A Novel Efficient Rate Control Algorithm for Hardware Implementation in JPEG2000

Alireza Aminlou aminlou@cad.ece.ut.ac.ir
Omid Fatemi omid@fatemi.net
Department of Electrical and Computer Engineering
University of Tehran, Tehran – Iran

ABSTRACT

In multimedia applications where images are extremely used, rate control method has significant role in image encoder performance, computational complexity and hardware implementation. We propose a simple rate control algorithm suitable for hardware implementation of JPEG2000 encoder with less computational complexity and area. The proposed algorithm which is based on exponential modeling of R-D curves employs distortion instead of slope values. The simulation results show similar performance compared to full search method with considerable reduction in hardware resources.

1. INTRODUCTION

Uncompressed multimedia data, especially video and image, require considerable storage capacity and transmission bandwidth which are outstripped by demands for multimedia applications. A variety of image encoders including DCT-based and wavelet-based have been proposed for image and video compression. Performance of DCT-based still image encoders degrade at low bit-rates. However, wavelet transform has emerged as a new transform candidate within the field of image compression. Similar to conventional DCT-based encoders, quantization can be employed in DWT-based encoders as a simple rate control method. However, in order to obtain better performance in low bit-rate, a variety of intensive rate control schemes have been proposed.

Recently, DWT-based encoders have been proposed like EZW (Embedded Zero-tree Wavelet compression) by Shapiro [1], SPIHT (Spatial Partitioning of Images into Hierarchical Trees) by Said and Pearlman [2], LZC (Layered Zero Coding) by Taubman and Zakhor [3] and EBCOT (Embedded Block Coding with Optimized Truncation) by Taubman [4] which are progressive image encoders. Because of advantages of EBCOT over the other encoders, it is selected as the core of modern JPEG2000 standard [5].

JPEG2000 encoder consists of a number of steps: DC-level shift, component transform, wavelet transform, entropy coder and rate control. In multimedia applications and mobile communications where images are used extremely, rate control method has significant effect in image encoder performance, computational complexity and hardware implementation.

In EBCOT encoder, RGB components of main image are transformed into YCrCb components. 2-D Discrete Wavelet Transform (DWT) decomposes each component into four distinct LL, HL, LH and HH subbands. LL which is the low-resolution subband can be decomposed into four subbands, recursively. Wavelet coefficients in each subband (S_i) are quantized and converted to sign-magnitude format.

EBCOT divides wavelet subbands (S_i) into smaller blocks (e.g.32×32) which are called codeblocks. Each codeblock (C_i) is coded independently and an embedded bit-stream is constructed which can be truncated to rates R_i^n, where i is codeblock index and n is truncation point index within a codeblock. The contribution from codeblock C_i to distortion in reconstructed image is denoted by D_i^n for each truncation point n. By assuming the independency of codeblocks, overall distortion of reconstructed image is the sum of the distortion values imposed by each codeblock ($\sum D_i^{n_i}$). D_i^n can be estimated in different distortion metrics. Mean Square Error (MSE) is a well-known model which can be used in calculation of distortion as shown in (1).

$$D_i^n = \omega_i^2 \sum_{k \in C_i} \left(s_i[k] - \hat{s}_i^n[k] \right)^2 \tag{1}$$

The data rate and quality are the main constraints in image compression applications. Therefore, rate control problem can be defined as either rate constrained or distortion constrained. The former satisfies the rate constraint while obtaining the lowest image distortion and the latter satisfies distortion condition while producing the minimum data-rate, as shown in (2) and (3), respectively.

$$R = \sum_i R_i^{n_i} \le R_{\max} \qquad D = \sum_i D_i^{n_i} = D_{\min} \tag{2}$$

$$D = \sum_i D_i^{n_i} \le D_{\max} \qquad R = \sum_i R_i^{n_i} = R_{\min} \tag{3}$$

Rate-distortion (R-D) optimization procedure determines optimal selection of the truncation point set, $\{n_i\}$, in order to minimize distortion (rate) subject to the constraint R_{max} (D_{max}). A variety of procedures have been proposed including EBCOT [4], incremental (INC) [6] and statistical modeling [7] to solve the bit allocation problem. However, these procedures can be, depending on the number of wavelet levels and the size of image, very time consuming and require a large amount of hardware resources. In this paper, we propose a new R-D

optimization algorithm which is based on exponential modeling of R-D curves for rate constrained and distortion constrained bit allocation problems with reduced hardware requirements.

This paper is organized as follows. Existing R-D optimization algorithms are introduced in section 2. Our R-D optimization algorithm is presented in section 3. Section 4 discusses hardware implementation issues. Finally, experimental results and conclusions are presented in sections 5 and 6, respectively.

2. R-D OPTIMIZATION ALGORITHMS

Recently, a variety of R-D optimization algorithms have been proposed in the literature. We note that existing algorithms can be applied only to convex curves. Convexity imposes that the slope of the curve should be strictly decreasing. If these algorithms are applied to non-convex curves, they do not converge or may not result in optimum point. In order to guarantee the convexity of the curves, Convex Hull Analysis (CHA) should be applied prior to the execution of the algorithms.

All of the R-D optimization algorithms are based on Lagrange's theorem which states that slopes of the curves in selected truncation points should be equal for optimum points. In other words, truncation points with the same slope lead to optimum result. The description of the main R-D optimization algorithms, namely Lagrange multiplier (EBCOT) and incremental computation (INC) are presented in the following sections.

2.1. Lagrange Multiplier (EBCOT)

Classical Lagrange multiplier approach has been proposed for continuous functions. But it was shown that generalized Lagrange's theorem can also be applied to discrete functions [8]. Hence, it is used in EBCOT encoder for R-D optimization in rate control module [4]. It states that any set of truncation points, $\{n_i^{\lambda}\}$, which minimizes unconstrained problem (4), is an optimum solution for constrained problems (2) and (3). The optimization of equation (4) results in the fact that slopes of curves at selected truncation points should be equal to $-\lambda$, so in equations (4) and (5), n_i^{λ} denotes the truncation points of each codeblock where their slopes are equal to $-\lambda$.

$$D(\lambda)+\lambda R(\lambda)=\sum\left(D_i^{n_i^{\lambda}}+\lambda R_i^{n_i^{\lambda}}\right) \tag{4}$$

$$R(\lambda)=\sum R_i^{n_i^{\lambda}} \le R_{\max} \tag{5}$$

In (5), greater values of λ lead to smaller values of rate and vise versa. To determine the optimum truncation points, $\{n_i^{\lambda}\}$, a minimum value should be found for λ which satisfies the constraint (5). Thus for various values of λ, truncation points with slope $-\lambda$ are selected and (5) is checked until minimum value of λ is found. The minimum value of λ is determined by a bisection search until a defined precision is obtained [9]-[10].

2.2. Incremental Computation (INC)

INC algorithm which is proposed by Westerink [6], adds the truncation point with the greatest slope to current set of selected truncation points. This process is repeated until sum of $R_i^{n_i}$ reaches to the constraint R_{max}, where n_i denotes selected truncation point for C_i. The steps of the algorithm are as follows:

Initially, selected truncation points of all curves are reset to zero, i.e. $n_i \leftarrow 0$ which corresponds to zero rate and maximum distortion. n_i+1 denotes the next candidate truncation point for each codeblock which relies on the corresponding convex R-D curve. Within the set $\{n_i+1\}$, the truncation point with the greatest slope is determined ($n_{max}+1$). Current truncation point of the corresponding codeblock (C_{max}) is replaced by the next one, i.e. $n_{max} \leftarrow n_{max}+1$. This process is repeated until the sum of $R_i^{n_i}$ ($D_i^{n_i}$) reaches its limit. The drawback of this algorithm is the search complexity because of the number of curves especially for high bit-rate applications.

3. PROPOSED R-D OPTIMIZATION ALGORITHM

We note that both of the previous algorithms result in the best performance. However they have computational complexity and large amount of execution time. We propose a fast R-D optimization algorithms, based on exponential modeling of R-D curves with comparable performance to full search and substantial reduction in execution time and hardware resources.

3.1. Optimization of Exponential Curves

It was shown in ECM (Exponential Curve Modeling) algorithm that R-D curves can be approximately modeled by exponential form as shown in (6) where parameters D_i and $\Re_i$ are determined by curve fitting technique [11].

$$D_i^n \cong \breve{D}_i^n = D_i \exp(-\frac{R_i^n}{\Re_i}) \tag{6}$$

Based on generalized Lagrange's theorem and exponential assumption or R-D curves, analytical analysis results in the fact that the distortion value in optimum truncation points for every curve is in proportion of the corresponding time constant ($\Re_i$) as expressed in (7) which shows the distortion share (DS_i) of a codeblock C_i.

$$D_i^{n_i} \cong DS_i = \Re_i \times \frac{D_{\max}}{\sum \Re_i} \tag{7}$$

We note that in our previous algorithm, ECM, calculation of D_i and $\Re_i$ are time consuming. On the other hand, only a small number of curves contribute in final optimal result in particular curves of LL subbands. Hence,

we have proposed a Simplified ECM (SECM) which reduces computational complexity and time delay [11].

SECM classifies curves as either significant or insignificant by comparing their DS_i and D_i^0 values. The curve that is Distortion share value is less than its maximum distortion (i.e. $DS_i < D_i^0$) is named significant curve. Distortion share value of an insignificant curve is more than its maximum distortion (i.e. $DS_i > D_i^0$). It means that the insignificant curve can not contribute in total distortion more than D_i^0, so, the extra distortion (DS_i - D_i^0) is wasted, while it could be used in the other curves. Insignificant curves are discarded by allocating zero bits for them corresponding to D_i^0 as shown in (8). When all the insignificant curves are processed, remaining distortion can be divided among the significant curves.

We note that time constant values of the curves are similar [11]. Hence, we have proposed to consider time constants of curves as equal. Therefore, the remaining distortion can be divided among the curves equally using (9) where D_r and N_r denote the remaining amount of distortion and the remaining number of curves.

$$n_i = 0 \Leftrightarrow (R_i^{n_i}, D_i^{n_i}) = (0, D_i^0) \tag{8}$$

$$D_i^{n_i} \cong DS_i = D_r / N_r \tag{9}$$

3.2. Simplified Incremental (SINC)

We note that ECM and SECM algorithms work best in distortion constrained problems. Based on the procedure of INC and the idea of SECM, we propose a new algorithm, Simplified INC (SINC), that can be used in either rate constrained or distortion constrained bit allocation problems. Simulation results show that the performance of SINC algorithm is comparable to full search with higher clock frequency and much less computational complexity and hardware resources.

According to the modeling of R-D curves in SECM, the slopes of truncation points in R-D curves are in proportion with their distortion values. Because of similarity in time constants of the curves, we assume they are equal. This assumption results in the fact that comparing slope values in INC algorithm can be replaced by comparing distortion values. We also note that in this case, optimum truncation points could satisfy the condition of equality of distortion values. We propose that distortion values to be considered instead of slope values to determine the most suitable truncation point as the next point to be included in the set of selected points. SINC algorithm can be applied to non-convex curves as well as convex curves because distortion values of R-D curves are strictly decreasing. The steps of SINC algorithm which is modified INC algorithm are as follows:

First, selected truncation points of all curves are initialized to zero, i.e. $n_i \leftarrow 0$ which corresponds to zero rate and maximum distortion. n_i+1 denotes the next candidate truncation point for each codeblock. At the second step, within the set $\{n_i\}$, the truncation point with the greatest distortion value (n_{max}) is determined. Then, current truncation point of the corresponding codeblock (C_{max}) is replaced by the next one, i.e. $n_{max} \leftarrow n_{max}+1$. This process is repeated until the sum of $R_i^{n_i}$ reaches its limit.

It should be noted that the next candidate truncation point (n_i+1) does not require to be on convex curve. But, in INC algorithm, the next candidate truncation point (n_i+1) should be on convex curves. Thus convex hull analysis which is a complex algorithm is not required to be applied before SINC algorithm. Furthermore, more number of truncation points of original curves result in more opportunities in selection of optimum points.

4. HARDWARE IMPLEMENTATIONS

All of the algorithms including CHA, EBCOT, INC, SECM and SINC have been implemented in dedicated hardware using VHDL, verified by ModelSim and synthesized by LeonardoSpectrum towards the APEX20k FPGAs.

INC and SINC are pipeline implemented and have four stages in the pipe. Hardware implementations of algorithms are compared in Table 1. The normalized number of required clock cycles is also reported where *cb* and *try* are the number of codeblocks and the number of iterations in EBCOT algorithm, respectively. CHA, EBCOT and INC require two multipliers; SECM needs only one multiplier while SINC does not require any multiplier.

Since INC and EBCOT can only be employed with convex curves, they require a more complex data structure for implementing CHA. Therefore for complete implementation, CHA area should be added to INC and EBCOT (i.e. 1005+1656=2661and 1005+1355=2360). In terms of chip real estate, SINC has the minimum area and the highest clock frequency which makes it suitable for hardware implementation of JPEG2000 encoding system for real-time applications.

Table 1 Synthesis properties of algorithms

	clk cycles	Mult	LCs	Area	clk freq
CHA	40/cb	2	1005	1141	27.4
INC	2/(cb×rate)	2	1656	1696	21.4
EBCOT	2/(cb×try)	2	1355	1453	28.8
SECM	1/cb	1	532	557	27.1
SINC	2/(cb×cb)	--	126	194	63.2

5. EXPERIMENTAL RESULTS

Performance of INC and SINC algorithms are compared using convex curves (_cnv), semi-convex curves (_dd0) and non-convex curves (_non). Semi-convex case, proposed in our algorithm, means that the truncation points which have no contribution in improvement of distortion are eliminated from the curves. This process can

be performed using a simple comparison (dd = 0) which has much less computational cost with respect to CHA, with substantial improvement in performance of optimization algorithms. As shown in Figure 1, performance of INC using convex curves (INC_cnv) is the best while it has the worst performance in non-convex case (INC_non). Performance of SINC is comparable to optimum algorithm in all the cases.

Practically, smaller codeblock size is employed in hardware implementation of JPEG2000 in order to reduce memory requirements. Performance of INC and SINC algorithms is illustrated in Figure 2 for small codeblock size (16×16). It is shown that for small size of codeblocks, performance of INC for non-convex (INC_non) and semi-convex (INC_dd0) curves degrades. However, SINC has a comparable performance with all types of the curves.

We note that obtained rate in SINC algorithm is very close to desired rate where rate ration is defined as the ratio of the obtained rate to the requested rate. As shown in Table 2, rate ratios for ECM and SECM algorithms are about 1.31 and 1.27 respectively, while it is 1.03 for SINC algorithm.

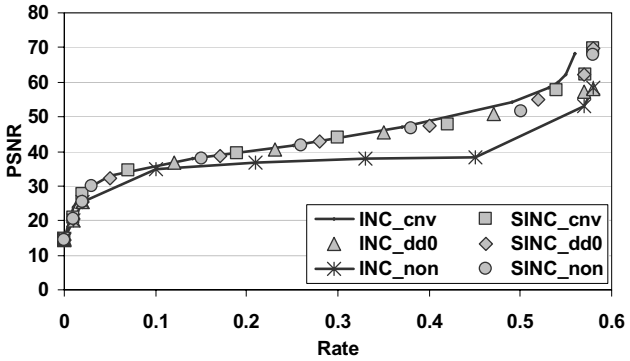

Figure 1 Performance of R-D optimization algorithms, codeblock size = 64x64, Number of resolution levels = 2

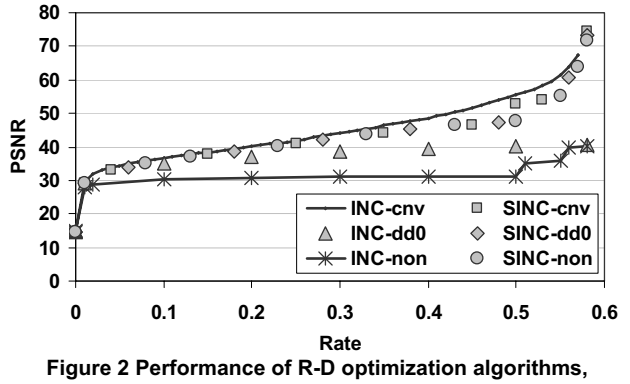

Figure 2 Performance of R-D optimization algorithms, codeblock size = 16x16, Number of resolution levels = 5

Table 2 Rate ration of R-D optimization algrithms

Algorithm	ECM	SECM	SINC	INC	EBCOT
Rate Ratio	1.31	1.27	1.03	1.00	1.00

6. CONCLUSIONS

In this paper, a new R-D optimization algorithm, Simplified INC (SINC), has been proposed which employs distortion values of truncation points in R-D curves instead of their slope values. SINC does not require to compute slopes of truncation points and can be applied to non-convex curves as well as convex curves. All of the algorithms have been implemented using VHDL and synthesized. Simulation results show that performance of SINC is comparable to full search algorithm with much less computational complexity and hardware resources. Hence, SINC is suitable for hardware implementation of JPEG2000 encoder for real time applications.

7. REFERENCES

[1] J. M. Shapiro, "An embedded hierarchical image coder using zerotrees of wavelet Coefficients," *IEEE Data Conference*, Snowbird, , pp. 214-223, 1993.

[2] A. Said and W. Pearlman, "A new, fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Transaction on Circuit and System Video Technology*, vol. 6, pp. 243-250, June 1996.

[3] D. Taubman and A Zakhor, "Multi-rate 3-D subband coding of video," *IEEE Transaction on Image Processing*, vol. 3, pp. 572-588, September 1994.

[4] D. Taubman, "High performance scalable image compression with EBCOT", *IEEE Transaction on Image Processing*, vol. 9, no. 7, July 2000.

[5] "JPEG2000 part I final draft international standard," ISO/IEC JTC1/SC29/WG1 N1890, Sept 2000.

[6] P.H. Westerink, J Biemond and D.E. Boekee, "An optimal bit allocation algorithm for subband coding", *In Proceeding of ICASSP'88*, pp. 757-760, 1988.

[7] P. Rault, C. Guillemot, "A simplified rate-distortion optimization procedure relying on statistical subband and noise modeling," *IEEE International Conference on Image Processing*, Chicago, USA, pp. 579-583, 1998.

[8] H. Everett III. "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources" *Operation Research*, vol. 11, pp. 399-417, 1963.

[9] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizes", *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 36, no. 9, pp. 1445-1453, September 1988.

[10] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distortion sense", *IEEE Transaction on Image Processing*, vol. 2, no. 2, pp. 160-175, April 1993.

[11] A. Aminlou and O. Fatemi, "Very fast bit allocation algorithm, based on simplified R-D curve modeling", *Proceedings of the 2003 10th IEEE International Conference on Electronics, Circuits and Systems*, vol. 1, pp. 112-115, December 2003.