review of related work. Then, in section 3 we present a simple analysis of subjective visual quality by experiments. Afterwards, we give the details of our proposed two-pass VBR control algorithm in section 4 and we show our experimental results in section 5. Finally, section 6 concludes the paper.

II. BACKGROUND AND RELATED WORK

In this section we introduce the principle of two-pass VBR coding scheme and give a quick review of related work.

A schematic overview of a typical two-pass VBR encoder is shown in Fig. 1. The principle of VBR encoding under the constraint of total bits budget is to maximize consistent video quality by taking bits from easy scenes and spending them on difficult ones. Therefore, it is preferred to acquire the characteristics of the entire video sequence in advance and utilize them to allocate bits properly for the actual encoding process. However, how to analyze the collected information to design suitable control techniques for optimal objectives is the kernel part for any VBR rate control algorithm, which actually determines the coding performance of an encoder.

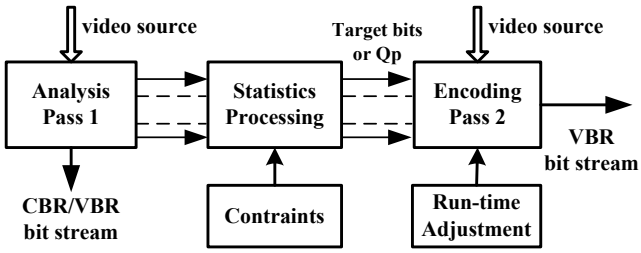

Fig. 1. The schematic overview of a two-pass VBR coding system. The statistics processing stage bridges the first and second pass, which actually determines the overall actions of the system.

A number of VBR rate control algorithms have been proposed for various video storage applications. We give a simple survey as follows.

Single-pass VBR rate control algorithms [3], [5], [8] are mainly used for real-time applications. Most of them assume that constant QP for the entire video sequence typically results in both good coding performance and uniform visual quality. The algorithm in [3] employs a simple bits-budget curve to adjust the QP. The algorithm in [8] utilizes a predefined statistical rate-complexity model to track the varying characteristics of a video source and adopts a nonlinear QP estimation approach to reduce the model mismatch. They gain the advantage of low computational complexity and low encoding delay, but they are hard to achieve a global optimal solution for coding performance due to the unavailability of future pictures in a real-time encoding scenario.

On the other hand, two-pass VBR encoding is quite effective from the perspective of coding performance, but it suffers from relatively high computational complexity, especially when the operational R-D model is adopted [2], [9]. In [2] Yu et al. establish their R-Q model by encoding each frame with all admissible QP values in the first pass. The original frames are used as the references during motion estimation to reduce the computational expense. Then, they follow an empirical bit allocation strategy to allocate fixed total bits among I, P and B frames with a predefined ratio (4: 2:1). In [9] Overmeire et al. proposed an off-line segment-based rate control approach. First, the video is divided into shots by activity analysis and cut detection. Then, each segment is encoded with a set of quantization scales to approximate its R-D characteristics. Based on operational R-D curves for each segment, all available bits are distributed properly among video segments to meet the constraint of constant PSNR.

To reduce computational complexity, some VBR control algorithms adopt analytic or statistical R-D models to control bit allocation [6] and employ the iterative QP selection algorithm to improve the model accuracy [7]. In [6] Wang and Woods formulated VBR coding as an explicit optimal bit allocation problem with constraints of distortion bounds on the individual frames. Based on the Lagrangian method and their statistical R-D model, they give the theoretical optimality conditions and propose a practical iterative solution to achieve constant PSNR among different frames. They also introduce weighted PSNR for consistent subjective visual quality. On the other hand, in [8] Yin et al. proposed another VBR coding algorithm for nearly constant PSNR. After performing the first pass encoding with fixed QPs, they take the PSNR fluctuation into account when calculate each picture's complexity. Then, these complexities are utilized to allocate bits to different frames. In addition, the iterative picture-level QP selection and adaptive macroblock quantization are also proposed to maintain low accumulated bit error and consistent visual quality.

Besides those R-D model based algorithms for constant PSNR, westerink et al. [1] described a two-pass MPEG-2 encoding system for constant perceived quality. Based on a number of perceptual experiments, they found that quantization scales from the CBR encoding pass can effectively track the varying characteristics of video content. Thus, they established a power function of QPs obtained from the first CBR pass to predict the optimal bit allocation for the second encoding pass. Finally, the linear R-Q model in TM5 is adopted to estimate QPs for the actual VBR compression.

In short, most of prior solutions are based upon the traditional R-D theory and the objective distortion metric. However, psychological research suggests that the human visual system does not favor PSNR or QP to measure the perceptual quality. The cited work [1] is one of the few that explicitly targets on perceived visual quality but we find that its efficiency largely depends on the CBR coding algorithm employed in the first pass. The large fluctuation of QPs also greatly influences the consistent visual quality.

III. EVALUATION OF PERCEPTUAL QUALITY

The ultimate objective of a VBR coding system has been defined as perceptually consistent visual quality. However, the commonly used distortion-based quality metric, such as MSE or PSNR, cannot evaluate the perceptual quality well. Therefore, in this section we perform several visual perception experiments in order to determine what constitutes consistent quality for the human observer.

A. Visual perception experiments

We choose six typical scenes from MPEG test sequences. They are different in characteristics, ranging from low to high spatial complexity and motion activity, as shown in the first column of Table I. Each of them was encoded individually by our H.264 encoder with fixed QPs ranging from 4 to 50. We evaluated the subjective visual quality by the similar method described in [1]. Five test viewers participated in our experiment and carefully chose the quantization level for each individual scene to makes them perceived to be of equal quality. Here we describe the visual quality with the subjective terms "good" and "fair". The former "good" indicates that the reconstructed frames of a compressed bit stream is nearly indistinguishable from the original sequence, while the latter means the reconstructed is satisfying without obvious annoying visual artifacts for most viewers. Our experimental results are shown in Table I.

TABLE I : CONSISTENT PERCEPTUAL QUALITY

Scene	Quality(Good)			Quality(Fair)		
	QP (Qstep)	Bitrate (kbps)	PSNR (db)	QP (Qstep)	Bitrate (kbps)	PSNR (db)
M&D	21(7)	331.0	43.23	26(13)	170.7	40.38
Foreman	25(11)	456.2	38.85	29(18)	275.3	36.42
Paris	27(14)	532.5	37.04	31(22)	337.9	34.13
Stefan	29(18)	672.4	34.53	33(28)	389.1	31.73
Football	29(18)	812.4	35.95	34(32)	430.1	32.78
Mobile	30(20)	772.5	32.11	35(36)	368.6	28.57

The statistics listed in Table I clearly demonstrate that neither QP nor PSNR can effectively measure the subjective visual quality. To make it more specific and detailed, we extract two pictures from compressed video streams, *foreman* and *mobilecal*, as shown in Fig.2. We are able to observe annoying blurring and slight blockness in Fig.2 (a), which degrades its subjective quality comparing with Fig.2 (b). We need to mention that both pictures were encoded with the same QP=35 and the picture (a) has an even higher PSNR than (b) (33.53 vs 28.28). It can be explained by the latter's high frequency texture pattern which is able to tolerate more distortion without loss of perceived quality.

(a) (b)

Fig. 2. Comparison of subjective quality. Picture (a) and (b) are the 70th coded frames (P type) of encoded foreman and mobilecal sequences with fixed QP=35 by H.264. (a): PSNR=33.53; (b): PSNR=28.28.

B. Modeling quantizer and complexity

Table I indicates that to reach the similar subjective visual quality, complex video scenes always have higher QPs and lower PSNR values than easy ones. It can be further explained by psychological vision research [11], [13], where evidences show that human visual error sensitivities vary with different spatial temporal frequencies and directional channels. Thus, we are less sensitive to errors where there are high spatial frequency patterns and fast temporal movements. Therefore, those scenes with high spatial temporal complexities can tolerate higher QPs and more distortion than easy scenes, but maintain close subjective visual quality.

Based on these facts, when choosing proper QPs for multiple video segments to achieve consistent perceptual quality, we are supposed to take their spatial temporal complexities into consideration. Typically, a coding unit's complexity can be defined as the product of the coded bit number and the corresponding Qstep.

$$X_i = R_i \times Q_i \tag{1}$$

Where *Ri* is the total bits including both texture and header bits (including motion vector bits). *Qi* is the quantization step (Qstep) which is indexed by QP in practical video standards. *Xi* is the spatial temporal complexity of current coding unit. Additionally, Equation (1) can take the form of (2) to serve as the linear R-1/Q model to control bit-production.

$$R_i(Q_i) = \frac{X_i}{Q_i} \tag{2}$$

When changing QPs in a small range, the linear model (2) has been proven to be reliable for a variety of video scenes with H.264 [17]. The actual R-1/Qstep curves of all six test sequences selected in our perception experiment are plotted in Fig.3, which further verifies the accuracy of this linear model.

We notice that the complexity X_i of each scene actually corresponds to the slope of its linear R-1/Qstep curve. It is a great advantage that it is nearly a constant characteristic for each scene, regardless of the change on Qsteps and bit rates, which makes it a suitable choice for our spatial temporal complexity.

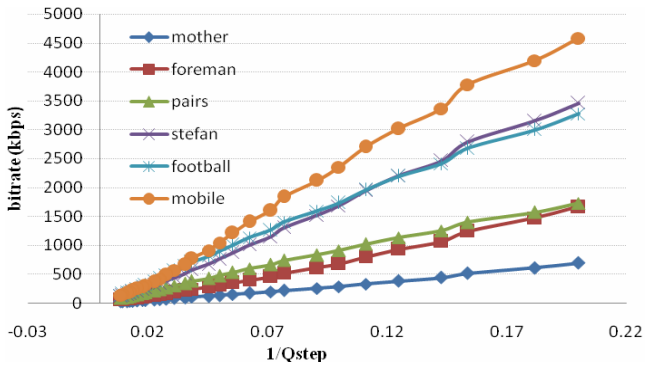

Fig. 3. The linear R-1/Qstep curves of our six test sequences. Different complexities result in different slops.

To illustrate the relationship between the complexity and the Qstep for nearly consistent perceptual quality, we calculated the complexity of each scene according to Table I and plotted these complexities with their corresponding Qsteps for both "good" and "fair" subjective quality, as shown in Fig.4. We can observe a large correlation exists between them. In this work, we choose a power function to fit the samples.

$$Q_i = k(X_i)^{\alpha} \tag{3}$$

Both curves of "good" and "fair" have close α values, typically within a range of 0.4-0.5 by our extensive experiments. Therefore, we argue that for multiple coding units if their complexities and corresponding Qsteps follow the basic trend of Equation (3), we expect to achieve a nearly perceptually consistent visual quality across these different coding units.

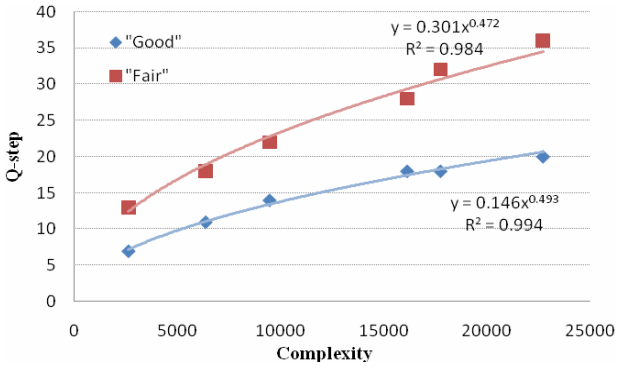

Fig.4. The large correlation between Q-step and spatial-temporal complexity. Power functions are employed to fit samples.

IV. PROPOSED TWO-PASS CODING SCHEME

In this section we give the implementation details of our proposed two-pass VBR rate control algorithm. We establish a practical *Q-Complexity* model for the second encoding pass to achieve sequence-level consistent visual quality. It is derived from the traditional R-D optimization framework and our statistical relation model (3) given in section 3. Furthermore, we employ a fast iterative search technique for the sequence level model parameter to meet the constraint of fixed storage capacity. A run-time bit production control method is also presented to compensate the model mismatch and to monitor the buffer status.

A. Principle of the algorithm

The key and fundamental problem for any VBR coding system is how to allocate the bits among different video segments or pictures. It is a classical optimal bit allocation problem. Let Q be a set of Qstep and N is the number of coding units, based on the traditional theory of RDO we can formulate the following optimal problem:

$$Q^* = \arg\min_{Q_i \in Q} \sum_{i=1}^{N} D_i(Q_i) \tag{4}$$

Subject to

$$\sum_{i=0}^{N} R_i(Q_i) \le R_{total} \tag{5}$$

In addition, one of our main objectives is to achieve perceptually consistent quality across the whole sequence. As a subjective term, visual quality is hard to measure and evaluate. Fortunately, according to our previous analysis in section 3, if the relation between the Qstep and the coding unit's complexity follows the basic trend of power-function curves shown in Fig.4, we can say that they have nearly equal perceptual quality. We take it as an implicit constraint to our optimal problem (4).

Before we solve Equation (4), we need to establish our *D-Q* and *R-Q* models. For the *D-Q* relation, it is well known that for a zero-mean independent and identically distributed source, the relation of the distortion versus the uniform Qstep could be approximated as

$$D_i(Q_i) = \frac{Q_i^2}{12} \tag{6}$$

In practice, DCT coefficients are more likely to be Laplacian distributed, but in our scheme we pay more attention to the perceptual distortion and do not care too much about the accuracy of the distortion model. Thus, we adopt Equation (6) as our *D-Q* model due to its simplicity. On the other hand, base on our analysis in section 3, we take the linear R-1/Qstep model to establish the *R-Q* relationship.

Afterwards, by means of the Lagrangian optimization and based on our established relations between the distortion, rate and quantizer, equation (4) can be rewritten as an equivalent unconstrained problem with the objective function defined as

$$J(\lambda, Q^*) = \sum_{i=0}^{N} \frac{Q_i^2}{12} + \lambda \sum_{i=0}^{N} \frac{X_i}{Q_i} \tag{7}$$

Where λ is a nonnegative real number as the Lagrangian parameter. The Lagrangian optimization is sometimes referred to as a constant slope solution in the sense that it implies the optimal operating point is located where the slopes of the R-D curves of the individual coding units are the same, as shown in Fig.5. The Lagrangian parameter λ is normally referred to as the slope, which is a constant parameter for a specific video sequence.

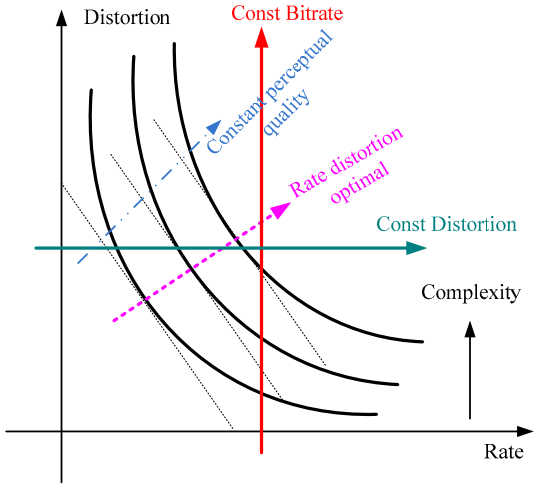

Fig.5. Rate-Distortion curves of coding units with different complexities. Several constraints such as constant bitrate, constant distortion and rate-distortion optimal are illustrated with lines in different color and shape.

If we assume that each coding unit is independent of the others, which is actually reasonable and practical when taking a GOP or a scene as the basic coding unit, we can solve (7) and get the following result.

$$Q_i = \sqrt[3]{6\lambda}(X_i)^{1/3} \tag{8}$$

We notice that Equation (8) also takes the same form of a power function, which coincides with our previous statistical correlation (3) between the Qstep and the complexity for perceptually consistent visual quality. Thus, by taking the practical coding scenarios into consideration, we extend Equation (8) to a more generic form as our picture level *quantizer-complexity* (Q-C) model:

$$Q_i = kw_i(\hat{X}_i)^p \tag{9}$$

We introduce several new parameters in Equation (9) to take additional coding techniques and strategies into account. We describe them in detail as follows:

1) The power p is a key factor to determine our model's behavior. It reflects the degree of how the complexity of a coding unit affects the corresponding Qstep. Smaller p results in less influence of the complexity. When p is set to zero, our model becomes a constant QP solution regardless of video content. While, if we set p to one, we obtain a CBR solution with constant bit rate $R=1/kw_i$. Furthermore, according to the result of RDO (8), we achieve the minimum total distortion when p equals 1/3. Finally, when p is set to 0.4-0.5 according to our perception analysis in section 3, multiple coding units with different characteristics are supposed to exhibit a similar perceptual quality. We illustrate these different control strategies in Fig.5. Considering we aim to achieve both good subjective quality and coding efficiency, we fix p=0.42 in our implementation.

2) It is noticeable that we introduce a new symbol $\hat{X}$ in our model (9), which is referred to as the *normalized complexity* for each individual picture. It adopts a weighted function:

$$\hat{X}_i = a\bar{X}_{scene} + bS_T X_i \qquad T \in I,P,B \tag{10}$$

Here $\bar{X}_{scene}$ represents the average complexity of a specific video scene. Based on the fact that each picture within a scene normally has similar spatial and temporal activities, we take this average complexity as the major part in our weighted function. Thus we set a and b to 0.7 and 0.3, respectively. In addition, due to their different coding techniques, video pictures of different type, typically I, P and B, always exhibit distinct R-D characteristics and result in complexities without the same modality. In order to eliminate their differences, we utilize the following normalize factors.

$$S_T = 1.0 \qquad T = I$$
$$S_T = \frac{N_T \sum X_I}{N_I \sum X_T} \qquad T \in P,B \tag{11}$$

3) The scale factor w_i in our model (9) can provide us an additional tool to adjust each picture's importance. One typical case is to assign different factors according to different picture types, which is due to the monotonicity property that for dependent prediction a better predictor will lead to more efficient coding. Therefore, we assign a little larger scale factor for those temporal predicated (P and B) pictures than those references (I). In our implementation we fix w_i to 1.0 for I type pictures and set w_i to 1.1 and 1.24 for P and B, respectively. More details about choosing the proper scale factors are discussed in [18].

4) As a sequence level parameter, k is the key to satisfy the constraint of total bit budget and reflects a specific level of subjective visual quality. We take equation (9) into (5) to obtain the following result.

$$k = \frac{\sum_i \frac{X_i}{w_i(\hat{X}_i)^p}}{R_{total}} \tag{12}$$

However, in practice Equation (12) disregards the following two aspects: the discreteness of admissible QPs and the buffer constraints imposed by the hypothetical reference decoder (HRD) [19]. Therefore, we propose a fast iterative search method to locate the optimal k. The search procedure consists of four steps:

a) First, k is initialized with Equation (12).

b) Next, an admissible Q is chosen according to current k and buffer status. It is later used to estimate bits based on our linear R-Q model.

$$Q_i \Leftarrow kw_i(\hat{X}_i)^p$$
$$R_{estimate} = \sum_{i=0}^{N} \frac{X_i}{Q_i} \tag{13}$$
$$\Delta R = R_{total} - R_{estimate}$$

c) Evaluate (14) to check whether the target bit budget is reached. If true, the search is stopped and returns current k. Otherwise, go to next step.

$$\left|\Delta R\right| \le \varepsilon_{threshold} \tag{14}$$

d) k is updated as equation (15), then go back to step (b) and continue a new iterative loop.

$$k_n \leftarrow k_{n-1}\left(1-\frac{\Delta R}{R_{total}}\right) \tag{15}$$

Since k is updated adaptively according to the distance to the target bit budget, it is an advantage that our search procedure usually converges quickly with ignorable computation overhead.

B. First pass encoding

The main purpose of our first pass encoding is to obtain the spatial temporal complexity of each individual picture. Based on our analysis in section 3, the complexity of a picture is considered to be a constant characteristic irrespective of the coding bit-rate and Qstep. Therefore, we gain an additional advantage that we can adopt any practical rate control method in our first pass. In contrast, previous algorithms always rely on specific coding strategies in their first pass, such as CBR in [1] and constant QP in [7]. Besides, in our first encoding pass we implement both scene change detection and dynamic GOP size for better coding efficiency.

C. Buffer constrain protection

The VBR encoding system is required to prevent buffer underflow according to the HRD [19]. Thus, we introduce a lower bound *LB* in our implementation, which is similar to the previous work [14].

$$LB_n = LB_{n-1} + \frac{u}{f_r} - b_{n-1} + \delta \tag{16}$$

Where u is the target bit rate, f_r is the frame rate, LB_n is the lower bound bits of picture n, b_{n-1} is the actual bits of picture $n-1$, and δ denotes a margin. During our iterative search and the actual VBR compression, we make use of *LB* to constraint our expected target bits in order to generate compliant bit streams.

D. Run-time bit production

It is inevitable for any bit-production model to have mismatch with actual outputs. Therefore, when performing the actual VBR encoding in the second pass, we monitor the accumulated bit production error and allow the target QP to fluctuate in a small range, as shown in the following equations

$$\begin{aligned}\Delta_n &= \Delta_{n-1} + b_{n-1} - b_{n-1,ideal} \\ Q_n &= \left(\frac{1}{1-k\Delta_n}\right) Q_{n,ideal}\end{aligned} \tag{17}$$

Where b_{n-1} is the actual bits of picture $n-1$, k is a toleration factor of bit-rate fluctuation, which is similar to [1].

E. Overall algorithm and second pass encoding

The coding efficiency of a VBR algorithm largely depends on two aspects: 1) accurate R-Q relations for each coding unit 2) an effective bit allocation strategy on the entire video sequence. To be more specific for our algorithm, we adopt the linear R-1/Qsetp model to approximate the real R-D characteristics and employ the Q-Complexity model to implicitly perform a sequence-level bit allocation for perceptually consistent visual quality. The main steps of our proposed VBR algorithm are described as follows:

1) Obtain each picture's spatial temporal complexity from our first coding pass and establish the R-Q function for each individual picture.
2) Detect the boundary of each scene and obtain its average complexity. Then, calculate each picture's normalized complexity according to our weighted functions.
3) Perform the iterative search for the proper sequence level parameter k to reach the total bit budget and update Qstep of each picture by our Q-Complexity model.
4) Run the second pass encoding with QPs obtained from previous step and perform the run-time bit production adjustment to generate the target VBR bit stream.

In conclusion, compared with previous work for constant QP or constant PSNR, our proposed rate control algorithm aims to generate the VBR bit-stream with perceptual consistent visual quality. We achieve our aim by mans of the novel Q-Complexity model and the fast iterative search technique. Our method is conceptually simple with very low computational complexity. The experiments in the next section demonstrate its efficiency. Furthermore, in this work we focus our attention on the picture-level rate control, but actually any macroblock level method can be further integrated to improve its performance. Visual mask and other perception optimized mechanisms can be further adopted to improve the perceptual quality, which is left to our future work.

V. EXPERIMENTAL RESULTS

To verify the effectiveness of our proposed algorithm, we implement the rate control in our H.264 encoder. We chose three test video clips with different resolutions to demonstrate its efficiency and robustness. Each clip consists of multiple scenes with different spatial and temporal activities. We compare our algorithm against the CBR and constant QP methods. We choose Yin's algorithm [15] as our CBR rate control for its accurate bit rate control and high coding efficiency. We also implement Westerink's two-pass VBR algorithm [1] to evaluate its coding performance, which is one of the few explicitly targeting on constant perceptual quality. In this paper we refer to Westerink's algorithm as Q-VBR (Qstep based) and our proposed as C-VBR (complexity-based).

We give more details about our experiments as follows. The CIF (352x288) video clip is a composite video including all six scenes used by our perception experiment in section 3. The SD clip (704x576) consists of four typical MPEG test sequence

segments (*crews*, *harbor*, *ice* and *soccer*). To evaluate HD coding performance, we extract a 300-frame-long video clip from a commercial HD (1920x1088) movie source with characteristics of dark areas, details textures and high motions. In our experiments, we coded all clips at 30fps, with GOP structure of "IBBP", GOP size of 16 and 4 reference frames. The loop filter and CABAC is also enabled. We set the search range of motion estimation to 32, 64 and 128 for the CIF, SD and HD clips, respectively. Besides the objective distortion metric of PSNR, we employ user studies to evaluate the subjective quality. We invited 8 colleagues, all with working experience in the video processing field, to directly score the resultant video streams. The score is in the range of 1 to 5, with 1 as the unacceptable worst and 5 as the perfect. Each compressed video stream is scored respectively without informing viewers any information about the bit rate and the rate control method. We average the scores given by each participant to obtain the mean opinion score (MOS), as shown in the last column in Table II. During our test, all video streams with resolutions of SD and HD are displayed on a 42' full-HD television to make the experiments close to the practical applications. The results are listed in Table II.

TABLE II : BITRATE AND QUALITY RESULTS

Video clips	Encoder	Actual Bitrate (kbps)	PSNR (db)	MOS (1-5)
Clip_CIF (500Kbps)	CBR	504.40	34.15	2.1
	Const-Q	504.32	34.27	2.9
	Q-VBR	502.57	34.52	3.3
	C-VBR	503.41	34.70	3.6
Clip_SD (2Mbps)	CBR	1997.23	37.31	2.6
	Const-Q	2001.95	37.23	3.4
	Q-VBR	2004.11	37.41	3.4
	C-VBR	2003.24	37.43	3.5
Clip_HD (8Mbps)	CBR	8016.07	42.18	3.0
	Const-Q	8083.51	42.26	3.3
	Q-VBR	8111.93	42.22	3.1
	C-VBR	8075.50	42.29	3.5

Table II demonstrates the superiority of our complexity based VBR algorithm. It got the highest MOS for the most appealing visual quality. From the perspective of PSNR, our method still yields a slight improvement due to its RDO consideration. The results from different resolutions and bit rates also prove its reliability for various applications. On the other hand, Table II illustrates the accuracy of our bit rate control. By means of the iterative search technique and the run-time bit production adjustment, our encoder successfully generates bit streams that are all able to fit the target storage capacity with a mismatch of less than 2%.

To further understand different visual quality of our four test algorithms, we plot their PSNR curves in Fig. 6. Although these curves cannot represent the subjective quality accurately, they do reflect the basic trend of quality fluctuation. The CBR curves vary greatly according to video content due to the constraint of constant bit rate. It spends too many bits on easy scenes and achieves overly high quality, but degrades complex scenes due to insufficient bit budget. According to previous work [1], the subjective visual quality of an entire video sequence is judged by the minimum quality across the whole sequence. Consequently, the subjective quality from the CBR rate control is always judged as the worst, as shown in Table II.

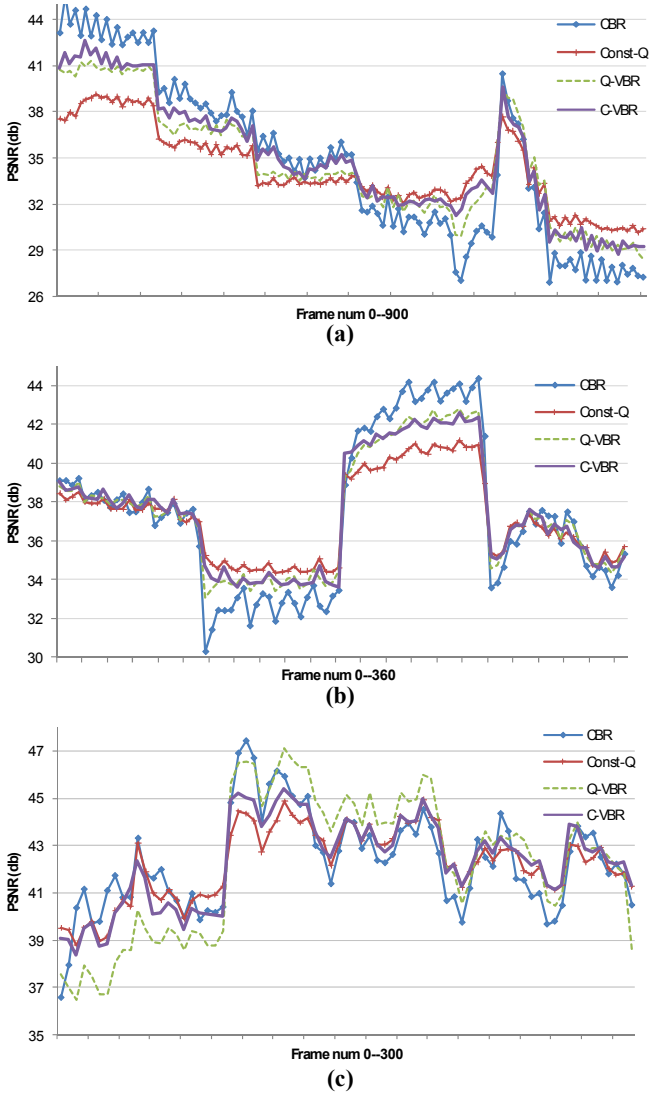

Fig.6. PSNR curves of four different rate control algorithms: CBR, Const-Q, Q-VBR, and our proposed C-VBR (a) CIF clip encoded at 500Kbps (b) SD clip encoded at 2Mbps (c) HD clip encoded at 8 Mbps

In contrast, all three other algorithms fall into the category of VBR. Their PSNR curves fluctuate much gently as shown in Fig. 6, which can be explained by the VBR coding principle of maximizing consistent video quality by taking bits from easy scenes and spending them on difficult ones. However, how to properly move bits among different scenes is crucial for rate control design. The Const-Q algorithm is relatively aggressive, which means it always takes too many bits from easy scenes thus greatly degrades their visual quality. Although these easy scenes may have a little higher PSNR

than the overall average, their subjective quality is relatively lower because that their low spatial temporal complexities make them hard to hide annoying artifacts or other distortion from the viewers. The other methods based on constant PSNR also suffer from the same problem. It could be even worse for this case since low complex scenes have to sacrifice too much quality to reach sequence level consistent PSNR. When talking about the Q-VBR method, we found that its PSNR fluctuation is a little higher than our C-VBR methods, which can be explained by its rate control strategy of only relying on the QPs from the first CBR pass. Comparing with our normalized complexity, the QPs of CBR always fluctuate in a relatively large range. To be more specific and detailed, we choose the CIF clip as an example to give an insight into each individual scene, as shown in Table III. Comparing with the Q-VBR method [1], our method has a lower PSNR variance, indicating a relatively smooth visual quality.

TABLE III: PSNR AND ITS VARIANCE FOR DIFFERENT RC METHODS

scene	PSNR(db) & (∂^2 of PSNR)			
	CBR	Const-Q	Q-VBR	C-VBR
Mother	43.41 (1.37)	38.51 (0.31)	41.33 (0.37)	40.78 (0.12)
Foreman	38.47 (1.32)	35.75 (0.24)	37.38 (0.57)	36.98 (0.25)
Pairs	35.16 (0.86)	33.59 (0.13)	34.67 (0.38)	33.89 (0.12)
Stefen	31.26 (1.29)	32.73 (0.27)	32.25 (0.67)	32.38 (0.32)
Football	32.49 (16.57)	34.45 (3.67)	33.65 (8.51)	34.64 (6.12)
Mobile	27.80 (0.93)	30.62 (0.28)	29.58(0.57)	29.65 (0.46)

VI. CONCLUSION

In this paper, we proposed a practical two-pass VBR rate control algorithm for perceptually consistent video quality. It is different from previous work aiming to constant objective distortion and it is especially suitable for DVD-like video storage applications. In our method we employ the spatial temporal complexity to evaluate the subjective visual quality. We further incorporate the statistical quality relation into our RDO based Q-Complexity model for the second pass VBR encoding. By means of the fast iterative search technique and the run-time bit production control mechanism, we are able to accurately fit the target storage capacity while maintaining a consistent perceived quality. Our experimental results from various test clips demonstrate its efficiency and reliability, indicating a noticeable improvement on both objective and subjective quality. There are several avenues for future work. We can incorporate the macro-block refinement into our method. Feature maps and saliency maps are also able to be merged into our system to achieve better perceptual visual quality.

ACKNOWLEDGMENT

We thank Lihua Zhu, Long Xu, Yadong Mu for their contributions and helpful comments to this work. Special thanks to all the colleagues participating in our experiments.

REFERENCES

[1] P.H.Westerink, R. Rajagopalan and C.A. Gonzales, "Two-pass MPEG-2 Variable-bit-rate encoding," *IBM Journal of Research and Development*, vol.43, no.4, July 1999

[2] Y. Yu, J. Zhou, Y. Wang, and C. W. Chen, "A novel two pass VBR coding algorithm for fixed-size storage application," *IEEE Transactions on Circuits and Systems for Video Technology.*, vol. 11, no. 3, pp. 345-356, March 2001

[3] D. Bagni, B. Biffi and R. Ramalho, "A constant-quality, single-pass VBR control for DVD recoders," *IEEE Transactions on Consumer Electronics.*, vol. 49, no. 3, Aug. 2003

[4] L. Texixeira and H. Ribeiro, "Analysis of a two step MPEG video system," *Proceedings of IEEE International Conference on Image Processing (ICIP'97)*, vol.1, pp.350-352, CA, October 26-29, 1997

[5] Byung Cheol Song, Kang Wook Chun, "A one-pass variable video coding for storage media," *IEEE Transactons on Consumer Electronics.*, vol. 49, pp689-692, Aug. 2003.

[6] K. Wang, J. W. Woods, "MPEG motion picture coding with long-term constraint on distortion variation," *IEEE Transactions on Circuits and System for Video Technology.*, vol. 18, no.3, March 2008

[7] H. B. Yin, X. Z. Fang, L. Chen, J. Hou, "A practical consistent-quality two-pass VBR video coding algorithm for digital storage application,", *IEEE Transactions on Consumer Electronics.*, vol. 50, no. 4, Nov, 2004

[8] A. Jagmohan and K. Ratakonda, "MPEG-4 one-pass VBR rate control for digital storage", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 5, May 2003

[9] L. Overmeire, L. Nachtergaele, F. Verdicchio, J. Barbarien, P. Schelkens, "Constant quality video coding using video content analysis", Signal processing: Image Communications vol. 20, pp 343-369, 2005

[10] B. Xie, W. Zeng, "A sequence-based rate control framework for consistent quality real-time video", *IEEE Transactions on Circuits and Systems for Video Technology*, vol.16, no.1, January 2006

[11] E. H. Adelson, J. R. Bergen, "Spatiotemporal energy models perception of motion", *Journal of the Optical Society of America*, vol.2, 1985

[12] N. Jayant, J. Johnston, R. Safranek, "Signal compression based on models of human perception", *Proceedings of the IEEE,* 1993, 81(10):1385-1422

[13] J. G. Robson, "Spatial and temporal contrast sensitivity functions of the visual system", *Journal of the Optical Society of America*, vol.59, pp. 1141-1142, 1966

[14] Li zhengguo, Gao Wen, Pang keng " Adaptive rate control with HRD consideration", JVT-H014, 8th meeting, Geneva ,2003

[15] P. Yin, J, Boyce, "A new rate control scheme for H.264 video coding" *Proceedings of IEEE International Conference on Image Processing(ICIP 2004), pp449-452*

[16] Y. Liu, Z. G. Li, " A novel rate control scheme for low delay video communication of H.264/AVC standard", *IEEE Transactions on Circuits and Systems for Video Technology*, vol.17, no.1, January 2007

[17] A.Ortega, "Optimal bit allocation under multiple rate constraints", *In proceedings of DCC*, pp.349-358, Mar,1996

[18] "Joint Scalable Video Model JSVM 1", Joint Video Team (JVT) 14th meeting: Hong-Kong, CN, 17-21 January, 2005, document JVT-N021.

[19] "Draft ITU-T recommenda-tion and final draft international standard of joint video specification", *ITU-T Rec. H.264 --ISO/IEC 14496-10 AVC*

Bo. Han received M.S degree in computer science from Peking University in 2004. He is currently a Ph.D. candidate in the Institute of Computer Science and Technology, Peking University, Beijing, China. His research interests include image and video coding and processing, real-time rendering and general purpose computation on graphics hardware.

Bingfeng. Zhou was born in 1963, Ph.D., professor, Ph.D. supervisor. His research interests include color image processing, multimedia system and image special effects, digital image halftone, image based rendering and modeling, virtual reality, and computer vision.