The Impact of Rate Control Algorithms on Video Codec Hardware Design

Sheu-Chih Cheng Hsueh-Ming Hang

Department of Electronics Engineering
National Chiao-Tung University,
Hsinchu, Taiwan 300, ROC.

abstract

This paper presents an evaluation of rate control algorithms from a system-level VLSI design viewpoint. Rate control in video coding has a significant influence on the coded bits and image quality. Many rate control algorithms have been proposed mainly focusing on the optimal rate-distortion performance without considering their overall performance on the VLSI implementation. However, a system-level designer should design an algorithm not only good in performance but also good in implementation. In this paper, three different types of popular rate control algorithms have been analyzed based on their picture quality, the internal buffer size and the hardware cost. The methodology and results presented here should provide useful guidelines for selecting an appropriate rate control algorithm for system-level VLSI design.

1 Introduction

The purpose of this paper aims at studying the impact of various types of rate control algorithms on system-level VLSI design. Since chip design and layout process are time-consuming and costly, it is very desirable to be able to predict the overall system performance of a high-level algorithm before the circuit layout is fully deployed. For digital video transmitted over a bandlimited channel, rate control is one of the critical elements to determine the picture quality and compression efficiency in a video coding system. To achieve the best PSNR performance, many rate control algorithms have thus been devised to optimize the rate-distortion performance without considering the problems of high computational complexity and/or large-size internal output buffer. However, in addition to good image quality, VLSI realization is an important factor in the system design. It is thus the purpose of this study is to compare various types of rate control algorithms from the viewpoint of hardware implementation.

Based on our survey, the generic processing structure proposed in this paper is adequate in implementing many different rate control algorithms to interests. Potentially, the MPEG compressed bits produced by the VLC unit can be greater than several thousands of bits in one macroblock (several hundreds Mbps). This poses impractical requirements on both I/O bandwidth and external memory. Hence, an internal output buffer is introduced to smooth out the data transmission rate between the VLC unit and the external RAM. However, it is often neglected by the algorithm designers.

2 Rate Control Algorithms

The goal of rate control algorithm is to efficiently distribute the coded bits properly to each coded image block at a given total bits budget. Most of the compression standards allow using different quantizer stepsize (*mquant*) for the DCT coefficients in different portions of a picture. In general, the rate control algorithm consists of two operations, namely, the bit allocation and the quantizer selection. The bit allocation unit estimates the number of bits available for coding the next picture and distributes the available bits to image blocks. The quantizer selection unit measures the image content of a macroblock and decides a proper *mquant*. Thus, a rate control algorithm is needed to determine the *mquant* for each macroblock to meet the given bits budget.

In the following analysis, we assume that r_i denotes the coded bits of the i-th macroblock. The macroblock *mquant* is determined by the chosen distortion d_i, the macroblock image activity a_i, and the buffer fullness b_i. The selected *mquant* (Q_i) is calculated based on r_i and $F(d, a, b)$ which is selected by the algorithm designer. For convenience, we thus define two terms for measuring the macroblock activity: (a) MV (minimum variance), which is referred to the minimum block variance of a macroblock, and (b) SCM (Sum

0-8186-8183-7/97 $10.00

of the DCT Coefficients in a Macroblock), which is the sum of the DCT coefficients inside a macroblock without the DC coefficient. To our understanding, the existing rate control algorithms can be classified into three groups according to their *mquant* determination strategy: (1) buffer-feedback method, (2) budget planning method, and (3) optimal bit allocation in a rate-distortion sense. The following sub-sections briefly describe the operations of various rate control algorithms.

2.1 Buffer-Feedback Method

To meet the exact target bits budget, the most straightforward rate control scheme is the buffer-feedback method. An example of buffer-feedback rate control algorithm is the well-known TM5 [1]. It contains an image content measuring mechanism.

In this algorithm, the Q_i is mainly determined by the b_i, and then it is adjusted by the normalized DCT variance of the image block. The *mquant* for the i-th macroblock is thus calculated:

$$Q_i = F_{TM5}(a_i, b_i) = \frac{b_i * 31}{r} \cdot \{N\_act(MV_i)\}, \quad (1)$$

where r is the "reaction parameter" and $N\_act(SMV_i)$ is the "normalized MV".

2.2 Budget Planning Method (BP)

This method is based on the assumption the coded bit counts can be predicted from the DCT activity. A bits model is thus constructed to predict the generated coding bits for each picture and macroblock from their DCT activity and *mquant*. In this paper, a linear bits model using the "sum of the absolute value of DCT coefficients without DCT term" (SCM) measure [3]. Two piecewise linear bits models are proposed to allocate bits to every picture frame and for every macroblock separately. Thus, the *mquant* of every macroblock is calculated by

$$Q_i = F_{budget}(a_i) = \frac{r_i - C_0}{C_1 \times SCM(a_i)}, \quad (2)$$

where r_i is the macroblock target bits budget. When the bits model is accurate, the actual coded bit count is close to the predicted one and the result is satisfactory. To improve the model accuracy, we can adjust the bits model dynamically in the process of coding. In this paper, the model parameters are updated by the LMS algorithm once per macroblock/frame.

2.3 Optimal Bit Allocation Method

The purpose of optimal bit allocation algorithm is to obtain the minimum distortion under the bits budget constraint. However, many popular compression standards use dependent coding, i.e., the set of available rate-distortion (R-D) operating points depends on the particular choice *mquant* in the previous macroblock/frame. To solve the dependent coding problem, several algorithms have been proposed using the multi-level dependency tree. However, it increases the frame coding delay and requires large buffers to store the calculated distortion measures and coded bits. In this paper, we like to examine the impact of these algorithms on VLSI design. Thus, the frame target bits in this algorithm is the same as that of the bit allocation strategy in the TM5 algorithm, but the *mquant* is obtained by selecting a minimum Lagrange cost, that is,

$$min_{Q_i}[\sum_{i=1}^{N_{mb}} D_i(Q_i) + \lambda \sum_{i=1}^{N_{mb}} r_i(Q_i)], \quad (3)$$

where λ (Lagrange multiplier) is updated using the algorithm in [2].

3 Complexity Analysis and Chip Area

3.1 Complexity Analysis

In VLSI implementation, the silicon area of a rate control algorithm can be approximated by

$$A_{total} = A_{op} + A_{ibuf} + A_{ext}, \quad (4)$$

where A_{op} is the area used for processing unit, A_{ibuf} is for the on-chip output buffer, and A_{ext} is for the additional hardware requirement. A generical processing structure allows a higher degree of flexibility and it is adequate for an efficient implementation of many different rate control algorithms. Thus, the silicon area of computation unit (A_{op}) is estimated by using the statistical results of the flexible programmable architectures ($100mm^2/GOP$) for video codec [4].

In the rate control algorithm, the required number of operations is mainly determined by the quantization selection unit, listed in Table 2. In this table, the *Operator type* is the operator used for a specific algorithm, and the *Processing rate* is the number of calculations required to compute the *mquant*. The parameter λ_{itr} represents the number of iterations for the optimal bit allocation algorithm to find the best value of λ. Its value is approximately 8 from the statistics of encoding the football picture sequence.

3.2 Internal Output Buffer

Because the huge number of compressed bits generated by the VLC unit must be transmitted to the external RAM, an internal output buffer is needed to smooth out the data transferring rate between the VLC unit and the external RAM. The size of internal output buffer is chosen to handle the worst

case of bandwidth requirement in the MMU (memory management unit). From the theoretical analysis of MPEG algorithm, five units can simultaneously issue memory request signals to the MMU to access the memory bus. They are the frame recorder, the motion estimator, the DCT unit, the frame memory processor, and the VLC unit. The size of internal output buffer is mainly determined by the allowed maximum data transmission rate which is restricted by the external memory bandwidth. Our analysis shows that the allowed output data rate of internal buffer is less than $500bits/t_{mb}$ (about 20 Mbps), when the external memory has a memory access time of $50ns$, the bus width (W) is chosen to be 60, and the motion estimator uses the three step search with its own buffer [5].

To test on a difficult picture, we synthesized a so-called *Gaussian* picture sequence. The CIF-size salesman picture sequence is placed in the middle of a CCIR frame and is surrounded by the white Gaussian random noise with a variance of 500. The simulation results of the on-chip output buffer fullness (measured at per macroblock interval) In this table, MAX is the peak value and Ave_{100} is the average of the 100 largest values. It is observed that the buffer size of our rate control algorithm is about 1/4 to 1/18 of that of TM5.

3.3 Chip Area Estimation

To implement the rate control operation, the optimal bit allocation algorithm requires additional quantization and VLC units to calculate the coded bits and other circuits to calculate the distortion at different quantization stepsizes. In this paper, we use the adder and multiplier operations to implement the VLC unit, the quantization unit, and other circuits. The statistical results of the dedicated architectures ($2mm^2/GOP$) for video codec [4] are used to estimate the area of these three units. In addition, the optimal bit allocation algorithm also requires an external on-chip buffer to store the coded bits and distortion at various *mquant* values.

A list of silicon area of the critical elements for various rate control algorithms is shown in Table 3. It is interesting to see that the area of internal output buffer plays an important role in the silicon area. From Table 3, we find that the silicon area of the optimal bit allocation algorithm is approximately 100 times larger than that of the budget planning algorithm.

4 Picture Quality

To compare the picture quality, the three-step search algorithm is used to reduce the computing time; the search range is 47 for P-pictures and 15 for B-pictures. Several sequences have been tested. Limited by space, only one of them are reported here. Figs. 1 shows the PSNR performance for the CCIR *football* image sequence with a coding rate of $5Mbits/s$. Unless there is a significant disadvantage in hardware cost, the optimal rate control algorithm offers the best PSNR performance. Figure 2 shows the PSNR performance in the *Test* picture sequence without the surrounding noise region. It is clear that the budget planning algorithm outperform all the other algorithms. However, the optimal bit allocation algorithm has the best overall PSNR performance when the surrounding noise borders are included. In general, the budget planning algorithm has a much lower hardware cost and only a slightly lower objective PSNR, but its subjective quality is as good as the other algorithms if not better.

5 Conclusion

The purpose of this study is not to propose a VLSI architecture for implementing a specific rate control algorithm but to evaluate various rate control strategies from the viewpoints of both VLSI design and coding performance. Our assessment on rate control algorithm is mainly based on silicon area and image quality. A distinct feature in our study is to include the internal output buffer into the silicon area. In addition to the rate-distortion performance, we should also consider the hardware implementation issue in designing a good rate control algorithm.

We suggest a budget planing algorithm with a linear bits model to allocate the macroblock and frame bits. Our preliminary experiments show that our simple budget planning rate control algorithm has a significant advantage in hardware cost while maintaining a comparable rate-distortion performance. This analysis should be able to provide useful guidelines to the system designers in choosing a suitable high-level algorithm for VLSI implementation.

References

[1] ISO/IEC JTCI/SC29/WG11, Doc. No. 400: *Test Model 5,* Apr. 1993.

[2] W.-Y. Lee and J.-B. Ra, "Fast algorithm for optimal bit allocation in a rate distortion sense," *Electron. Leet.,* vol. 32, No. 20, Sept. 1996.

[3] W.-Y. Sun, H.-M. Hang, and C.-B. Fong, "Scene adaptive parameters selection for MPEG syntax based HDTV coding," *Int'l Workshop on HDTV '93*, Ottawa, Canada, Oct. 1993.

[4] P. Pirsch, N. Demassieux, and W. Gehrke, "VLSI architectures for video compression — a survey," *Proc. of the IEEE*, vol. 83, no. 2, Feb. 1995.

Table 1: On-chip buffer size for various CCIR picture sequences

Picture sequence \ Algorithm / Item	TM5		BP		OB	
	Max	Ave_{50}	Max	Ave_{50}	Max	Ave_{50}
Football	2469	1762	349	310	3621	2870
Flowergarden	4679	4253	416	385	58863	57718
Test picture	6839	6370	326	295	22255	20409
internal buffer	6370		385		57718	

[5] S.-C. Cheng and H.-M. Hang, "A comparison of block-matching algorithms mapped to systolic-array implementation," to appear in *IEEE Trans. CAS for Video Tech.*,

Table 2: Implementation complexity of the processing units

Algorithm(1)		Test Model 5		
Operator type eqn. (1)	Processing rate (N_{op}/s)	Q_i	mul	$2 \cdot N_{mb} \cdot f_r$
		MV	add	$2 \times 4 \cdot n \cdot N_{mb} \cdot f_r$
			mul	$4 \cdot n \cdot N_{mb} \cdot f_r$
		N_act_j	add	$2 \cdot n \cdot N_{mb} \cdot f_r$
			mul	$1 \cdot N_{mb} \cdot f_r$
Algorithm(2)		Budget planning algorithm		
Operator type eqn. (2)	Processing rate (N_{op}/s)	SCM	add	$2 \times 6 \cdot n \cdot N_{mb} \cdot f_r$
		r_i	mul	$2 \cdot N_{mb} \cdot f_r$
			log	$1 \cdot N_{mb} \cdot f_r$
		Q_i	add	$1 \cdot N_{mb} \cdot f_r$
			mul	$2 \cdot N_{mb} \cdot f_r$
		model adjust	add	$6 \cdot N_{mb} \cdot f_r$
			mul	$2 \cdot N_{mb} \cdot f_r$
Algorithm(3)		Optimal bit allocation		
Operator type eqn. (3)	Processing rate (N_{op}/s)	Q_i	add	$2 \times 32 \cdot \lambda_{itr} \cdot N_{mb} \cdot f_r$
			mul	$32 \cdot \lambda_{itr} \cdot N_{mb} \cdot f_r$
		λ	add	$2 \cdot f_r$
			mul	$3 \cdot f_r$
			log	$1 \cdot f_r$

Table 3: Estimated silicon area for various rate control algorithms

items	TM5		BP			OBA		
N_{op} (MIPS)	add	mul	add	mul	log	add	mul	lo
	20.8	10.5	30.9	0.24	0.04	20.7	10.4	0.0(
A_{op}	9.4		3.5			9.3		
A_{ibuf}	5.3		0.3			54		
N_{ext} (MIPS)	-		-			add 995	mul 1.5k	bi 69
A_{ext}	0		0			204		
A_{total}	14.73		3.79			267.3		

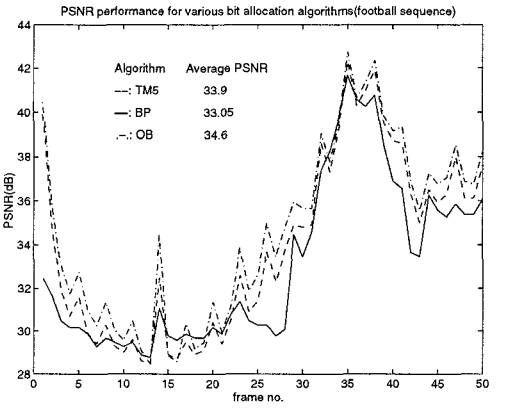

Figure 1: PSNR performance of various rate control algorithms for the CCIR *Football* picture sequence

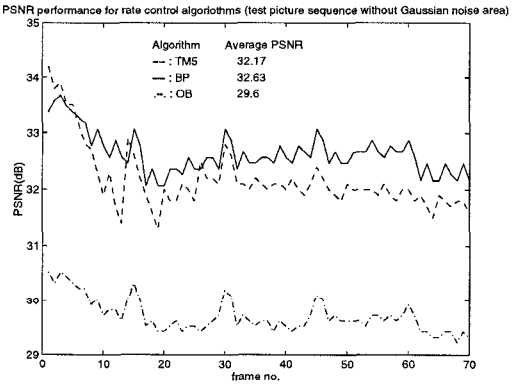

Figure 2: PSNR performance of the salesman portion inside the CCIR (*Test*) picture sequence