BUFFER-CONSTRAINED R-D OPTIMIZED RATE CONTROL FOR VIDEO CODING

Lifeng Zhao and C.-C. Jay Kuo

Integrated Media Systems Center and Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089-2564
{lifengzh, cckuo}@sipi.usc.edu

ABSTRACT

Buffer-constrained R-D optimized rate control for video coding is investigated in this work. A frame level bit allocation is first presented based on a model of the relationship between the rate (R) and nonzero (NZ) coefficients. With the modelled R-NZ relationship, a quality feedback scheme is proposed to generate VBV (Video Buffer Verifier) compliant bitstream with assured video quality. Then, a R-D optimized macroblock level rate control is described by jointly selecting the quantization parameter and the coding mode of macroblocks in I, B and P pictures for both progressive and interlaced video. To avoid the irregularly large MV or one single isolated coefficient, we extend the set of coding modes of MB by including zero MV and zero texture bits as two more candidates. Finally, fast heuristics are developed to reduce the computational complexity of R-D data generation and the Viterbi algorithm (VA) in R-D optimization, which achieves coding results close to the optimal one at a much lower computational cost.

1. INTRODUCTION

One fundamental problem in the encoder design is the selection of coding parameters to maximize visual quality under constraints imposed by the computational complexity, delay, bandwidth and/or loss factors. For a buffer-constrained CBR coding, the optimal encoder bit allocation problem was studied in [1] with a forward dynamical programming technique known as the Viterbi Algorithm (VA) over a discrete set of quantizers. Instead of optimizing bit allocation among frames, bit allocation can also be optimized among macro-blocks by selecting different quantization steps and/or coding modes for the P frame coding in the H.263 standard [2, 4, 5]. Wiegand *et al.* [4] proposed a method to select one from four possible modes, *i.e.* uncoded, intra, inter and inter-4V (Inter MB with four motion vectors), for the coding of MBs in a P frame to optimize the R-D tradeoff. A joint coding-mode and quantization-step selection method was considered in [2, 5] to encode the P frame with the R-D optimization. To reduce the computational complexity, Mukherjee *et al.* [2] proposed the M-best search scheme, in which the M least-cost paths are retained as survivors at each state in a trellis and carried over to the next step. Schuster *et al.* [5] restricted the range of quantization parameters to be between 8 and 12 for a speed-up.

In this work, a frame level bit allocation is first presented in Sec. 2 based on the model of the relationship between the rate (R) and nonzero (NZ) coefficients. With the modelled R-NZ relationship, a quality feedback scheme is proposed to generate VBV (Video Buffer Verifier) compliant bitstream with assured video quality. Then, a R-D optimized macroblock level rate control is proposed in Sec. 3 by jointly selecting the quantization parameter and the coding mode of macroblocks in I, B and P pictures for both progressive and interlaced video. Furthermore, fast heuristics are developed to reduce the computational complexity of R-D data generation and the Viterbi algorithm (VA) in R-D optimization. Experimental results are shown in Sec. 4, where the coding results of the simplified algorithm are demonstrated to be close to the optimal one at a much lower computational cost.

2. PROPOSED FRAME-LEVEL BIT ALLOCATION

In this section, we propose a frame layer bit allocation scheme that targets at real time CBR video encoding with a strict buffer constraint. With the quality estimation and feedback, we can mitigate the effect of inaccurate estimation of frame complexity. Moreover, we can use the VBV buffer adjustment to handle temporally difficult pictures.

2.1. Non-Zero (NZ) Based R-Q Model

If the rate control scheme can estimate target QP accurately based on the given bit target, the rate control task will become easy. However, most previous work that performs a direct model of the relationship between the rate and QP does not work well. It was observed by He *et al.* [3] that there is a strong correlation between the rate and the percentage of non-zero coefficients. They claimed that the number of bits spent for a coding unit is proportional to the number of non-zero coefficients in this unit. In other words, the average bits spent for a single non-zero (NZ) coefficient is constant, *i.e.*

$$R = C \times N_{NZ}. \qquad (1)$$

Consequently, they suggested to perform the histogram analysis of non-zero coefficients at each QP. They also assumed that ratio C is the same as the previous frame. Hence, based on given frame target, we only need one table lookup operation to obtain the target QP.

By examining this problem in detail, we have the following two observations. Fist, since DCT coefficients are run-length coded, if there are a smaller number of non-zero coefficients, there are better chances to have a longer run. For example, when QP is changed from 1 to 31, the run value of each non-zero coefficient potentially becomes larger while the average level becomes smaller. These two factors have an opposite effect on the final bit rate spent on texture coding. Therefore, one natural question is that whether these two factors cancel each other. If it is valid, we should get the exact R-NZ relationship as given by (1). Otherwise, the bit cost of each

NZ coefficients should vary with QP. Second, (1) only estimates bits used by the texture part (to be more specific, bits spent for the coding of DCT coefficients). However, the I frame consists of bits for the syntax header and textures, P and B frames also consists of bits for motion vectors. To estimate the bit rate for the entire frame at different QP, the syntax header of I, P, B frames and motion vectors of P and B frames should also be taken into account. Clearly, the number of syntax bits such as the MB-coded is changing with QP. Although bits for motion vectors are not varying with QP, the MB type may change for MB with the zero motion vector since it becomes the skipped MB when all coefficients become zero.

As a result of the above two observations, instead of modelling the relationship between texture bits and the number of non-zero coefficients, we attempt to model the relationship between the bits used for the syntax header plus the texture part and the number of NZ coefficients below. Thus, the derived model can be used to estimate the bits required by the I frame directly, while the required bits for the P and B frames can be estimated by adding the pre-calculated motion vector bits for inter frames.

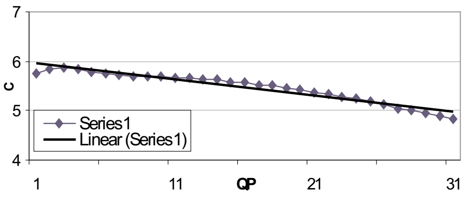

Fig. 1. The average number of bits per non-zero coefficient in intra frames for the “Football” sequence.

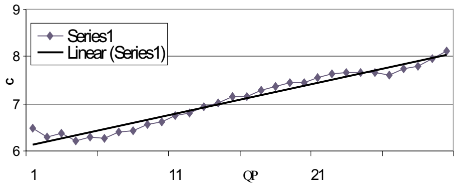

Fig. 2. The average number of bits per non-zero coefficient in inter frames for the “Tempete” sequence.

By performing extensive simulation on different contents, we observe that that the average cost per non-zero coefficients is linearly decreasing as QP increases from 1 to 31 for intra frames. This fact is shown in Fig. 1 which is also typical fig. obtained from our extensive experiments. For inter frames, it is supported by extensive experiments that if the number of non-zero coefficients is larger than two times of the number of MB in one frame, the linear relationship between cost (C) per NZ and QP is almost certain. If the number of non-zero coefficients is too small, then the bits spent for the skip frame and the MB header part become dominant and the linear relationship of C and QP becomes less obvious. From the fig., we can see the relationship of C and QP match very well with linear model especially for QP larger than 5. It is also observed that the average cost per non-zero coefficient does not have significant difference for different video contents. It is almost always between 4 and 8 for intra frame and between 5 and 12 for inter frame. Being different from the I-frame case, C is linearly increasing as QP increases from 1 to 31 in the inter-frame. The average cost C per non-zero coefficient can be written as

$$C = a \times QP + b, \quad (2)$$

where parameter a is positive for inter frames and negative for intra frames.

2.2. Description of Frame-level Bit Allocation

The proposed buffer constrained real-time CBR video rate control scheme is shown in Fig.3 and described below.

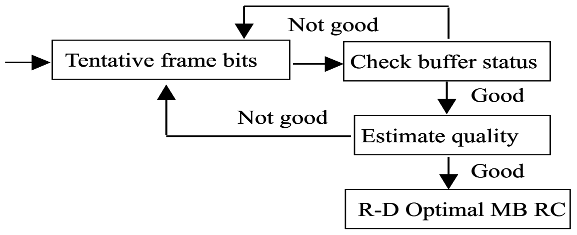

Fig. 3. The overview of the proposed rate control scheme

Step 1: Initial frame rate determination

Based on the given GOP level bit allocation, we first choose a target frame bit allocation similar to TM5 based on the frame complexity defined as $X_i = R_i \times QP_i$ (*i.e.* if the frame complexity is thought to be QP invariant, it is actually the first order R-Q model). Furthermore, let us assume that the target frame has the same complexity as the preceding frames of the same frame type. Based on these assumptions, it is straightforward to obtain the bit allocation formula which is the same as that in TM5.

Step 2: Frame rate modification via buffer validity check

The obtained bit allocation from the first step may or may not be allowed by the VBV buffer. We can check the validity of this bit allocation scheme with 3, where B_{max} is the maximal buffer size, C is the channel rate, T is the frame duration and B_i is the buffer fullness before encoding the frame i. If it is allowed, go to step 3. Otherwise, we choose the value closest to the target and allowed by the VBV buffer.

$$T * C - B_i < R_i < B_{max} + T * C - B_i \quad (3)$$

Step 3: Quality estimation

One sliding window is preserved to measure the average quality, with which we will compare the quality of the current frame k. Suppose the QP’s of previous W frames are QP_{k-w}, QP_{k-w+1}, $\cdots$, QP_{k-1}, and their mean is equal to m_{qp} and their variance is σ_{qp}. Two thresholds are defined such that the lower bound is $QP_{dn} = m_{qp} - 2\sigma_{qp}$ and the upper bound is $QP_{up} = m_{qp} + 2\sigma_{qp}$. Based on the R-NZ model discussed above, we can estimate the average QP for I frame directly. By supposing standard mode decision (i.e., MV bits can be calculated independent of QP), similarly, average QP for B and P frames can be also estimated for given bit budget. If $QP_k \in [QP_{dn}, QP_{up}]$, the task is done and we can proceed to the next frame. Otherwise, it means that the current bit allocation scheme is either more than sufficient (with $QP_k < QP_{dn}$) or less than the minimal rate (with $QP_k > QP_{up}$). To determine QP_k, we may solve the following minimization problem

$$\overset{argmin}{QP_k}((B_{des} - B_{k-1}) + \lambda * (QP_k - m_{qp})), \quad (4)$$

where λ is the weighting factor between the buffer discrepancy and the quality smoothness. It is set to $B_{des}/15$ in our implementation.
Step 4: Final frame level bit calculation

After getting QP_k, since all R-D data are available, we can easily calculate R_k corresponding to QP_k. Then, we use R_k as the bit budget of the macroblock level rate control, and determine the best QP and the mode for this MB as described in section 3.

3. PROPOSED MB LEVEL RATE CONTROL

In this section, we focus on two choices in MB level rate control: (1) the selection of the quantization parameter and (2) the selection of the optimal coding mode. Since they are closely related to the standard syntax, they should be handled separately for different standards. In the following, we will take MPEG-4 as examples, and the idea presented can be readily applied to MPEG-2, H.263 and H.26L standards.

3.1. MB Dpendency and Its Simplification

R-D data generation, which is computationally expensive, is highly dependent on the dependency among MBs. The MB dependency can come from either the standard syntax requirement or the algorithm itself. For example, QP can only vary in the range of (-2,2) from the QP of the previous MB, and only $INTER-1MV$ MB or $INTRA$ MB can have a different QP from the previous MB since $Dquant$ (QP difference of MB) is not allowed in other modes defined by the syntax. The dependency imposed by algorithm includes the AC and DC prediction and predictive encoding of motion vectors. Besides, if the OBMC technique is adopt, both rate and distortion of a MB depends on its left, top and top right neighbors.

As a result of these constraints, the rate and distortion of the target MB i is determined by considering its own property as well as the coding options (χ_{i-l}, χ_{i-t} and χ_{i-tr}) of its three neighboring MBs (*i.e.* left MB, top MB and top right MB). If the Lagrangian approach is adopted, the resulting cost function will be the function with several variables

$$J_i(\chi_i, \lambda | \chi_{i,l}, \chi_{i,t}, \chi_{i,tr}) \quad = \quad R_i(\chi_i | \chi_{i,l}, \chi_{i,t}, \chi_{i,tr}) + \lambda * D_i(\chi_i | \chi_{i,l}, \chi_{i,t}, \chi_{i,tr}), \quad (5)$$

where χ_i is the 2-D feature vector set that consists of quantization parameter QP and coding mode M such that $\chi = QP \times M$.

Without decomposing the rate and distortion terms, it will be extremely difficult to find the exact solution to ${}^{argmin}_{\chi} \sum J_i$ due to the coupled dependency among MBs even with VA. To make this problem tractable, we simplify the dependency using the following assumptions.

First, for the AC prediction in the I frame, we suppose that the left, top MB has the same QP as current MB. Second, in the P and B frames, there is a rare chance for the current and its left, top, top-right neighbors to be all intra-coded. Hence, the AC prediction can be disabled in the P and B frames. Third, in the P and B frames, to determine the coding modes of the neighboring left, top and top right MBs, the concept of "causal optimality" is introduced, by which we take the optimal mode determined by current Lagrange parameter and QP of previous frame as

$${}^{argmin}_{M} J_i = R_i(QP_{prev}, M) + \lambda * D_i(QP_{prev}, M) \quad (6)$$

With the above assumptions, we restrict the dependency of the current MB i to its immediately preceding MB only, which can be efficiently handled with VA. Hence, (5) can be rewritten as

$$J_i(\chi_i, \lambda | \chi_{i-1}) = R_i(\chi_i | \chi_{i-1}) + \lambda D_i(\chi_i | \chi_{i-1}), \quad (7)$$

where $\chi = QP \times M$, QP is equal to $[1, ..., 31]$ by definition and the mode $M = [m_1, ..., m_n]$ of the MB depends on the frame type.

In MPEG-4, the possible coding modes of I, B, and P frames are stated below. For the I frame, only the intra-coded MB is possible. For the P frame, MB can be coded as the skip, intra coded, inter coded with 1 MV, inter coded with 4 MV, inter coded with 2 MV (field prediction), and the special case that MV is zero. For the B frame, MB can be encoded as the direct mode, backward-1MV, forward-1MV, interpolated-2MV, backward-2MV, forward-2MV, and interpolated-4MV.

Moreover, for inter-coded MB, it may consist of motion vectors and DCT data. It is the flexibility of the encoder to decide whether to encode DCT coefficients even though there are non-zero coefficients. By skipping high cost non-zero coefficients with a small increase in distortion, we may get a better R-D performance than directly encoding these expensive coefficients. Hence, for the inter-MB mode in B and P frames, both options (with or without DCT coefficients) are examined. Moreover, the inclusion of these modes does not incur additional complexity in the R-D data generation.

The complexity of R-D data generation as shown in (7) can be further simplified to make it grow linearly instead of exponentially with the dependency depth. Generally speaking, the rate of MB i includes the rate for MV and the texture. The rate of MV is invariant to the MB prediction type or the quantization parameter. Hence, only one fixed lookup table is needed to generate the MV rate. The rate of the MB texture can be decomposed to terms $R_{i,start\_code}$ (constant), $R_{i,AC}$, (depending on its own coding option), R_{i,MB_type}, $R_{i,DC}$, and $R_{i,CBPC}$ (depending on QP difference of MBs).

To summarize, the rate of the MB texture can be decomposed into one term, which is determined by its own coding option χ_i, and several lookup tables with predefined entries. Hence, R-D data generation can be performed independently while MB dependency can be addressed with the Viterbi coding procedure. It reduces the complexity of R-D data generation significantly. That is, the complexity grows linearly instead of exponentially with the number of MBs.

3.2. Simplified MB-Level Rate Control

The complexity of the R-D optimized rate control comes from two parts: (1) R-D data generation and (2) the Viterbi algorithm. To reduce the complexity, instead of generating the R-D data corresponding to each QP and each mode, we may generate the R-D data for the most probable M modes and N QPs. On the other hand, we still should keep the coding efficiency of the fully optimized scheme as much as possible.

Based on the R-NZ model described in Sec. 2, the average QP can be estimated as QP_{avg}, then we limit the QP range for both R-D data generation and subsequent Viterbi coding as

$$QP \in (\lfloor QP_{avg} + 0.5 \rfloor - 2, \lfloor QP_{avg} + 0.5 \rfloor + 2) \quad (8)$$

Only 5 R-D data points are needed in the above range. If the complexity of the R-D data generation and Viterbi coding grows linearly with the QP range, it represents a complexity reduction factor of 31/5*31/5 = 38.4 for R-D data generation and Viterbi coding.

To select the K most probable coding modes for the current MB, we calculate

$$J_{Mi} = R_{Mi} + \lambda_{prev} \times D_{Mi},$$

and the mode M_i with the K smallest J_{Mi} are selected for that MB. Here, λ_{prev} is the Larangian parameter selected by the previous frame. Indeed, it can be further simplified so that only modes with the K smallest distortions (in the SAD or MSE sense) are preserved and their rates and distortions are calculated. Then, both R-D data generation and its subsequent Viterbi coding are simplified by factor of N/K.

3.3. Summary of the Algorithm

1: After the DCT transform, calculate the histogram $hist(QP)$ of NZ coefficients at each QP, if B or P frame, selected the mode via standard mode decision.
2: Obtain $C(5)$ and $C(10)$ as two sampling points with pseudo coding and find the target QP based on the given bit target B_{tgt}

$$\underset{QP}{argmin}\left(\frac{B_{tgt}}{C(QP)} - hist(QP)\right). \qquad (9)$$

3: Perform the R-D data generation from $QP_{tgt} - 2$ to $QP_{tgt} + 2$ for every mode.
4: If P or B frame, based on λ_{prev}, prune the possible coding modes to K (*e.g.* 3) modes
5: Perform Viterbi coding in the limited QP range for these K modes.

4. EXPERIMENTAL RESULTS

In this section, we examine the performance of the R-D optimized MPEG-2 and MPEG-4 encoders with MPEG-2 and MPEG-4 standard reference software.

In Fig. 4, we show the PSNR comparison of our encoder with the reference encoder for sequence ”mobile and Calendar” at 4 Mbps. Our R-D optimized macro-block level rate control scheme has a coding gain of about 1.9 dB. The simplified MPEG-2 encoder utilizes the information of the number of non-zero coefficients to limit the possible QP range to 5 and then select 3 best modes for R-D data generation and Viterbi coding. The performance gap between the fully R-D optimized version and the simplified version is very small. The PSNR result of the simplified R-D optimized MPEG-2 encoder is compared with that of the fully optimized scheme is also shown in Fig. 4. We see that the average PSNR degradation is about 0.15dB.

We compare our R-D optimized MPEG-4 codec with the MPEG-4 reference codec with TM5 rate control in Fig. 5. The GOP structure is IPPP. The PSNR comparison is presented for the 15fps Foreman CIF sequence at 192kbps. It is found that more than 1 dB gain is achieved with the proposed R-D optimized MB level rate control. The comparison of the simplified RD optimization versus the R-D optimization is also presented in Fig.5. As expected, the simplified one performs very closely to the RD optimal one in both cases (less than 0.05dB difference).

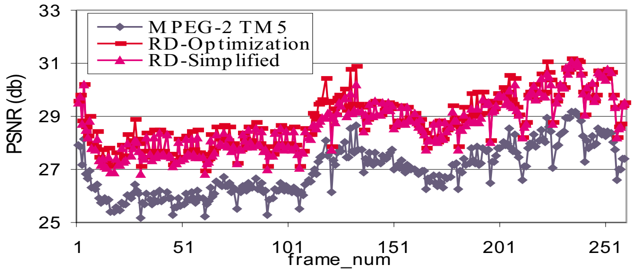

Fig. 4. The PSNR comparison of the TM5 reference encoder with the proposed R-D optimized encoder and simplified RD optimized encoder for the Football sequence.

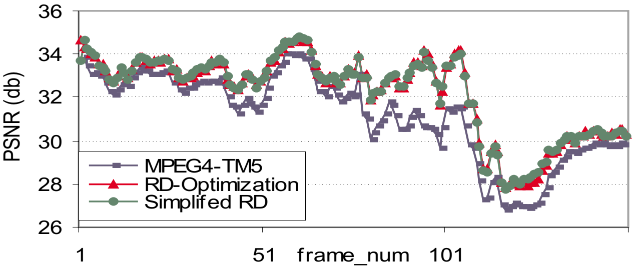

Fig. 5. The PSNR comparison of the reference TM5, the simplified R-D rate control method, and the R-D optimized rate control method for the 15fps Foreman sequence at 192kbps, where $PSNR_{TM5-avg} = 31.15dB$, $PSNR_{smp-rd-avg} = 32.21dB$ and $PSNR_{RD-avg} = 32.22db$.

5. REFERENCES

[1] A. Ortega, K. Ramchandran, and M. Vetterli,“Optimal trellis-based buffered compression and fast approximation”, in *IEEE Trans. Image Proc.*, vol.3, pp.26-40, Jan. 1994.

[2] D. Mukherjee, and S.K. Mitra, “Combined mode selection and macroblock quantization step adaptation for the H.263 video encoder”, in *Proc. of International Conf. Image Proc.*, pp. 37-40, Sept. 1997, Santa Barbara, USA.

[3] Z. He and S.K. Mitra, “A unified rate-distortion analysis framework for transform coding”, in *IEEE Trans. CSVT.*, vol. 11, No. 12, pp. 1221-1236, Dec. 2001.

[4] T. Wiegand, M. Lightstone, D. Mukherjee, T. G. Campbell and S.K.Mitra, “Rate-distortion optimized mode selection for very low bit rate video coding and the emerging H.263 standard”, in *IEEE Trans. CSVT.*, vol. 6, No. 2, pp. 182-190, Apr. 1996.

[5] G.M.Schuster and A.K. Katsaggelos, “A theory for the optimal bit allocation between displacement vector field and displaced frame difference”, in *IEEE Journal on Selected Areas in Comm.*, vol. 15, No. 9, pp. 1739-1751, Dec. 1997.