DELAY-POWER-RATE-DISTORTION MODEL FOR H.264 VIDEO CODING

Chenglin Li[1,2]*, Dapeng Wu*[1]*, Hongkai Xiong*[2]

[1]Department of Electrical and Computer Engineering, University of Florida, FL 32611, USA
[2]Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

ABSTRACT

In a video encoding system, the encoding performance would be determined by not only the rate-distortion (R-D) behavior, but also the encoding time and power consumption. To analyze and control the R-D behavior of the video encoding system under the encoding delay and energy constraints, in this paper, we extend the traditional R-D model to a novel delay-power-rate-distortion (d-P-R-D) model by including another two dimensions (the encoding time and power consumption), which quantifies the relationship among source encoding delay, rate, distortion and power consumption for IPPPP coding mode in H.264/AVC. Here, I and P stand for intra-coding and predictive coding, respectively. The model accuracy has been verified through experiments and compensated by considering the statistics of both the current frame and the previous frame.

Index Terms— Delay-power-rate-distortion model, video coding, H.264/AVC.

1. INTRODUCTION

Video encoding systems have experienced extensive growth in the last decades and been utilized for a wide range of applications. In order to optimize the R-D performance, conventionally, the bit rate and quantization distortion models have been extensively studied for the video encoder [1] [2] [3]. However, when the subsequent video transmission is jointly taken into account, video encoding delay and video encoding power consumption become two new constraints which, as well, would affect the overall R-D performance of the video communication systems and need to be applied to adjust the R-D behavior in the video encoder. From the perspective of the delay and power consumption, the constraints on the video encoder are two-fold. First, to save the amount of energy and reduce the probability of errors in data transmission, efficient video compression is required to significantly reduce the amount of data to be transmitted. Second, more efficient video compression often requires higher computational complexity, and thus larger power consumption and longer processing time in computing. As implied by these two conflicting aspects, there is a tradeoff among delay d, power consumption P, rate R, and video distortion D for the practical design of the video encoder.

This work was supported in part by National Science Foundation under grant ECCS-1002214, CNS-1116970 and the Joint Research Fund for Overseas Chinese Young Scholars under grant No. 61228101.

To derive the power-rate-distortion model for the video encoding system, [4] summarizes the encoding complexity of H.263 video encoder as three modules: motion estimation (ME), PRE-coding (transform, quantization, inverse quantization and inverse transform), and entropy coding, analyzes the relationship among the encoding complexity, rate, and distortion, and then use the power consumption level instead to represent the encoding complexity. However, this P-R-D model is proposed only for H.263 video encoder. Such model needs to be re-derived since H.264/AVC utilizes the tree structured motion compensation with seven inter modes, causing ME consumes much more encoding complexity than the other two modules. Note that [4] also fails to consider the dimension of the encoding time that is related to the encoding complexity.

In this paper, we develop accordingly an analytic framework to model the d-P-R-D behavior of the video encoder. Specifically, four dimensions (rate, distortion, delay and power) that jointly determine the performance of the H.264/AVC video encoder are derived as functions of several coding parameters (search range and number of reference frames in ME, and quantization step size), respectively. Here, without loss of generality, the coding structure of the H.264/AVC encoder is chosen to be IPPPP coding mode, which is also reasonable, since as will be introduced, the ME module for inter-coded P-frames takes the major part of the entire encoding complexity. The model accuracy has also been validated and compensated by considering the statistics of both the current frame and the previous frame. The rest of the paper is organized as follows. In Sec. 2, the d-P-R-D source coding model is derived for IPPPP coding mode in H.264/AVC. Sec. 3 verifies the model accuracy based on experiments. Some concluding remarks are given in Section 4.

2. D-P-R-D SOURCE CODING MODEL

2.1. Source Rate and Distortion Model

In the rate-distortion model [1], [3], both source rate and distortion for a hybrid video coder, e.g., H.264/AVC encoder, are based on the distribution of transformed residuals which is mainly determined by the ME accuracy and quantization distortion. For i.i.d. zero-mean Laplacian source under the uniform quantizer [3], the closed form functions of entropy of quantized transform coefficients and quantization distortion are derived as functions of the standard deviation of the transformed residuals σ and the quantization step size Q (or quantization parameter QP, there is a one-to-one mapping between Q and QP [5]), given by:

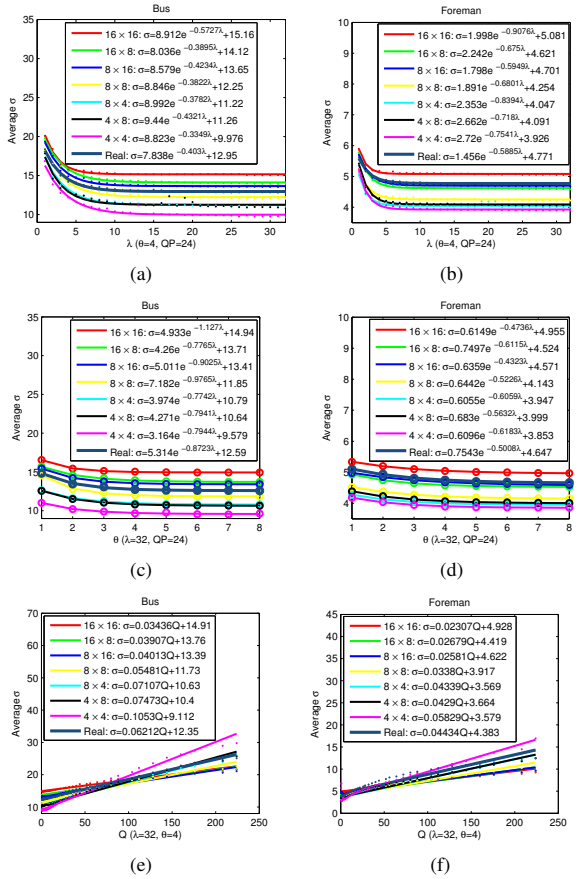


Fig. 1. σ vs. λ, θ, and Q, respectively.

$$R(\Lambda, Q) = H(\Lambda, Q) = -P_0 \log_2 P_0 + (1 - P_0)\left[\frac{\Lambda Q \log_2 e}{1 - e^{-\Lambda Q}} - \log_2(1 - e^{-\Lambda Q}) - \Lambda Q \gamma \log_2 e + 1\right] \quad (1)$$

$$D(\Lambda, Q) = \frac{\Lambda Q e^{\gamma \Lambda Q}(2 + \Lambda Q - 2\gamma \Lambda Q) + 2 - 2e^{\Lambda Q}}{\Lambda^2 (1 - e^{\Lambda Q})} \quad (2)$$

where $\Lambda = \sqrt{2}/\sigma$ is the Laplace parameter; γQ represents the rounding offset and γ is a parameter between $(0, 1)$, such as $1/6$ for H.264/AVC inter frame coding [1]; $P_0 = 1 - e^{-\Lambda Q(1-\gamma)}$ is the probability of quantized transform coefficient being zero.

To further analyze the ME accuracy, in H.264/AVC, σ depends on the following coding parameters. 1) Macroblock (MB) coding mode. For every MB in an inter-coded frame, ME has to choose one search mode from skip mode and the other seven inter modes with MB partitions of 16×16, 16×8, 8×16, 8×8, 8×4, 4×8 and 4×4 (indicated by index 1 to 7 as in JM configuration, i.e., assigning index 1 to 16×16 inter mode, index 2 to 16×8 inter mode, etc.). 2) ME search range λ and number of reference frames θ. Based on one dimensional search range λ, $(2\lambda + 1)^2$ specifies the number of pixels in a rectangular 2-D spatial search area in a reference frame for a MB to find the best match, such a 2-D search area times θ can form a 3-D search cube in ME. 3) Quantization step size Q. Since the reference frames are previously encoded, decoded and reconstructed frames, different Q will affect the distortion of the reference frames, and thus impact on σ.

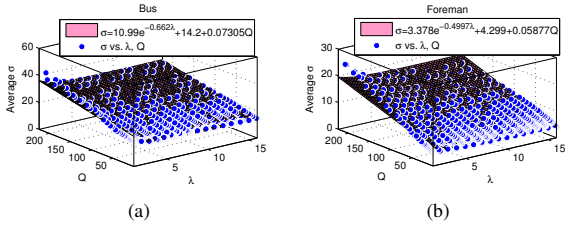
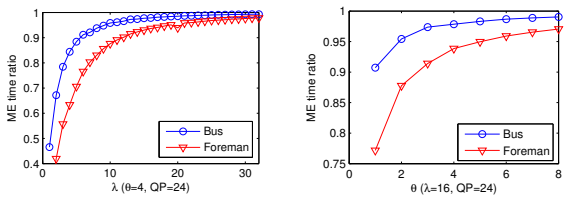


Fig. 2. Two dimensional fitting of σ vs. λ and Q.

To develop the source rate and distortion model, we need to find the closed form function of $\sigma(\lambda, \theta, Q)$. However, due to the lack of any prior knowledge of the exact function form, the basic method employed here is to draw the relationship of σ versus λ, θ, and Q and then fit such relationship with some known function forms. The JM18.2 [6] encoder is used to implement IPPPP coding mode for two video sequences, Bus (QCIF) and Foreman (CIF), and collect the average standard deviation of transformed residuals σ given different λ, θ and Q. Note that in practical H.264/AVC encoding, each MB can choose a specific inter mode based on R-D Optimization (RDO). For comparison, we can exclude the potential influence of MB coding mode in ME by enforcing all MBs select the same coding mode from the aforementioned eight inter modes except skip mode.

Since λ, θ and Q are independently tuned parameters in JM18.2 configurations, we first separately evaluate their impacts on the average σ. For all the seven inter modes and the real mode selection where each MB can choose the best inter mode based on RDO, respectively, Fig. 1 illustrates the curves of σ vs. λ, σ vs. θ, and σ vs. Q, while the other two coding parameters are fixed. It can be seen from Figs. 1(a) and 1(b) that every curve obtained under one of the seven inter modes or the real mode selection can be fitted by an exponential function with a constant vertical translation. In addition, as mode index increases from 1 to 7, the MB partition size used to find the best match becomes smaller, which leads to smaller predictive residuals and thus smaller σ under the same λ. Accordingly, σ vs. λ curves would slightly drop along σ direction as mode index increase. The curve of the real mode selection locates in the middle of these curves, since under the real mode selection, each MB can select a specific inter mode out of the seven inter modes. Likewise, Figs. 1(c) and 1(d) show that σ vs. θ curves can also be fitted by an exponential function plus a constant, while Figs. 1(e) and 1(f) illustrate that σ vs. Q curves can be simply fitted by a linear function.

As shown in Fig. 1, compared with the other two parameters, θ has little contribution to the change of σ. Therefore, in the case of the real mode selection, when all the three parameters, λ, θ and Q



(a) (b)

Fig. 3. Impact of λ and θ on the MET ratio.

are jointly taken into account, we could only investigate the function $\sigma(\lambda, Q)$ for given $\theta = \theta_0$ and use it instead to approximate the function $\sigma(\lambda, \theta, Q)$, for the sake of simplicity. Fig. 2 illustrates the two dimensional fitting results of function $\sigma(\lambda, Q)$ with $\theta = 1$. Considering both the exponential relationship with λ and linear relationship with Q, the two dimensional fitting function of σ can be represented in the form of

$$\sigma(\lambda, Q) = ae^{-b\lambda} + c + dQ \approx \sigma(\lambda, \theta, Q) \tag{3}$$

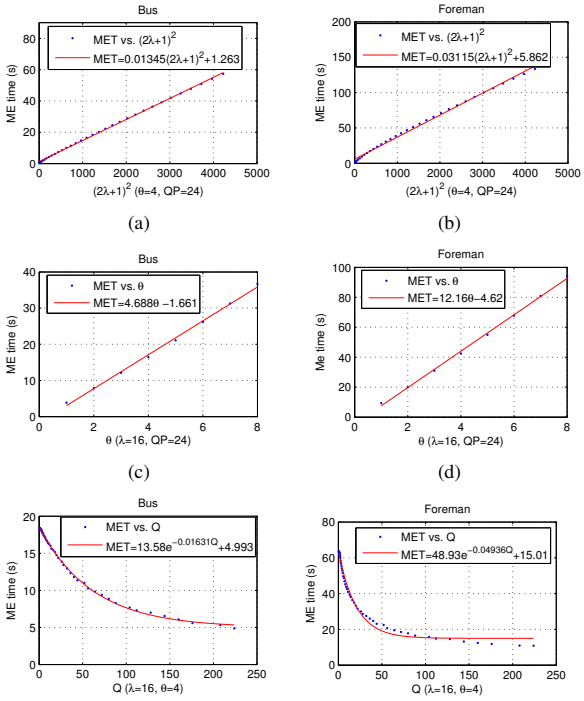
where a, b, c and d are fitting parameters. Plugging Eq. (3) and $\Lambda = \sqrt{2}/\sigma$ into Eqs. (1) and (2), both source rate and distortion can be further expressed as functions of λ and Q.

2.2. Encoding Time and Power Model

In order to achieve higher compression efficiency, H.264/AVC utilizes tree structured motion compensation with seven inter modes, which causes ME as the most time consuming part within all the three segments of the encoder. As justification, Fig. 3 illustrates the ratio of ME time to the entire encoding time, with λ and θ varying, respectively. It can be seen that ME takes majority of the entire encoding time, e.g., for Bus sequence, ME takes more than 90% of the entire encoding time for λ greater than 5 in Fig. 3(a) and any value of θ in Fig. 3(b). Such ME ratio grows and approaches to 1 with the increment of λ and θ. Therefore, it is reasonable to approximate the entire encoding time by the motion estimation time for IPPPP coding mode. It should be noted that the ME time ratio in Fig. 3 is based on exhaustive full search, which is used throughout this paper to guarantee achieving the optimal motion vector and thus minimum predictive residuals.

The relationship between the ME time (MET) and λ, θ and Q, respectively, are further investigated in Fig. 4, which indicates that MET can be fitted by a linear function of $(2\lambda+1)^2$, and a linear function of θ, The aforementioned linear relationship is straightforward, since $(2\lambda + 1)^2 \cdot \theta$ can form a 3-D search cube for a specific MB in ME and accordingly represent the total number of sum of absolute difference (SAD) operations per MB, which is proportional to the total ME complexity. However, MET is also a decreasing function of Q, which can be fitted by an exponential function plus a constant.

Theoretically, MET of an inter-coded P frame can be derived as the total number of CPU clock cycles consumed by SAD computations of that frame divided by the number of clock cycles per



(a) (b) (c) (d) (e) (f)

Fig. 4. MET vs. $(2\lambda + 1)^2$, θ, and Q, respectively.

second. Therefore, the encoding time for an inter-coded P frame approximated by the motion estimation time is

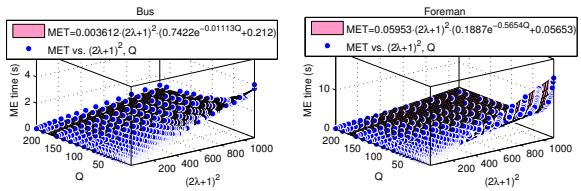
$$d(\lambda, \theta, Q) \approx MET(\lambda, \theta, Q) = \frac{N(2\lambda + 1)^2\theta \cdot \alpha(Q) \cdot c_0}{f_{CLK}} \tag{4}$$

where N is the number of MBs in a frame; $(2\lambda+1)^2\theta$ is the theoretical total number of SAD computations in a 3-D search cube for each MB; $\alpha(Q)$ is defined as the ratio of the actual number of SAD computations implemented by JM codec to the theoretical total number of SAD computations and thus $N(2\lambda + 1)^2\theta \cdot \alpha(Q)$ represents the actual number of SAD computations of a frame; c_0 is the number of clock cycles of one SAD computation for a given CPU; f_{CLK} is the clock frequency of that CPU. By introducing the dynamic voltage scaling model [7] [8] to dynamically control the power consumption of the microprocessor on a portable device in hardware design, f_{CLK} can be further related to the CPU power consumption:

$$P = k \cdot f_{CLK}^3 \tag{5}$$

where k is a constant defined in a specific dynamic voltage scaling model and determined by both the supply voltage and the effective switched capacitance of the circuits [9].

Next, we will experimentally justify the theoretical encoding time model proposed in Eq. (4). Fig. 5 illustrates the relationship of MET vs. λ and Q, with $\theta = 1$. Comparing the two dimensional fitting results with Eq. (4), the fitted functions comply with the functional form of $d(\lambda, \theta, Q|\theta = 1)$. For example, it can be observed



(a) (b)

Fig. 5. Two dimensional fitting of MET vs. $(2\lambda+1)^2$ and Q.

from comparison between Eq. (4) and the fitted function in Fig. 5(a) that, $\frac{N \cdot \alpha(Q) \cdot c_0}{f_{CLK}} = 0.003612 \cdot (0.7422e^{-0.01113Q} + 0.212)$, where $N = 99$ for Bus qcif sequence.

3. MODEL ACCURACY VERIFICATION

According to [3] [10], the transform coefficients in video encoder are not i.i.d., i.e., the 16 coefficients in a 4×4 integer transform show a decreasing variance in the well-known zigzag scan order as used in H.264/AVC. Specifically, suppose $(x, y),\ x, y \in \{0, 1, 2, 3\}$ is the position of a specific coefficient in the two dimensional transform domain of the 4×4 integer transform, the variance $\sigma^2_{(x,y)}$ is derived based on the average variance σ^2 of all positions [3].

$$\sigma^2_{(x,y)} = 2^{-(x+y)} \cdot \sigma^2_{(0,0)} = 2^{-(x+y)} \cdot \frac{1024}{225}\sigma^2 \tag{6}$$

Then the source rate and distortion model can be improved by considering different variances of different coefficients.

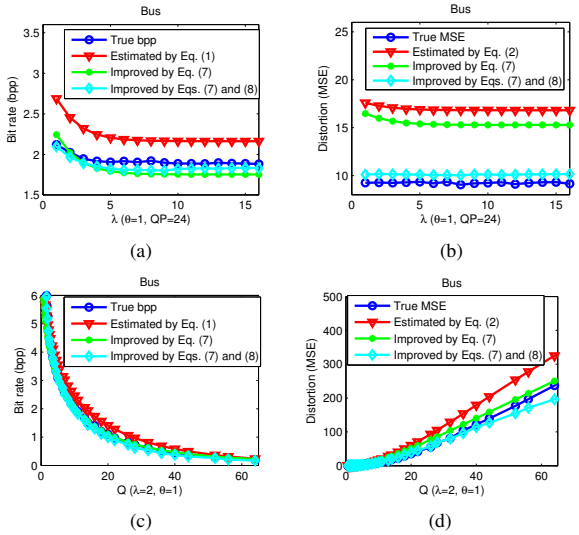
$$R = \frac{1}{16}\sum_{x=0}^{3}\sum_{y=0}^{3} R(\Lambda, Q)_{(x,y)},\ D = \frac{1}{16}\sum_{x=0}^{3}\sum_{y=0}^{3} D(\Lambda, Q)_{(x,y)} \tag{7}$$

Since Laplacian distribution may deviate significantly from the true residual histogram, the derived source and distortion models themselves are not accurate enough. By using the statistics from the previous frame, we can compensate the mismatch between the assumed Laplacian distribution and the true residual histogram for better estimation, where such mismatches between two consecutive frames (k and $k+1$) are assumed to be almost the same [3]:

$$R_t^k = \frac{R_t^{k-1} R_l^k}{R_l^{k-1}},\ D_t^k = \frac{D_t^{k-1} D_l^k}{D_l^{k-1}} \tag{8}$$

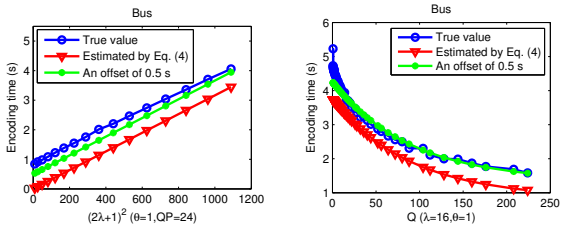
where l and t denote the estimated value under Laplacian distribution assumption and the true value, respectively. In Fig. 6, the accuracy of the proposed source rate and distortion model is tested in comparison to the true bit per pixel (bpp) and MSE values. The compensated source rate and distortion models based on Eq. (7), and the further improvement by jointly considering Eq. (8), are also illustrated.

On the other hand, the entire encoding time is approximated by its major part, ME module. The other two minor parts, PRE-coding and entropy coding time, though have little contribution to the encoding time and are thus neglected in Eq. (4), may result in the



(a) (b) (c) (d)

Fig. 6. Compensation for source rate and distortion models.



(a) (b)

Fig. 7. Compensation for encoding time model.

estimation error of the encoding time model. In a video sequence, however, the change of PRE-coding and entropy coding time as varying λ, θ or Q is trivial compared with that of the motion estimation time. Therefore, we can modify the encoding time model slightly by adding a small offset to compensate these two parts. Fig. 7 shows the true encoding time and the estimated encoding time by Eq. (4). It can be seen that the difference between the encoding time model (4) and its true value is relatively small, and as the encoding time increases such difference would become less significant.

4. CONCLUSION

In this paper, we derived the analytic delay-power-rate-distortion model for IPPPP coding mode in H.264/AVC to investigate the relationship among video encoding time, power, rate and distortion. Experimental results have verified and compensated the accuracy of the proposed d-P-R-D model, which can provide a theoretical guideline for the system design and performance optimization in video encoding systems. Our future work will focus on the d-P-R-D model based rate control problems for the video encoding system.

5. REFERENCES

[1] X. Li, N. Oertel, A. Hutter, and A. Kaup, "Laplace distribution based Lagrangian rate distortion optimization for hybrid video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 2, pp. 193–205, Feb. 2009.

[2] S. Ma, W. Gao, and Y. Lu, "Rate-distortion analysis for H. 264/AVC video coding and its application to rate control," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 12, pp. 1533–1544, Dec. 2005.

[3] Z. Chen and D. Wu, "Rate-distortion optimized cross-layer rate control in wireless video communication," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 3, pp. 352–365, Mar. 2012.

[4] Z. He, Y. Liang, L. Chen, I. Ahmad, and D. Wu, "Power-rate-distortion analysis for wireless video communication under energy constraints," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 5, pp. 645–658, May 2005.

[5] I. E. G. Richardson, "H.264/MPEG-4 Part 10: Transform and Quantization," 2003.

[6] K. Sühring, "H.264/AVC reference software JM18.2," Feb. 2012.

[7] R. Min, T. Furrer, and A. Chandrakasan, "Dynamic voltage scaling techniques for distributed microsensor networks," in *Proceedings of IEEE Computer Society Workshop on VLSI*, 2000, pp. 43–46.

[8] J. R. Lorch, A. J. Smith, et al., "Improving dynamic voltage scaling algorithms with PACE," *ACM SIGMETRICS Performance Evaluation Review*, vol. 29, no. 1, pp. 50–61, Jun. 2001.

[9] T. D. Burd and R. W. Brodersen, "Processor design for portable systems," *The Journal of VLSI Signal Processing*, vol. 13, no. 2, pp. 203–221, 1996.

[10] E.Y. Lam and J.W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Transactions on Image Processing*, vol. 9, no. 10, pp. 1661–1666, Oct. 2000.