Hardware Acceleration of Rate-Distortion Optimized Quantization Algorithm

Yusuke Funayama, Takashi Kambe†, Gen Fujita‡
Graduate School of Science and Engineering, Kindai University
3-4-1 Kowakae, Higashi-Osaka City, Osaka, Japan
†Depart. of Electric and Electronic Engineering, Kindai University
‡Faculty of Information and Communication Engineering,
Osaka Electro-Communication University, Osaka, Japan

Abstract— Rate-distortion optimized quantization (RDOQ) is an important technology for improving video coding performance. RDOQ is able to determine the optimal value among multiple quantization candidates based on rate-distortion (RD). This paper describes a hardware design of the RDOQ processing for a 4×4 block for inter-frame prediction using high-level synthesis technology [1] based on the improved RDOQ algorithm [4]. The hardware design results are also evaluated.

I. Introduction

Although the Rate-Distortion Optimized Quantization (RDOQ) in both H.264/AVC and H.265/HEVC improves the compression rate, processing times are nevertheless high and implementing it as hardware is complex.

In this paper we implement the algorithm in hardware using high-level synthesis and improve its performance.

II. Rate-Distortion optimized quantization (RDOQ)

The quantization process in both H.264/AVC and H.265/HEVC is implemented using the Rate-Distortion Optimized Quantization (RDOQ) technique. RDOQ is able to determine the optimal quantization level among multiple quantization candidates by minimizing the sum of rate-distortion (RD) costs in each block. RDOQ consists of five steps to and is able to achieve higher encoding performance than conventional approaches.

1. Three l_i^{float} rounding candidates, (a) zero level l_i^0, (b) floor rounding level l_i^{floor}, and (c) ceiling rounding level l_i^{ceil} are enumerated and their distortion values D_i^j are calculated. Here, i denotes the coefficient number, l_i^{float} denotes the quotient (before rounding) of DCT coefficient c_i divided by quantization step Q_{step}, and j denotes the rounding candidate level.
2. The last non-zero (LNZ) coefficient is found in zigzag scanning order to minimize RD cost.
3. The CBP (Coded Block Pattern) is estimated.
4. The bit-rate estimation of each rounding candidate and its RD cost is calculated using equation (1).

$$J_i^j = err_i^j + \lambda \times bits_i^j \quad (1)$$

   where, $bits_i^j$ represents the number of bits obtained by performing entropy coding on the quantized level l_i^j, err_i^j indicates the quantization error if the coefficient c_i is quantized to value l_i^j, and λ is constant for each quantization parameter (QP).
5. Determine whether all coefficients are set to zero by comparing steps 1 to 4 with the sum of RD costs.

RDOQ has a higher encoding performance than conventional techniques but it involves complex processes such as optimizing the quantization value for each quantization coefficient and updating the context of each coefficient. Also, because coefficients for bit-rate calculation are dependent on each other, hardware acceleration using parallel processing is difficult.

[2] proposed reduction of quantization level candidates and simplification of bit-rate estimation. In [4], we improved the bit-rate estimation approach in [2] further, and proposed three new improvements, (a) acceleration of distortion calculation, (b) reduction of LNZ coefficients search range, and (c) decision about whether all coefficients are zero.

III. Hardware Acceleration using high-level synthesis technology

Fig. 1 shows the architecture of the RDOQ processing based on the proposed RDOQ algorithm [4]. The target performance is real time processing for a 4×4 block for inter-frame prediction using high-level synthesis technology [1]. If this is achieved, the entire RDOQ process for a 720×480 frame will also be able to process in real time using a hierachical parallel processing technique.

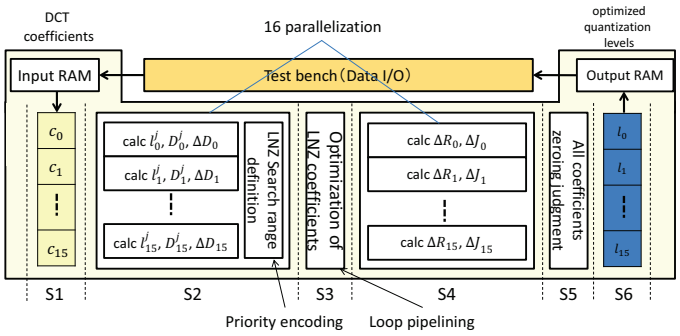

Fig. 1. Block diagram of RDOQ processing module

A. *Parallel processing of RD cost calculation*

Because [4] uses only the estimated bit-rate from equations for all coefficients, the proposed RDOQ algorithm can calculate the RD costs of all coefficients in parallel. However, when enumerating rounding candidates and deciding the distortion calculation LNZ coefficient search range there are data dependencies among coefficients. To deal with this problem, a search range flag is set at rounding candidates enumeration and the LNZ coefficient search range is determined using this flag after distortion calculation. This modification eliminates the data dependency between the two processes and all the coefficients can be processed in parallel (S2 and S4 in Fig. 1) All flags are bit-concatenated in reverse coefficient number order and the search range is decided using a one-cycle priority encoder in S2. One-cycle loop pipelining is also applied for each calculation at S3.

B. *Function based pipelining of RDOQ process*

The RDOQ process is implemented using six-stage functional pipelining. The pipeline consists of (S1)the input of DCT coefficients from RAM, (S2) rounding candidates enumeration and distortion calculation, (S3) optimization of LNZ coefficient, (S4) bit-rate estimation and RD cost calculation, (S5) decision of all coefficients to zero, and (S6) the output of the optimized quantization levels to RAM. The stages communicate using synchronization signals.

IV. Hardware design results

Five kinds of circuit, shown in Table I were designed to evaluate each of the acceleration methods. The gate level logic was synthesized from RT level HDL using Synopsys' Design Compiler and mapped to Hitachi 0.18 μm CMOS library cells. The clock frequency of all the circuits was 100MHz. The processing time of RDOQ calculation is for one frame(6248 4×4 blocks).

TABLE I
Design results

Acceleration method	Circuit size [gates]	Processing time [μs]
v0 (Sequential)	213,706	100,419
v1 (v0 + 16 parallel processing of S2 and S4)	778,654	30,654
v2 (v1 + loop pipelining of S3)	794,006	28,556
v3 (v2 + Priority encoding in S2)	785,035	28,049
v4 (v3 + functional pipelining)	830,464	4,038

Circuit v0 is a sequential Bach C description of the proposed algorithm. Circuit v1 executes 16 processes in parallel to calculate the 4×4 block coefficients at S2 and S4 in v0. In circuit v2 one cycle loop pipelining is applied at S3 in v1 by moving multi-cycle multiplications out of loop. Circuit v3 implements a priority encoder for search range reduction to v2. Circuit v4 adds functional pipelining of the RDOQ process to v3 and is able to process a 4×4 block in real time. Using this circuit, RDOQ on a 720×480 frame can be performed in real time.

V. Conclusion

In this paper, we implemented an improved RDOQ algorithm in hardware using the Bach C high-level synthesis tool. We designed an RDOQ processing circuit module capable of processing a 4×4 block in real time and plan to extend our methods to the whole RDOQ process for larger frame sizes.

Acknowledgements

The authors would like to thank the Bach system development group in SHARP Corporation, Electronic Components and Devices Development Group. This work is supported by the VLSI Design and Education Center (VDEC), the University of Tokyo in collaboration with Synopsys Corporation.

References

[1] K.Okada, et al. : "Hardware Algorithm Optimization Using Bach C,," IEICE Trans. Fundamentals vol.E85-A, No.4, pp835-841, 2002.

[2] Jing He, Fuzheng Yang: "High-speed implementation of rate-distortion optimized quantization for H.264/AVC," ©Springer-Verlag London 2013.

[3] Sullivan, G. J., et al.: "Overview of the High Efficiency Video Coding (HEVC) Standard", Circuits and Systems for Video Technology, IEEE Transactions on Circuits and Systems for Video Technology on 22(12), pp.1649-1668, 2012.

[4] Moriguchi, G., et al.: "An Improved Rate-Distortion Optimized Quantization Algorithm and its Hardware Implementation," Proceedings of SASIMI 2015, pp.409-414, 2015.